Invent. math. 152, 37–55 (2003)
DOI: 10.1007/s00222-002-0266-3

Inventiones mathematicae

Entropy and the combinatorial dimension

S. Mendelson[1], **R. Vershynin**[2]

[1] Research School of Information Sciences and Engineering, The Australian National University, Canberra, ACT 0200, Australia (e-mail: shahar.mendelson@anu.edu.au)
[2] Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta T6G 2G1, Canada (e-mail: rvershynin@math.ualberta.ca)

Oblatum 10-XII-2001 & 4-IX-2002
Published online: 8 November 2002 –

Abstract. We solve Talagrand's entropy problem: the L_2-covering numbers of every uniformly bounded class of functions are exponential in its shattering dimension. This extends Dudley's theorem on classes of $\{0, 1\}$-valued functions, for which the shattering dimension is the Vapnik-Chervonenkis dimension. In convex geometry, the solution means that the entropy of a convex body K is controlled by the maximal dimension of a cube of a fixed side contained in the coordinate projections of K. This has a number of consequences, including the optimal Elton's Theorem and estimates on the uniform central limit theorem in the real valued case.

1. Introduction

The fact that the covering numbers of a set are exponential in its linear algebraic dimension is fundamental and simple. Let A be a class of functions bounded by 1, defined on a set Ω. If A is a finite dimensional class then for every probability measure on μ on Ω,

$$N(A, t, L_2(\mu)) \leq \left(\frac{3}{t}\right)^{\dim(A)}, \quad 0 < t < 1, \tag{1}$$

where $\dim(A)$ is the linear algebraic dimension of A and the left-hand side of (1) is the covering number of A, the minimal number of functions needed to approximate any function in A within an error t in the $L_2(\mu)$-norm. This inequality follows by a simple volumetric argument (see e.g. [Pi] Lemma 4.10) and is, in a sense, optimal: the dependence both on t and on the dimension is sharp (except, perhaps, for the constant 3).

The linear algebraic dimension of A is often too large for (1) to be useful, as it does not capture the "size" of A in different directions but

only determines in how many directions A does not vanish. The aim of this paper is to replace the linear algebraic dimension by a combinatorial dimension originated from the classical works of Vapnik and Chervonenkis [VC71], [VC81].

We say that a subset σ of Ω is t-shattered by a class A if there exists a level function h on σ such that, given any subset σ' of σ, one can find a function $f \in A$ with $f(x) \leq h(x)$ if $x \in \sigma'$ and $f(x) \geq h(x) + t$ if $x \in \sigma \setminus \sigma'$. The *shattering dimension* of A, denoted by $\mathrm{vc}(A, t)$ after Vapnik and Chervonenkis, is the maximal cardinality of a set t-shattered by A. Clearly, the shattering dimension does not exceed the linear algebraic dimension, and is often much smaller. Our main result states that the linear algebraic dimension in (1) can be essentially replaced by the shattering dimension.

Theorem 1. *Let A be a class of functions bounded by* 1*, defined on a set* Ω*. Then for every probability measure μ on* Ω*,*

$$N(A, t, L_2(\mu)) \leq \left(\frac{2}{t}\right)^{K \cdot \mathrm{vc}(A,\, ct)}, \quad 0 < t < 1, \tag{2}$$

where K and c are positive absolute constants.

There also exists a (simple) reverse inequality complementing (2): for some measure μ, one has $N(A, t, L_2(\mu)) \geq 2^{K \cdot \mathrm{vc}(A,\, ct)}$, where K and c are some absolute constants, see e.g. [T02].

The origins of Theorem 1 are rooted in the work of Vapnik and Chervonenkis, who first understood that entropy estimates are essential in determining whether a class of functions obeys the uniform law of large numbers. The subsequent fundamental works of Koltchinskii [K] and Ginè and Zinn [GZ] enhanced the link between entropy estimates and uniform limit theorems (see also [T96]).

In 1978, R. Dudley proved Theorem 1 for classes of $\{0, 1\}$-valued functions ([Du], see [LT] 14.3). This yielded that a $\{0, 1\}$-class obeys the uniform law of large numbers (and even the uniform Central Limit Theorem) if and only if its shattering dimension is finite for $0 < t < 1$. The main difficulty in proving such limit theorems for general classes has been the absence of a uniform entropy estimate of the nature of Theorem 1 ([T88], [T92], [T96], [ABCH], [BL], [T02]). However, proving Dudley's result for general classes is considerably more difficult due to the lack of the obvious property of the $\{0, 1\}$-valued classes, namely that if a set σ is t-shattered for some $0 < t < 1$ then it is automatically 1-shattered.

In 1992, M. Talagrand proved a weaker version of Theorem 1: under some mild regularity assumptions, $\log N(A, t, L_2(\mu)) \leq K \cdot \mathrm{vc}(A, ct) \log^M(\frac{2}{t})$, where K, c and M are some absolute constants ([T92], [T02]). Theorem 1 is Talagrand's inequality with the best possible exponent $M = 1$ (and without regularity assumptions).

Talagrand's inequality was motivated not only by limit theorems in probability, but to a great extent by applications to convex geometry. A subset

B of $\mathbb{R}^n$ can be viewed as a class of real valued functions on $\{1, \dots, n\}$. If B is convex and, for simplicity, symmetric, then its shattering dimension $\mathrm{vc}(B, t)$ is the maximal cardinality of a subset σ of $\{1, \dots, n\}$ such that $P_\sigma(B) \supset [-\frac{t}{2}, \frac{t}{2}]^\sigma$, where P_σ denotes the orthogonal projection in $\mathbb{R}^n$ onto $\mathbb{R}^\sigma$. In the general, non-symmetric, case we allow translations of the cube $[-\frac{t}{2}, \frac{t}{2}]^\sigma$ by a vector in $\mathbb{R}^\sigma$.

The following entropy bound for convex bodies is then an immediate consequence of Theorem 1. Recall that $N(B, D)$ is the covering number of B by a set D in $\mathbb{R}^n$, the minimal number of translates of D needed to cover B.

Corollary 2. *There exist positive absolute constants K and c such that the following holds. Let B be a convex body contained in $[0, 1]^n$, and D_n be the unit Euclidean ball in $\mathbb{R}^n$. Then for $0 < t < 1$*

$$N(B, t\sqrt{n}D_n) \le \left(\frac{2}{t}\right)^{Kd},$$

where d is the maximal cardinality of a subset σ of $\{1, \dots, n\}$ such that

$$P_\sigma(B) \supseteq h + [0, ct]^\sigma \text{ for some vector } h \text{ in } \mathbb{R}^n.$$

As M. Talagrand notices in [T02], Theorem 1 is a "concentration of pathology" phenomenon. Assume one knows that a covering number of the class A is large. All this means is that A contains many well separated functions, but it tells nothing about the structure these functions form. The conclusion of (2) is that A must shatter a large set σ, which detects a very accurate pattern: one can find functions in A oscillating on σ in all possible $2^{|\sigma|}$ ways around fixed levels. The "largeness" of A, *apriori* diffused, is *a fortiori* concentrated on the set σ.

The same phenomenon is seen in Corollary 2: given a convex body B with large entropy, one can find an entire cube in a coordinate projection of B, the cube that certainly *witnesses* the entropy's largeness.

When dualized, Corollary 2 solves the problem of finding the best asymptotics in Elton's Theorem. Let $x_1, \dots, x_n$ be vectors in the unit ball of a Banach space, and $\varepsilon_1, \dots, \varepsilon_n$ be Rademacher random variables (independent Bernoulli random variables taking values 1 and -1 with probability $1/2$). By the triangle inequality, the expectation $\mathbb{E}\|\sum_{i=1}^n \varepsilon_i x_i\|$ is at most n, and assume that $\mathbb{E}\|\sum_{i=1}^n \varepsilon_i x_i\| \ge \delta n$ for some number $\delta > 0$.

In 1983, J. Elton [E] proved an important result that there exists a subset σ of $\{1, \dots, n\}$ of size proportional to n such that the set of vectors $(x_i)_{i\in\sigma}$ is equivalent to the ℓ_1 unit-vector basis. Specifically, there exist numbers $s, t > 0$, depending only on δ, such that

$$|\sigma| \ge s^2 n \text{ and } \Big\| \sum_{i\in\sigma} a_i x_i \Big\| \ge t \sum_{i\in\sigma} |a_i| \text{ for all real numbers } (a_i). \tag{3}$$

Several steps have been made towards finding the best possible s and t in Elton's Theorem. A trivial upper bound is $s, t \le \delta$ which follows from the

example of identical vectors and by shrinking the usual ℓ_1 unit-vector basis. As for the lower bounds, J. Elton proved (3) with $s \sim \delta/\log(1/\delta)$ and $t \sim \delta^3$. A. Pajor [Pa] removed the logarithmic factor from s. M. Talagrand [T92], using his inequality discussed above, improved t to $\delta/\log^M(1/\delta)$. In the present paper, we use Corollary 2 to solve this problem by proving the optimal asymptotics: $s, t \sim \delta$.

Theorem 3. *Let $x_1, \dots, x_n$ be vectors in the unit ball of a Banach space, satisfying*

$$\mathbb{E}\Big\| \sum_{i=1}^n \varepsilon_i x_i \Big\| \geq \delta n \text{ for some number } \delta > 0.$$

Then there exists a subset $\sigma \subset \{1, \dots, n\}$ of cardinality $|\sigma| \geq c\delta^2 n$ such that

$$\Big\| \sum_{i\in\sigma} a_i x_i \Big\| \geq c\delta \sum_{i\in\sigma} |a_i| \text{ for all real numbers } (a_i),$$

where c is a positive absolute constant.

Furthermore, there is an interplay between the size of σ and the isomorphism constant – they can not attain their worst possible values together. Namely, we prove that s and t in (3) satisfy in addition to $s, t \gtrsim \delta$ also the lower bound $s \cdot t \log^{1.6}(2/t) \gtrsim \delta$, which, as an easy example shows, is optimal for all δ within the logarithmic factor. The power 1.6 can be replaced by any number greater than 1.5. This estimate improves one of the main results of the paper [T92] where this phenomenon in Elton's Theorem was discovered and proved with a constant (unspecified) power of logarithm.

The paper is organized as follows. In the remaining part of the introduction we sketch the proof of Theorem 1; the complete proof will occupy Sect. 2. Section 3 is devoted to applications to Elton's Theorem and to empirical processes.

Here is a sketch of the proof of Theorem 1. Starting with a set A which is separated with respect to the $L_2(\mu)$-norm, it is possible find a coordinate $\omega \in \Omega$ (selected randomly) on which A is diffused, i.e. the values $\{f(\omega),\ f \in A\}$ are spread in the interval $[-1, 1]$. Then there exist two nontrivial subsets A_1 and A_2 of A with their set of values $\{f(\omega),\ f \in A_1\}$ and $\{f(\omega),\ f \in A_2\}$ well separated from each other on the line. Continuing this process of separation for A_1 and A_2, etc., one can construct a dyadic tree of subsets of A, called a separating tree, with at least $|A|^{1/2}$ leaves. The "largeness" of the class A is thus captured by its separating tree.

The next step evoked from a beautiful idea in [ABCH]. First, there is no loss of generality in discretizing the class: one can assume that Ω is finite (say $|\Omega| = n$) and that the functions in A take values in $\frac{t}{6}\mathbb{Z} \cap [-1, 1]$. Then, instead of producing a large set σ shattered by A with a certain level function h, one can count the number of different pairs (σ, h) for which σ is shattered

by A with the level function h. If this number exceeds $\sum_{k=0}^{d}\binom{n}{k}(\frac{12}{t})^k$ then there must exist a set σ of size $|\sigma| > d$ shattered by A (because there are $\binom{n}{k}$ possible sets σ of cardinality k, and for such a set there are at most $(\frac{12}{t})^k$ possible level functions).

The only thing remaining is to bound below the number of pairs (σ, h) for which σ is shattered by A with a level function h. One can show that this number is bounded below by the number of the leaves in the separating tree of A, which is $|A|^{1/2}$. This implies that $|A|^{1/2} \leq \sum_{k=0}^{d}\binom{n}{k}(\frac{12}{t})^k \sim (\frac{n}{td})^d$, where $d = \mathrm{vc}(A, ct)$. The ratio $\frac{n}{d}$ can be eliminated from this estimate by a probabilistic extraction principle which reduces the cardinality of Ω.

Acknowledgements. The first author was supported by an Australian Research Council Discovery grant. The second author thanks Nicole Tomczak-Jaegermann for her constant support. He also acknowledges a support from the Pacific Institute of Mathematical Sciences, and thanks the Department of Mathematical Sciences of the University of Alberta for its hospitality. Finally, we would like to thank the referee for his valuable comments and suggestions.

2. The Proof of Theorem 1

For $t > 0$, a pair of functions f and g on Ω is *t-separated in $L_2(\mu)$* if $\|f - g\|_{L_2(\mu)} > t$. A set of functions is called t-separated if every pair of distinct points in the set is t-separated. Let $N_{\mathrm{sep}}(A, t, L_2(\mu))$ denote the maximal cardinality of a t-separated subset of A. It is standard and easily seen that

$$N(A, t, L_2(\mu)) \leq N_{\mathrm{sep}}(A, t, L_2(\mu)) \leq N\left(A, \frac{t}{2}, L_2(\mu)\right).$$

This inequality shows that in the proof of Theorem 1 we may assume that A is t-separated in the $L_2(\mu)$ norm, and replace its covering number by its cardinality.

We will need two probabilistic results, the first of which is straightforward.

Lemma 4. *Let X be a random variable and X' be an independent copy of X. Then*

$$\mathbb{E}|X - X'|^2 = 2\mathbb{E}|X - \mathbb{E}X|^2 = 2\inf_a \mathbb{E}|X - a|^2.$$

The next lemma is a small deviation principle. Denote by $\sigma(X)^2 = \mathbb{E}|X - \mathbb{E}X|^2$ the variance of the random variable X.

Lemma 5. *Let X be a random variable with nonzero variance. Then there exist numbers $a \in \mathbb{R}$ and $0 < \beta \leq \frac{1}{2}$, so that letting*

$$p_1 = \mathbb{P}\left\{X > a + \tfrac{1}{6}\sigma(X)\right\} \text{ and}$$
$$p_2 = \mathbb{P}\left\{X < a - \tfrac{1}{6}\sigma(X)\right\},$$

one has either $p_1 \geq 1 - \beta$ and $p_2 \geq \frac{\beta}{2}$, or $p_2 \geq 1 - \beta$ and $p_1 \geq \frac{\beta}{2}$.

Proof. Recall that a median of X is a number M_X such that $\mathbb{P}\{X \geq M_X\} \geq 1/2$ and $\mathbb{P}\{X \leq M_X\} \geq 1/2$; without loss of generality we may assume that $M_X = 0$. Therefore $\mathbb{P}\{X > 0\} = 1 - \mathbb{P}\{X \leq 0\} \leq 1/2$ and similarly $\mathbb{P}\{X < 0\} \leq 1/2$.

By Lemma 4,

$$\begin{aligned}\sigma(X)^2 \leq \mathbb{E}|X|^2 &= \int_0^\infty \mathbb{P}\{|X| > \lambda\}\, d\lambda^2 \\ &= \int_0^\infty \mathbb{P}\{X > \lambda\}\, d\lambda^2 + \int_0^\infty \mathbb{P}\{X < -\lambda\}\, d\lambda^2 \end{aligned} \tag{4}$$

where $d\lambda^2 = 2\lambda\, d\lambda$.

Assume that the conclusion of the lemma fails, and let c be any number satisfying $\frac{1}{3} < c < \frac{1}{\sqrt{8}}$. Divide $\mathbb{R}_+$ into intervals I_k of length $c\sigma(X)$ by setting

$$I_k = \Big(c\sigma(X)k,\ c\sigma(X)(k+1)\Big], k = 0, 1, 2, \ldots$$

and let $\beta_0, \beta_1, \beta_2, \ldots$ be the non-negative numbers defined by

$$\mathbb{P}\{X > 0\} = \beta_0 \leq 1/2,\ P\{X \in I_k\} = \beta_k - \beta_{k+1}, k = 0, 1, 2, \ldots$$

We claim that

$$\text{for all } k \geq 0,\ \beta_{k+1} \leq \frac{1}{2}\beta_k. \tag{5}$$

Indeed, assume that $\beta_{k+1} > \frac{1}{2}\beta_k$ for some k and consider the intervals $J_1 = (-\infty, c\sigma(X)k]$ and $J_2 = (c\sigma(X)(k+1), \infty)$. Then $J_1 = (-\infty, 0] \cup (\bigcup_{0 \leq l \leq k-1} I_l)$, so

$$\mathbb{P}\{X \in J_1\} = (1 - \beta_0) + \sum_{0 \leq l \leq k-1} (\beta_l - \beta_{l+1}) = 1 - \beta_k.$$

Similarly, $J_2 = \bigcup_{l \geq k+1} I_l$ and thus

$$\mathbb{P}\{X \in J_2\} = \sum_{l \geq k+1} (\beta_l - \beta_{l+1}) = \beta_{k+1} > \frac{1}{2}\beta_k.$$

Moreover, since the sequence (β_k) is non-increasing by its definition, then $\beta_k \geq \beta_{k+1} > \frac{1}{2}\beta_k \geq 0$ and $\beta_k \leq \beta_0 \leq \frac{1}{2}$. Then the conclusion of the lemma would hold with a being the middle point between the intervals J_1 and J_2 and with $\beta = \beta_k$, which contradicts the assumption that the conclusion of the lemma fails. This proves (5).

Now, one can apply (5) to estimate the first integral in (4). Note that whenever $\lambda \in I_k$,

$$\mathbb{P}\{X > \lambda\} \leq \mathbb{P}\{X > c\sigma(X)k\} = \mathbb{P}\Big(\bigcup_{l \geq k} I_l\Big) = \beta_k.$$

Then

$$\int_0^\infty \mathbb{P}\{X > \lambda\}\, d\lambda^2 \le \sum_{k\ge 0} \int_{I_k} \beta_k \cdot 2\lambda\, d\lambda$$
$$\le \sum_{k\ge 0} \beta_k \cdot 2c\sigma(X)(k+1)\ \mathrm{length}(I_k). \tag{6}$$

Applying (5) inductively, it is evident that $\beta_k \le (\frac{1}{2})^k \beta_0 \le \frac{1}{2^{k+1}}$, and since $\mathrm{length}(I_k) = c\sigma(X)$, (6) is bounded by

$$2c^2\sigma(X)^2 \sum_{k\ge 0} \frac{k+1}{2^{k+1}} = 4c^2\sigma(X)^2 < \frac{1}{2}\sigma(X)^2.$$

By an identical argument one can show that the second integral in (4) is also bounded by $\frac{1}{2}\sigma(X)^2$. Therefore

$$\sigma(X)^2 < \frac{1}{2}\sigma(X)^2 + \frac{1}{2}\sigma(X)^2 = \sigma(X)^2,$$

and this contradiction completes the proof. □

2.1. Constructing a separating tree

Let A be a finite class of functions on a probability space (Ω, μ), which is t-separated in $L_2(\mu)$. Throughout the proof we will assume that $|A| > 1$. One can think of the class A itself as a (finite) probability space with the uniform measure on it, that is, each element x in A is assigned probability $\frac{1}{|A|}$.

Lemma 6. *Let A be a t-separated subset of $L_2(\mu)$. Then, there exist a coordinate i in Ω and numbers $a \in \mathbb{R}$ and $0 < \beta \le 1/2$, so that setting*

$$N_1 = \left|\left\{x \in A :\ x(i) > a + \tfrac{1}{12}t\right\}\right| \ \textit{and}$$
$$N_2 = \left|\left\{x \in A :\ x(i) < a - \tfrac{1}{12}t\right\}\right|,$$

one has either $N_1 \ge (1-\beta)|A|$ and $N_2 \ge \frac{\beta}{2}|A|$, or vice versa.

Proof. Let x, x' be random points in A selected independently according to the uniform (counting) measure on A. By Lemma 4,

$$\mathbb{E}\|x - x'\|^2_{L_2(\mu)} = \mathbb{E}\int_\Omega |x(i) - x'(i)|^2\, d\mu(i) = \int_\Omega \mathbb{E}|x(i) - x'(i)|^2\, d\mu(i)$$
$$= 2\int_\Omega \mathbb{E}|x(i) - \mathbb{E}x(i)|^2\, d\mu(i) \tag{7}$$
$$= 2\int_\Omega \sigma(x(i))^2\, d\mu(i)$$

where $\sigma(x(i))^2$ is the variance of the random variable $x(i)$ with respect to the uniform measure on A.

On the other hand, with probability $1 - \frac{1}{|A|}$ we have $x \neq x'$ and, whenever this event occurs, the separation assumption on A implies that $\|x - x'\|_{L_2(\mu)} \geq t$. Therefore

$$\mathbb{E}\|x - x'\|^2_{L_2(\mu)} \geq \Big(1 - \frac{1}{|A|}\Big)t^2 \geq \frac{t^2}{2}$$

provided that $|A| > 1$.

Together with (7) this proves the existence of a coordinate $i \in \Omega$, on which

$$\sigma(x(i)) \geq \frac{t}{2}, \tag{8}$$

and the claim follows from Lemma 5 applied to the random variable $x(i)$. □

This lemma should be interpreted as a separation lemma for the set A. It means that one can always find two nontrivial subsets of A and a coordinate in Ω, on which the two subsets are separated with a "gap" proportional to t.

Based on Lemma 6, one can construct a large separating tree in A. Recall that a *tree of subsets* of a set A is a finite collection T of subsets of A such that, for every pair $B, D \in T$ either B and D are disjoint or one of them contains the other. We call D a *son* of B if D is a maximal (with respect to inclusion) proper subset of B that belongs to T. An element of T with no sons is called a *leaf*.

Definition 7. *Let A be a class of functions on Ω and $t > 0$. A t-*separating tree *T of A is a tree of subsets of A such that every element $B \in T$ which is not a leaf has exactly two sons B_+ and B_- and, for some coordinate $i \in \Omega$,*

$$f(i) > g(i) + t \ \textit{ for all } f \in B_+,\ g \in B_-.$$

Proposition 8. *Let A be a finite class of functions on a probability space (Ω, μ). If A is t-separated with respect to the $L_2(\mu)$ norm, then there exists a $\frac{1}{6}t$-separating tree of A with at least $|A|^{1/2}$ leaves.*

Proof. By Lemma 6, any finite class A which is t-separated with respect to the $L_2(\mu)$ norm has two subsets A_+ and A_- and a coordinate $i \in \Omega$ for which $f(i) > g(i) + \frac{1}{6}t$ for every $f \in A_+$ and $g \in A_-$. Moreover, there exists some number $0 < \beta \leq 1/2$ such that

$$|A_+| \geq (1 - \beta)|A| \ \text{ and } \ |A_-| \geq \frac{\beta}{2}, \text{ or vice versa.}$$

Thus, A_+ and A_- are sons of A which are both large and well separated on the coordinate i.

The conclusion of the proposition will now follow by induction on the cardinality of A. The proposition clearly holds for $|A| = 2$. Assume it holds for every t-separated class of cardinality bounded by N, and let A be a t-separated class of cardinality $N + 1$. Let A_+ and A_- be the sons of A as above; since $\beta > 0$, we have $|A_+|, |A_-| \leq N$. Moreover, if A_+ has a $\frac{1}{6}t$-separating tree with N_+ leaves and A_- has a $\frac{1}{6}t$-separating tree with N_- leaves then, by joining these trees, A has a $\frac{1}{6}t$-separating tree with $N_+ + N_-$ leaves, the number bounded below by $|A_+|^{1/2} + |A_-|^{1/2}$ by the induction hypothesis. Since $\beta \leq 1/2$,

$$\begin{aligned} |A_+|^{\frac{1}{2}} + |A_-|^{\frac{1}{2}} &\geq \big((1-\beta)|A|\big)^{\frac{1}{2}} + \left(\frac{\beta}{2}|A|\right)^{\frac{1}{2}} \\ &= \left[(1-\beta)^{\frac{1}{2}} + \left(\frac{\beta}{2}\right)^{\frac{1}{2}}\right]|A|^{\frac{1}{2}} \geq |A|^{\frac{1}{2}} \end{aligned}$$

as claimed. □

The exponent $1/2$ has no special meaning in Proposition 8. It can be improved to any number smaller that 1 at the cost of reducing the constant $\frac{1}{6}$.

2.2. Counting shattered sets

As explained in the introduction, our aim is to construct a large set shattered by a given class. We will first try to do this for classes of integer-valued functions.

Let A be a class of integer-valued functions on a set Ω. We say that a couple (σ, h) is a *center* if σ is a finite subset of Ω and h is an integer-valued function on σ. We call the cardinality of σ the dimension of the center. For convenience, we introduce (the only) 0-dimensional center $(\emptyset, \emptyset)$, which is the *trivial center*.

Definition 9. *The* set A shatters a center (σ, h) *if the following holds:*

- *either* (σ, h) *is trivial and* A *is nonempty,*
- *or, otherwise, for every choice of signs* $\theta \in \{-1, 1\}^\sigma$ *there exists a function* $f \in A$ *such that for* $i \in \sigma$

$$\begin{cases} f(i) > h(i) & \text{when } \theta(i) = 1, \\ f(i) < h(i) & \text{when } \theta(i) = -1. \end{cases} \tag{9}$$

It is crucial that both inequalities in (9) are strict: they ensure that whenever a d-dimensional center is shattered by A, one has $\mathrm{vc}(A, 2) \geq d$. In fact, it is evident that $\mathrm{vc}(A, 2)$ is the maximal dimension of a center shattered by A.

Proposition 10. *The number of centers shattered by* A *is at least the number of leaves in any* 1*-separating tree of* A.

Proof. Given a class B of integer-valued functions, denote by $s(B)$ the number of centers shattered by B. It is enough to prove that if B_+ and B_- are the sons of an element B of a 1-separating tree in A then

$$s(B) \geq s(B_+) + s(B_-). \tag{10}$$

By the definition of the 1-separating tree, there is a coordinate $i_0 \in \Omega$, such that $f(i_0) > g(i_0) + 1$ for all $f \in B_+$ and $g \in B_-$. Since the functions are integer-valued, there exists an integer t such that

$$f(i_0) > t \text{ for } f \in B_+ \text{ and } g(i_0) < t \text{ for } g \in B_-.$$

If a center (σ, h) is shattered either by B_+ or by B_-, it is also shattered by B. Next, assume that (σ, h) is shattered by both B_+ and B_-. Note that in this case $i_0 \notin \sigma$. Indeed, if the converse holds then σ contains i_0 and hence is nonempty. Thus the center (x, σ) is nontrivial and there exist $f \in B_+$ and $g \in B_-$ such that $t < f(i_0) < h(i_0)$ (by (9) with $\theta(i_0) = -1$) and $t > g(i_0) > h(i_0)$ (by (9) with $\theta(i_0) = 1$), which is impossible. Consider the center $(\sigma', h') = (\sigma \cup \{i_0\}, h \oplus t)$, where $h \oplus t$ is the extension of the function h onto the set $\sigma \cup \{i_0\}$ defined by $(h \oplus t)(i_0) = t$.

Observe that (σ', h') is shattered by B. Indeed, since B_+ shatters (σ, h), then for every $\theta \in \{-1, 1\}^\sigma \times \{1\}^{\{i_0\}}$ there exists a function $f \in B_+$ such that (9) holds for $i \in \sigma$. Also, since $f \in B_+$, then automatically $f(i_0) > t = h'(i_0)$. Similarly, for every $\theta \in \{-1, 1\}^\sigma \times \{-1\}^{\{i_0\}}$, there exists a function $f \in B_-$ such that (9) holds for $i \in \sigma$ and automatically $f(i_0) < t = h'(i_0)$.

Clearly, (σ', h') is shattered by neither B_+ nor by B_-, because $f(i_0) > t = h'(i_0)$ for all $f \in B_+$, so (9) fails if $\theta(i_0) = -1$; a similar argument holds for B_-.

Summarizing, $(\sigma, h) \to (\sigma', h')$ is an injective mapping from the set of centers shattered by both B_+ and B_- into the set of centers shattered by B but not by B_+ or B_-, which proves our claim. □

Combining Propositions 8 and 10, one bounds from below the number of shattered centers.

Corollary 11. *Let A be a finite class of integer-valued functions on a probability space (Ω, μ). If A is 6-separated with respect to the $L_2(\mu)$ norm then it shatters at least $|A|^{1/2}$ centers.*

To show that there exists a large dimensional center shattered by A, one must assume that the class A is bounded in some sense, otherwise one could have infinitely many low dimensional centers shattered by the class. A natural assumption is the uniform boundedness of A, under which we conclude a preliminary version of Theorem 1.

Proposition 12. *Let (Ω, μ) be a probability space, where Ω is a finite set of cardinality n. Assume that A is a class of functions on Ω into $\{0, 1, \ldots, p\}$,*

which is 6*-separated in* $L_2(\mu)$*. Set* d *to be the maximal dimension of a center shattered by* A*. Then*

$$|A| \leq \left(\frac{pn}{d}\right)^{Cd}, \tag{11}$$

where C *is an absolute constant. In particular, the same assertion holds for* $d = \mathrm{vc}(A, 2)$.

Proof. By Corollary 11, A shatters at least $|A|^{1/2}$ centers. On the other hand, the total number of centers whose dimension is at most d that a class of $\{0, 1, \dots, p\}$-valued functions on Ω can shatter is bounded by $\sum_{k=0}^{d} \binom{n}{k} p^k$. Indeed, for every k there exist at most $\binom{n}{k}$ subsets $\sigma \subset \Omega$ of cardinality k and, for each σ with $|\sigma| = k$ there are at most p^k level functions h for which the center (σ, h) can be shattered by such a class. Therefore $|A|^{1/2} \leq \sum_{k=0}^{d} \binom{n}{k} p^k$ (otherwise there would exist a center of dimension larger than d shattered by A, contradicting the maximality of d). The proof is completed by approximating the binomial coefficients using Stirling's formula. □

Actually, the ratio n/d can be eliminated from (11) (perhaps at the cost of increasing the separation parameter 6). To this end, one needs to reduce the size of Ω without changing the assumption that the class is "well separated". This is achieved by the following probabilistic extraction principle.

Lemma 13. *There is a positive absolute constant* c *such that the following holds. Let* Ω *be a finite set with the uniform probability measure* μ *on it. Let* A *be a class of functions bounded by* 1*, defined on* Ω*. Assume that for some* $0 < t < 1$

$$A \text{ is } t\text{-separated with respect to the } L_2(\mu) \text{ norm.}$$

If $|A| \leq \frac{1}{2}\exp(ct^4 k)$ *for some positive number* k*, there exists a subset* $\sigma \subset \Omega$ *of cardinality at most* k *such that*

$$A \text{ is } \frac{t}{2}\text{-separated with respect to the } L_2(\mu_\sigma) \text{ norm,}$$

where μ_σ *is the uniform probability measure on* σ.

As the reader guesses, the set σ will be chosen randomly in Ω. We will estimate probabilities using a version of Bernstein's inequality (see e.g. [VW], or [LT] 6.3 for stronger inequalities).

Lemma 14 (Berstein's inequality). *Let* $X_1, \dots, X_n$ *be independent random variables with zero mean. Then, for every* $u > 0$*,*

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{n} X_i\Big| > u\Big\} \leq 2\exp\Big(-\frac{u^2}{2(b^2 + a\,u/3)}\Big),$$

where $a = \sup_i \|X_i\|_\infty$ *and* $b^2 = \sum_{i=1}^{n} \mathbb{E}|X_i|^2$.

Proof of Lemma 13. For the sake of simplicity we identify Ω with $\{1, 2, \dots, n\}$. The difference set $S = \{f - g \mid f \neq g, f, g \in A\}$ has cardinality $|S| \leq |A|^2$. For each $x \in S$ we have $|x(i)| \leq 2$ for all $i \in \{1, ..., n\}$ and $\sum_{i=1}^n |x(i)|^2 \geq t^2 n$. Fix an integer k satisfying the assumptions of the lemma and let $\delta_1, \dots, \delta_n$ be independent $\{0, 1\}$-valued random variables with $\mathbb{E}\delta_i = \frac{k}{2n} =: \delta$. Then for every $z \in S$

$$\begin{aligned}\mathbb{P}\Big\{\sum_{i=1}^n \delta_i |x(i)|^2 \leq \frac{t^2\delta n}{2}\Big\} &\leq \mathbb{P}\Big\{\Big|\sum_{i=1}^n \delta_i |x(i)|^2 - \delta \sum_{i=1}^n |x(i)|^2\Big| > \frac{t^2\delta n}{2}\Big\} \\ &= \mathbb{P}\Big\{\Big|\sum_{i=1}^n (\delta_i - \delta)|x(i)|^2\Big| > \frac{t^2\delta n}{2}\Big\} \\ &\leq 2\exp\Big(-\frac{ct^4\delta n}{1+t^2}\Big) \leq 2\exp(-ct^4 k),\end{aligned}$$

where the last line follows from Bernstein's inequality for $a = \sup_i \|X_i\| \leq 2$ and

$$b^2 = \sum_{i=1}^n \mathbb{E}|X_i|^2 = \sum_{i=1}^n |x(i)|^4\, \mathbb{E}(\delta_i - \delta)^2 \leq 16\delta n.$$

Therefore, by the assumption on k

$$\mathbb{P}\Big\{\exists x \in S : \Big(\frac{1}{k}\sum_{i=1}^n \delta_i |x(i)|^2\Big)^{1/2} \leq \frac{t}{2}\Big\} \leq |S| \cdot 2\exp(-ct^4 k) < 1/2.$$

Moreover, if σ is the random set $\{i \mid \delta_i = 1\}$ then by Chebyshev's inequality,

$$\mathbb{P}\{|\sigma| > k\} = \mathbb{P}\Big\{\sum_{i=1}^n \delta_i > k\Big\} \leq 1/2,$$

which implies that

$$\mathbb{P}\Big\{\exists x \in S : \|x\|_{L_2(\mu_\sigma)} \leq \frac{t}{2}\Big\} < 1.$$

This translates into the fact that with positive probability the class A is $\frac{t}{2}$-separated with respect to the $L_2(\mu_\sigma)$ norm. □

Proof of Theorem 1. One may clearly assume that $|A| > 1$ and that the functions in A are defined on a finite domain Ω, so that the probability measure μ on Ω is supported on a finite number of atoms. Next, by splitting these atoms (by replacing an atom ω by, say, two atoms ω_1 and ω_2, each carrying measure $\frac{1}{2}\mu(\omega)$ and by defining $f(\omega_1) = f(\omega_2) = f(\omega)$ for $f \in A$), one can make the measure μ almost uniform without changing neither the covering numbers nor the shattering dimension of A. Therefore, assume that the domain Ω is $\{1, 2, \dots, n\}$ for some integer n, and that μ is the uniform measure on Ω.

Fix $0 < t \leq 1/2$ and let A be a $2t$-separated in the $L_2(\mu)$ norm. By Lemma 13, there is a set of coordinates $s \subset \{1, ..., n\}$ of size $|\sigma| \leq \frac{C \log |A|}{t^4}$ such that A is t-separated in $L_2(\mu_\sigma)$, where μ_σ is the uniform probability measure on σ.

Let $p = \lfloor 7/t \rfloor$, define $\tilde{A} \subset \{0, 1, ..., p\}^\sigma$ by

$$\tilde{A} = \left\{ \left(\left\lfloor \frac{7 f(i)}{t} \right\rfloor \right)_{i \in \sigma} \mid f \in A \right\},$$

and observe that $\tilde{A}$ is 6-separated in $L_2(\mu_\sigma)$. By Proposition 12,

$$|A| = |\tilde{A}| \leq \left(\frac{p|\sigma|}{d} \right)^{Cd}$$

where $d = \mathrm{vc}(\tilde{A}, 2)$, implying that

$$|A| \leq \left(\frac{C \log |A|}{dt^5} \right)^{Cd}.$$

By a straightforward computation,

$$|A| \leq \left(\frac{1}{t} \right)^{Cd},$$

and our claim follows from the fact that $\mathrm{vc}(\tilde{A}, 2) \leq \mathrm{vc}(A, t/7)$. □

Remark. Theorem 1 also holds for the $L_p(\mu)$ covering numbers for all $0 < p < \infty$, with constants K and c depending only on p. The only minor modification of the proof is in Lemma 4, where the equations would be replaced by appropriate inequalities.

3. Applications: Gaussian processes and convexity

The first application is a bound on the expectation of the supremum of a Gaussian processes indexed by a set A. Such a bound is provided by Dudley's integral in terms of the L_2 entropy of A; the entropy, in turn, can be majorized through Theorem 1 by the shattering dimension of A. The resulting integral inequality improves the main result of M. Talagrand in [T92].

If A be a class of functions on the finite set I, then a natural Gaussian process $(X_a)_{a \in A}$ indexed by elements of A is

$$X_a = \sum_{i \in I} g_i \, a(i)$$

where g_i are independent standard Gaussian random variables.

Theorem 15. *Let A be a class of functions bounded by* 1*, defined on a finite set I of cardinality n. Then $E = \mathbb{E}\sup_{a\in A} X_a$ is bounded as*

$$E \le K\sqrt{n}\int_{cE/n}^{1} \sqrt{\mathrm{vc}(A,t)\cdot\log(2/t)}\,dt,$$

where K and c are absolute positive constants.

The nonzero lower limit in the integral will play an important role in the application to Elton's Theorem.

The first step in the proof is to view A as a subset of $\mathbb{R}^n$. Dudley's integral inequality can be stated as

$$E \le K\int_0^\infty \sqrt{\log N(A, tD_n)}\,dt,$$

where D_n is the unit Euclidean ball in $\mathbb{R}^n$, see [Pi] Theorem 5.6. The lower limit in this integral can be improved by a standard argument. This fact was first noticed by A. Pajor.

Lemma 16. *Let A be a subset of $\mathbb{R}^n$. Then $E = \mathbb{E}\sup_{a\in A} X_a$ is bounded as*

$$E \le K\int_{cE/\sqrt{n}}^{\infty} \sqrt{\log N(A, tD_n)}\,dt,$$

where K is an absolute constant.

Proof. Fix positive absolute constants c_1, c_2 whose values will be specified later. There exists a subset $\mathcal{N}$ of A, which is a $(\frac{c_1E}{\sqrt{n}})$-net of A with respect to the Euclidean norm and has cardinality $|\mathcal{N}| \le N(A, \frac{c_1E}{2\sqrt{n}}D_n)$. Then $A \subset \mathcal{N} + \frac{c_1E}{2\sqrt{n}}D_n$, and one can write

$$E = \mathbb{E}\sup_{a\in A} X_a \le \mathbb{E}\max_{a\in\mathcal{N}} X_a + \mathbb{E}\sup_{a\in\frac{c_1E}{\sqrt{n}}D_n} X_a. \tag{12}$$

The first summand is estimated by Dudley's integral as

$$\mathbb{E}\max_{a\in\mathcal{N}} X_a \le K\int_0^\infty \sqrt{\log N(\mathcal{N}, tD_n)}\,dt. \tag{13}$$

On the interval $(0, \frac{c_2E}{\sqrt{n}})$,

$$\begin{aligned} K\int_0^{\frac{c_2E}{\sqrt{n}}} \sqrt{\log N(\mathcal{N}, tD_n)}\,dt &\le K\frac{c_2E}{\sqrt{n}}\cdot\sqrt{\log|\mathcal{N}|} \\ &\le K\frac{c_2E}{\sqrt{n}}\cdot\sqrt{\log N\big(A, \tfrac{c_1E}{2\sqrt{n}}D_n\big)}. \end{aligned}$$

The latter can be estimated using Sudakov's inequality [D,Pi], which states that $\varepsilon\sqrt{\log(N, \varepsilon D_n)} \le K\,\mathbb{E}\sup_{a\in A} X_a$ for all $\varepsilon > 0$. Indeed,

$$K\tfrac{c_2E}{\sqrt{n}} \cdot \sqrt{\log N\big(A, \tfrac{c_1E}{2\sqrt{n}}D_n\big)} \le K_1(2c_2/c_1)\,\mathbb{E}\sup_{a\in A} X_a = K_1(2c_2/c_1)E \le \tfrac{1}{4}E,$$

if we select $c_2 = c_1/8K_1$. Combining this with (13) implies that

$$\mathbb{E}\max_{x\in\mathcal{N}} X_a \le \frac{1}{4}E + K\int_{\frac{c_2E}{\sqrt{n}}}^{\infty} \sqrt{\log N(A, tD_n)}\,dt \tag{14}$$

because $\mathcal{N}$ is a subset of A.

To bound the second summand in (12), we apply the Cauchy-Schwarz inequality to obtain that for any $t > 0$,

$$\mathbb{E}\sup_{a\in tD_n} X_a \le t\cdot\mathbb{E}\Big(\sum_{i\in I} g_i^2\Big)^{1/2} \le t\sqrt{n}.$$

In particular, if $c_1 < 1/4$ then

$$\mathbb{E}\sup_{a\in\frac{c_1E}{\sqrt{n}}D_n} X_a \le c_1E \le \frac{1}{4}E.$$

This, (12) and (14) imply that

$$E \le K_2\int_{\frac{c_2E}{\sqrt{n}}}^{\infty} \sqrt{\log N(A, tD_n)}\,dt,$$

where K_2 is an absolute constant. □

Proof of Theorem 15. By Lemma 16,

$$E \le K\int_{cE/\sqrt{n}}^{\infty} \sqrt{\log N(A, tD_n)}\,dt.$$

Since $A \subset [-1, 1]^n \subset \sqrt{n}D_n$, the integrand vanishes for $t \ge \sqrt{n}$. Hence, by Theorem 1

$$\begin{aligned} E &\le K\int_{cE/\sqrt{n}}^{\sqrt{n}} \sqrt{\log N(A, tD_n)}\,dt \\ &= K\sqrt{n}\int_{cE/n}^{1} \sqrt{\log N(A, t\sqrt{n}D_n)}\,dt \\ &\le K_1\sqrt{n}\int_{cE/n}^{1} \sqrt{\mathrm{vc}(A, c_1t)\cdot\log(2/t)}\,dt. \end{aligned}$$

The absolute constant $0 < c_1 < 1/2$ can be made 1 by a further change of variable. □

The main consequence of Theorem 15 is Elton's Theorem with the optimal dependence on δ.

Theorem 17. *There is an absolute constant c for which the following holds. Let $x_1, \dots, x_n$ be vectors in the unit ball of a Banach space. Assume that*

$$\mathbb{E}\Big\| \sum_{i=1}^n g_i x_i \Big\| \geq \delta n \ \text{ for some number } \delta > 0.$$

Then there exist numbers $s, t \in (c\delta, 1)$, and a subset $\sigma \subset \{1, \dots, n\}$ of cardinality $|\sigma| \geq s^2 n$, such that

$$\Big\| \sum_{i\in\sigma} a_i x_i \Big\| \geq t \sum_{i\in\sigma} |a_i| \ \text{ for all real numbers } (a_i). \tag{15}$$

In addition, the numbers s and t satisfy the inequality $s \cdot t \log^{1.6}(2/t) \geq c\delta$.

Before the proof, recall the interpretation of the shattering dimension of convex bodies. If a set $B \subset \mathbb{R}^n$ is convex and symmetric then $\mathrm{vc}(B, t)$ is the maximal cardinality of a subset σ of $\{1, \dots, n\}$ such that $P_\sigma(B) \supset [-\frac{t}{2}, \frac{t}{2}]^\sigma$. Indeed, every convex symmetric set in $\mathbb{R}^n$ can be viewed as a class of functions on $\{1, ..., n\}$. If σ is t-shattered with a level function h then for every $\sigma' \subset \sigma$ there is some $f_{\sigma'}$ such that $f_{\sigma'}(i) \geq h(i) + t$ if $i \in \sigma'$ and $f_{\sigma'} \leq h$ on $\sigma \backslash \sigma'$. By selecting for every such σ' the function $(f_{\sigma'} - f_{\sigma\backslash\sigma'})/2$ and since the class is convex and symmetric, it follows that $P_\sigma(B) \supset [-\frac{t}{2}, \frac{t}{2}]^\sigma$, as claimed.

Taking the polars, this inclusion can be written as $\frac{t}{2}(B^\circ \cap \mathbb{R}^\sigma) \subset B_1^n$, where B_1^n is the unit ball of ℓ_1^n. Denoting by $\|\cdot\|_{B^\circ}$ the Minkowski functional (the norm) induced by the body B°, one can rewrite this inclusion as the inequality

$$\Big\| \sum_{i\in\sigma} a_i e_i \Big\|_{B^\circ} \geq \frac{t}{2} \sum_{i\in\sigma} |a_i| \ \text{ for all real numbers } (a_i),$$

where (e_i) is the standard basis of $\mathbb{R}^n$. Therefore, to prove Theorem 17, one needs to bound below the shattering dimension of the dual ball of a given Banach space.

Proof of Theorem 17. By a perturbation argument, one may assume that the vectors $(x_i)_{i\leq n}$ are linearly independent. Hence, using an appropriate linear transformation one can assume that $X = (\mathbb{R}^n, \|\cdot\|)$ and that $(x_i)_{i\leq n}$ are the unit coordinate vectors $(e_i)_{i\leq n}$ in $\mathbb{R}^n$. Let $B = (B_X)^\circ$ and note that the assumption $\|e_i\|_X \leq 1$ implies that $B \subset [-1, 1]^n$.

Set

$$E = \mathbb{E}\Big\| \sum_{i=1}^n g_i x_i \Big\|_X = \mathbb{E} \sup_{b\in B} \sum_{i=1}^n g_i\, b(i).$$

By Theorem 15,

$$\delta n \le E \le K\sqrt{n}\int_{c\delta}^{1}\sqrt{\mathrm{vc}(B,t)\cdot\log(2/t)}\,dt.$$

Consider the function

$$h(t) = \frac{c_0}{t\log^{1.1}(2/t)}$$

where the absolute constant $c_0 > 0$ is chosen so that $\int_0^1 h(t)\,dt = 1$. It follows that there exits some $c\delta \le t \le 1$ such that

$$\sqrt{\mathrm{vc}(B,t)/n\cdot\log(2/t)} \ge \delta h(t).$$

Hence

$$\mathrm{vc}(B,t) \ge \frac{c_0\delta^2}{t^2\log^{3.2}(2/t)}n.$$

Therefore, letting $s^2 = \mathrm{vc}(B,t)/n$, it follows that $s\cdot t\log^{1.6}(2/t) \ge \sqrt{c_0}\delta$ as required, and by the discussion preceding the proof there exists a subset σ of $\{1,\dots,n\}$ of cardinality $|\sigma| \ge s^2 n$ such that (15) holds with $t/2$ instead of t. The only thing remaining is to check that $s \gtrsim \delta$. Indeed, $s \ge \frac{\sqrt{c_0}\delta}{t\log^{1.6}(2/t)} \ge c_1\delta$, because $t \le 1$. □

Remarks. 1. As the proof shows, the exponent 1.6 can be reduced to any number larger than 3/2.

2. The relation between s and t in Theorem 17 is optimal up to a logarithmic factor for all $0 < \delta < 1$. This is seen from by the following example, shown to us by Mark Rudelson. For $0 < \delta < 1/\sqrt{n}$, the constant vectors $x_i = \delta\sqrt{n}\cdot e_1$ in $X = \mathbb{R}$ show that st in Theorem 17 can not exceed δ. For $1/\sqrt{n} \le \delta \le 1$, we consider the body $D = \mathrm{conv}(B_1^n \cup \frac{1}{\delta\sqrt{n}}D_n)$ and let $X = (\mathbb{R}^n, \|\cdot\|_D)$ and $x_i = e_i$, $i = 1,\dots,n$. Clearly, $\mathbb{E}\|\sum g_i x_i\|_X \ge \mathbb{E}\|\sum \varepsilon_i e_i\|_D = \delta n$. Let $0 < s,t < 1$ be so that (15) holds for some subset $\sigma \subset \{1,\dots,n\}$ of cardinality $|\sigma| \ge s^2 n$. This means that $\|x\|_D \ge t\|x\|_1$ for all $x \in \mathbb{R}^\sigma$. Dualizing, $\frac{t}{\delta\sqrt{n}}\|x\|_2 \le t\|x\|_{D^\circ} \le \|x\|_\infty$ for all $x \in \mathbb{R}^\sigma$. Testing this inequality for $x = \sum_{i\in\sigma} e_i$, it is evident that $\frac{t}{\delta\sqrt{n}}\sqrt{|\sigma|} \le 1$ and thus $st \le \delta$.

We end this article with an application to empirical processes. A key question is when a class of functions satisfies the Central Limit Theorem uniformly in some sense. Such classes of functions are called *uniform Donsker classes*. We will not define these classes formally but rather refer the reader to [D,VW] for an introduction to the subject. It turns out that the uniform Donsker property is related to uniform estimates on covering numbers via the Koltchinskii-Pollard entropy integral.

Theorem 18. [D] *Let F be a class of functions bounded by* 1. *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 condition in mind, it is natural to try to seek entropy estimates which are "dimension free", that is, do not depend on the size of the sample. In the $\{0, 1\}$-valued case, such bounds where first obtained by Dudley who proved Theorem 1 for these classes (see [LT] Theorem 14.13) which implied through Theorem 18 that every VC class is a uniform Donsker class.

Theorem 1 solves the general case: the following corollary extends Dudley's result on the uniform Donsker property from $\{0, 1\}$ classes to classes of real valued functions.

Corollary 19. *Let F be a class of functions bounded by* 1 *and assume that the integral*

$$\int_0^1 \sqrt{\mathrm{vc}(F, t) \log \frac{2}{t}}\, dt$$

converges. Then F is a uniform Donsker class.

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

This result has an advantage over Theorem 18 because in many cases it is easier to compute the shattering dimension of the class rather than its entropy (see, e.g. [AB]).

References

[AB] Anthony, M., Bartlett, P.L.: Neural Network Learning. Theoretical Foundations, Cambridge University Press 1999

[ABCH] Alon, N., Ben-David, S., Cesa-Bianchi, N., Hausser, D.: Scale sensitive dimensions, uniform convergence and learnability. J. ACM **44**, 615–631 (1997)

[BL] Bartlett, P.L., Long, P.M.: Prediction, learning, uniform convergence, and scale-sensitive diensions. J. Comput. Syst. Sci. **56**, 174–190 (1998)

[BKT] Bourgain, J., Kalton, N., Tzafriri, L.: Geometry of finite-dimensional subspaces and quotients of L_p. Geometric aspects of functional analysis, 138–175 (1987–88), Lect. Notes Math., 1376, Springer, Berlin 1989

[DGZ] Dudley, R.M., Giné, E., Zinn, J.: Uniform and universal Glivenko–Cantelli classes. J. Theor. Probab. **4**, 485–510 (1991)

[Du] Dudley, R.M.: Central limit theorems for empirical measures. Ann. Probab. **6**, 899–929 (1978)

[D] Dudley, R.M.: Uniform central limit theorems. Camb. Stud. Adv. Math. **63**, Cambridge University Press 1999

[E] Elton, J.: Sign-embeddings of ℓ_1^n. Trans. Am. Math. Soc. **279**, 113–124 (1983)

[GZ] Giné, E., Zinn, J.: Some limit theorems for empirical processes. Ann. Probab. **12**, 929–989 (1984)

[GZ91] Giné, E., Zinn, J.: Gaussian charachterization of uniform Donsker classes of functions. Ann. Probab. **19**, 758–782 (1991)

[K] Koltchinskii, V.I.: On the central limit theorem for empirical measures. Theory Probab. Math. Stat. **24**, 71–82 (1981)

[LT] Ledoux, M., Talagrand, M.: Probability in Banach spaces. Springer 1991

[Pa] Pajor, A.: Sous espaces ℓ_1^n des espaces de Banach. Hermann, Paris 1985

[Pi] Pisier, G.: The volume of convex bodies and Banach space geometry. Camb. Tracts Math. **94**, Cambridge University Press 1989

[T88] Talagrand, M: The Glivenko-Cantelli problem. Ann. Probab. **15**, 837–870 (1987)

[T92] Talagrand, M.: Type, infratype, and Elton-Pajor Theorem. Invent. Math. **107**, 41–59 (1992)

[T96] Talagrand, M.: The Glivenko-Cantelli problem, ten years later. J. Theor. Probab. **9**, 371–384 (1996)

[T02] Talagrand, M.: Vapnik-Chervonenkis type conditions and uniform Donsker classes of functions. Ann. Probab., to appear

[TJ] Tomczak-Jaegermann, N.: Computing 2-summing norm with few vectors. Ark. Mat. **17**, 273–277 (1979)

[VW] Van der Vaart, A., Wellner, J.: Weak convergence and empirical processes. Springer-Verlag 1996

[VC71] Vapnik, V., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Appl. **16**, 264–280 (1971)

[VC81] Vapnik, V., Chervonenkis, A.: Necessary and sufficient conditions for the uniform convergence of empirical means to their expectations. Theory Probab. Appl. **3**, 532–553 (1981)