Sharp estimate on the supremum of a class of sums of small i.i.d. random variables

Péter Major

Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences

Abstract

We take a class of functions $\mathcal{F}$ with polynomial covering numbers on a measurable space $(X, \mathcal{X})$ together with a sequence of independent, identically distributed X-space valued random variables $\xi_1, \dots, \xi_n$, and give a good estimate on the tail distribution of $\sup_{f\in\mathcal{F}} \sum_{j=1}^{n} f(\xi_j)$ if the expected values $E|f(\xi_1)|$ are very small for all $f \in \mathcal{F}$. In a subsequent paper [5] we give a sharp bound for the supremum of normalized sums of i.i.d. random variables in a more general case. But the proof of that estimate is based on the results in this work.

Keywords: Uniform covering numbers, Classes of functions with polynomially increasing covering numbers, Vapnik–Červonenkis classes, Hoeffding inequality
2010 MSC: 60F10, 60G50, 60G60

1. Introduction.

This work is part of a more general investigation about the supremum of (normalized) sums of bounded, independent and identically distributed random variables if the class of random variables whose sums we investigate have some nice properties. It turned out that it is useful to investigate first the case when the expectations of the absolute values of these random variables are very small, and this is the subject of the present paper. In paper [5] we shall get good estimates in the general case when the expectations of these

Email address: major.peter@renyi.mta.hu (Péter Major)
URL: http://www.renyi.hu/~major (Péter Major)

absolute values may be relatively large with the help of the main result in this paper.

First I recall the notion of uniform covering numbers and classes of functions with polynomial covering numbers, since they play an important role in our investigation. Then I formulate the main result of this paper, and make some comments that may help in understanding its content.

Definition of uniform covering numbers with respect to L_1-norm. Let a measurable space $(X, \mathcal{X})$ be given together with a class of measurable, real valued functions $\mathcal{F}$ on this space. The uniform covering number of this class of functions at level ε, $\varepsilon > 0$, with respect to the L_1-norm is $\sup_\nu \mathcal{N}(\varepsilon, \mathcal{F}, L_1(\nu))$, where the supremum is taken for all probability measures ν on the space $(X, \mathcal{X})$, and $\mathcal{N}(\varepsilon, \mathcal{F}, L_1(\nu))$ is the smallest integer m for which there exist some functions $f_j \in \mathcal{F}$, $1 \le j \le m$, such that $\min_{1 \le j \le m} \int |f - f_j| \, d\nu \le \varepsilon$ for all $f \in \mathcal{F}$.

Definition of a class of functions with polynomially increasing covering numbers. We say that a class of functions $\mathcal{F}$ has polynomially increasing covering numbers with parameter D and exponent L if the inequality

$$\sup_\nu \mathcal{N}(\varepsilon, \mathcal{F}, L_1(\nu)) \le D\varepsilon^{-L} \tag{1}$$

holds for all $0 < \varepsilon \le 1$ with the number $\sup_\nu \mathcal{N}(\varepsilon, \mathcal{F}, L_1(\nu))$ introduced in the previous definition.

The main result of this work is the following Theorem 1.

Theorem 1. *Let $\mathcal{F}$ be a finite or countable class of functions on a measurable space $(X, \mathcal{X})$ which has polynomially increasing covering numbers with some parameter $D \ge 1$ and exponent $L \ge 1$, and $\sup_{x \in X} |f(x)| \le 1$ for all $f \in \mathcal{F}$. Let $\xi_1, \dots, \xi_n$, $n \ge 2$, be a sequence of independent and identically distributed random variables with values in the space $(X, \mathcal{X})$ with a distribution μ, and assume that the inequality $\int |f(x)| \mu(\, dx) \le \rho$ holds for all $f \in \mathcal{F}$ with a number $0 < \rho \le n^{-200}$. Put $S_n(f) = S_n(f)(\xi_1, \dots, \xi_n) = \sum_{j=1}^n f(\xi_j)$ for all $f \in \mathcal{F}$. The inequality*

$$P\left(\sup_{f \in \mathcal{F}} |S_n(f)| \ge u \right) \le D\rho^{Cu} \quad \textit{for all } u > 41L \tag{2}$$

holds with some universal constant $1 > C > 0$. *We can choose e.g.* $C = \frac{1}{50}$.

I introduce an example that may help in understanding better the content of Theorem 1. In particular, it gives some hints why a condition of the type $u > \bar{C}L$ had to be imposed in formula (2). (We imposed this condition with $\bar{C} = 41$.)

Let us take a set $X = \{x_1, \dots, x_N\}$ with a large number N together with the uniform distribution μ on it, i.e. let $\mu(x_j) = \frac{1}{N}$ for all $1 \le j \le N$, and define the following class of function $\mathcal{F}$ on X. Fix a positive integer L, and let the class of functions $\mathcal{F}$ consist of the indicator functions of all subsets of X containing no more than L points. Let us fix a number n, and choose for all numbers $j = 1, \dots, n$ a point of the set X choosing each point with the same probability $\frac{1}{N}$ independently of each other. Let ξ_j denote the element of X we chose at the j-th time. In such a way we defined a sequence of independent random variables $\xi_1, \dots, \xi_n$ on X with distribution μ, and a class of functions $\mathcal{F}$ consisting of non-negative functions bounded by 1 such that $\int f(x)\mu(\,dx) \le \frac{L}{N}$ for all $f \in \mathcal{F}$. Let us introduce the random sums $S_n(f) = \sum_{j=1}^{n} f(\xi_j)$ for all $f \in \mathcal{F}$. We shall estimate first the probability $P_n = P\left(\sup_{f\in\mathcal{F}} S_n(f) \ge n\right)$ and then the probability $P_{u,n} = P\left(\sup_{f\in\mathcal{F}} S_n(f) \ge u\right)$ for $u \le n$.

It is not difficult to see that $P_n = 1$ if $n \le L$, and $P_n \le \binom{N}{L}(\frac{L}{N})^n \le C^L \rho^{n-L}$ with $\rho = \frac{L}{N}$ if $n > L$, where C is a universal constant. The number C can be chosen as a constant for which the inequality $p^p \le C^p p!$ holds for all positive integers p. We can choose for instance $C = 4$. In the proof of the above estimate we may exploit that X has $\binom{N}{L}$ subsets containing exactly L points, and the event $\sup_{f\in\mathcal{F}} S_n(f) \ge n$ may occur only if there is a subset of X of cardinality L which contains all points ξ_j, $1 \le j \le n$. Also the estimate $P_{u,n} \le \binom{n}{u} P_u \le C^L n^u \rho^{u-L}$ holds, because the event $\sup_{f\in\mathcal{F}} S_n(f) \ge u$ can only occur if there are some indices $1 \le j_1 < j_2 < \dots < j_u \le n$ such that all points ξ_{j_s}, $1 \le s \le u$, are contained in a subset of X of cardinality L. The probability of such an event is P_u for all sequences $1 \le j_1 < j_2 < \dots < j_u \le n$, and there are $\binom{n}{u}$ such sequences.

We show that if $N \ge n^{201}$ and $n \ge 41L$, then the above model satisfies the conditions of Theorem 1, and compare the bound we got for $P_{u,n}$ in our previous calculation with the estimate of Theorem 1 in this example. We

shall apply Theorem 1 for the class of functions $\mathcal{F}$ consisting of the indicator functions of all subsets containing at most L points of a set X. To apply Theorem 1 I show that the above defined $\mathcal{F}$ is a class of functions with polynomially increasing covering numbers with exponent L and an appropriate parameter D. Then we can apply the estimate of Theorem 1 for the probability $P_{u,n}$. To check the above stated property of $\mathcal{F}$ I recall the definition of Vapnik–Červonenkis classes together with a classical result about their properties.

Definition of Vapnik–Červonenkis classes. Let a set X be given, and let us select a class $\mathcal{D}$ of subsets of this set X. We call $\mathcal{D}$ a Vapnik–Červonenkis class if there exist two real numbers B and K such that for all positive integers n and subsets $S(n) = \{x_1, \dots, x_n\} \subset X$ of cardinality n of the set X the collection of sets of the form $S(n) \cap D$, $D \in \mathcal{D}$, contains no more than Bn^K subsets of $S(n)$. We call B the parameter and K the exponent of this Vapnik–Červonenkis class.

It is not difficult to see that the subsets of a set X containing at most L points constitute a Vapnik–Červonenkis class with exponent $K = L$ and an appropriate parameter B. (Some calculation shows that we can choose $B = \frac{1.5}{L!}$.) I recall a classical result (see e.g. [7] Chapter 2, 25 Approximation Lemma) by which the indicator functions of the sets in a Vapnik–Červonenkis class constitute a class of functions with polynomially increasing covering numbers. (Actually the work [7] uses a slightly different terminology, and it presents a more general result.) In the book [7] it is proved that if the parameter and exponent of a Vapnik–Červonenkis class are B and K, then the class of functions consisting of its indicator functions has polynomially increasing covering numbers with parameter $D = \max(B^2, n_0)$ and exponent $L = 2K$ with an appropriate constant $n_0 = n_0(K)$. Moreover, it is not difficult to see by slightly modifying the proof that this exponent can be chosen as $L = (1 + \varepsilon)K$ and an appropriate parameter $D = D(K, L, \varepsilon)$ for arbitrary $\varepsilon > 0$. Actually there are some improved versions of this result that supply slightly better estimates for this parameter and exponent (see [2] or [8] Theorem 2.6.4), but this improvement does not play an important role in our considerations.

The above argument shows that the class of functions $\mathcal{F}$ considered in the above example has polynomially increasing covering numbers with exponent $2L$ and an appropriate parameter D. Its exponent can be chosen even as $(1 + \varepsilon)L$ with an appropriate parameter $D(\varepsilon)$ for all $\varepsilon > 0$. This means in

particular that Theorem 1 can be applied to estimate the probability $P_{u,n}$ if the numbers L, N and n are appropriately chosen. It can be proved that both Theorem 1 and our previous argument provide an estimate of the form $P_{u,n} \le \rho^{\alpha u}$ with a universal constant $0 < \alpha < 1$, only the parameter α is different in these two estimates. (Observe that $\rho = \frac{L}{N} \ge \int f(x)\mu(\,dx)$ for all $f \in \mathcal{F}$ in this example.). To see that we proved such an estimate for $P_{u,n}$ which implies the inequality $P_{u,n} \le \rho^{\alpha u}$ under the conditions of Theorem 1 observe that $\rho^{u-L} \le \rho^{40u/41}$, and $n^u \le \rho^{-u/200}$. Moreover, it can be seen that if we are not interested in the value of the universal parameter α, then this estimate is sharp. I also remark that we can give a useful estimate for $P_{u,n}$ in this example (and not only the trivial bound $P_{u,n} \le 1$) only if $u > L$.

The main content of Theorem 1 is that a similar picture arises if the supremum of the sums we consider is defined with the help of an arbitrary class of functions with polynomially increasing covering numbers. Namely, Theorem 1 states that if $\mathcal{F}$ is a class of functions with polynomially increasing covering numbers with some exponent L and parameter D that satisfies some natural conditions, then there are universal constants $0 < \alpha < 1$, $C_1 > 1$ and $C_2 > 0$ such that $P\left(\sup_{f\in\mathcal{F}} S_n(f) > u\right) \le D\rho^{-\alpha u}$ if $n \ge C_1 L$ and $\rho \le n^{-C_2}$. Here we applied the notations of Theorem 1. We also gave an explicit value for these universal parameters in Theorem 1, but we did not try to find a really good choice. It might be interesting to show on the basis of the calculation of the present paper that we can choose $C_1 = 1 + \varepsilon$ or $\alpha = 1 - \varepsilon$ with arbitrary small $\varepsilon > 0$ if the remaining universal constants are appropriately chosen.

As the above considered example shows the estimate of Theorem 1 holds only if $u \ge CL$ with a number $C \ge 1$. The other condition of Theorem 1 by which $\rho \le n^{-C_2}$ with a sufficiently large number $C_2 > 0$ can be weakened. Actually this is the topic of paper [5] which is a continuation of the present work. In paper [5] I shall consider such classes of functions $\mathcal{F}$ with polynomially increasing covering numbers for which the parameter ρ considered in Theorem 1 can be relatively large. On the other hand, in [5] we shall consider only such classes of functions $\mathcal{F}$ whose elements have the 'normalizing property' $\int f(x)\mu(\,dx) = 0$ for all $f \in \mathcal{F}$. In the present work we did not impose such a normalization condition, because in the case $\rho \le n^{-\alpha}$ with some $\alpha > 1$ the lack of normalization has a negligible effect.

Theorem 1 will be proved with the help of Theorem 1A formulated below.

After its formulation I shall explain that Theorem 1A can be considered as a special case of Theorem 1.

Theorem 1A. *Let $X = \{x_1, \dots, x_N\}$ be a finite set of N elements, and let $\mathcal{X}$ be the σ-algebra consisting of all subsets of X. Let μ denote the uniform distribution on X, i.e. let $\mu(A) = \frac{|A|}{N}$ for all sets $A \subset X$, where $|A|$ denotes the cardinality of the set A. Let $\mathcal{F}$ be a class of functions with polynomially increasing covering numbers with some parameter $D \ge 1$ and exponent $L \ge 1$ on the measurable space $(X, \mathcal{X})$ such that $0 \le f(x) \le 1$ for all $x \in X$ and $f \in \mathcal{F}$, and $\int f(x)\mu(\,dx) \le \frac{\rho}{2}$ for all $f \in \mathcal{F}$ with some $\rho > 0$ which satisfies the inequality $\rho \le \min(\frac{1}{1000}, L^{-20})$.*

For the sake of a simpler argument we shall assume that the number N has the special form $N = 2^k N_0$ with some integer $k \ge 0$ and a number N_0 that satisfies the inequality $\frac{1}{16}\rho^{-3/2} < N_0 \le \frac{1}{8}\rho^{-3/2}$. (Actually we could choose an arbitrary number $N \ge \frac{1}{16}\rho^{-3/2}$ in Theorem 1A, but this special choice of N makes our argument simpler.)

Introduce for all numbers $p = 1, 2, \dots$ the p-fold direct product X^p of the space X together with the p-fold product measure μ_p of the uniform distribution μ on X, i.e. let each sequence $x^{(p)} = (x_{s_1}, \dots, x_{s_p})$, $x_{s_j} \in X$, $1 \le j \le p$, have the weight $\mu_p(x^{(p)}) = \frac{1}{N^p}$ with respect to the measure μ_p.

Given a function $f \in \mathcal{F}$ and a positive integer p let us define the set $B_p(f) \subset X^p$ for all $p \ge 2$ by the formula

$$B_p(f) = \{x^{(p)} = (x_{s_1}, \dots, x_{s_p})\colon\ x^{(p)} \in X^p, \quad f(x_{s_j}) = 1 \quad \textit{for all } 1 \le j \le p\}, \tag{3}$$

and put

$$B_p = B_p(\mathcal{F}) = \bigcup_{f \in \mathcal{F}} B_p(f). \tag{4}$$

If $p \ge 2L$ and $p \le \rho^{-1/100}$, then there exist some universal constants $C_1 > 0$ and $1 > C_2 > 0$ such that

$$\mu_p(B_p) = \mu_p(B_p(\mathcal{F})) \le C_1 D \rho^{C_2 p}. \tag{5}$$

We can choose for instance $C_1 = 2$ and $C_2 = \frac{1}{4}$.

In Theorem 1A we considered a special case of the problem discussed in Theorem 1. We took a space of the form $X = \{x_1, \dots, x_N\}$ with the uniform distribution μ on it, and considered a class of functions with polynomially increasing covering numbers and some additional special properties.

If we apply Theorem 1 with the choice $p = n$, then the event $B_p(\mathcal{F})$ defined in (4) agrees with the event $\sup_{f\in\mathcal{F}} S_n(f) \ge n$, and formula (5) implies the estimate (2) with the special choice $u = n$ for the system X, $\mathcal{F}$, μ considered in Theorem 1A.

It may be worth mentioning that in Theorem 1A we considered a class of functions $\mathcal{F}$ which contains functions f with the property $0 \le f \le 1$, while the event $B_p(\mathcal{F})$ whose probability we have estimated depended only on the sets where these functions f take value 1. Hence the event $B_p(\mathcal{F})$ would not change if we replaced all functions $f \in \mathcal{F}$ by the smaller functions $f \cdot I_{\{f=1\}}$, where $I_{\{f=1\}}$ denotes the indicator function of the set $\{x\colon\ f(x) = 1\}$. Nevertheless, it was useful to formulate the result in the form as we did, because by replacing the functions f by $f \cdot I_{\{f=1\}}$ we may get such a class of functions which has not polynomially increasing covering numbers. We shall prove Theorem 1 with the help of Theorem 1A formulated in the present form.

In Section 2 I make some remarks and discuss some examples which may explain the motivation behind the investigation of this paper. Theorem 1A will be proved by means an appropriate induction procedure in Section 3. Theorem 1 will be proved in Section 4 with the help of Theorem 1A and a good approximation.

2. A discussion of the results in this paper.

To understand the results of this work it may be useful to consider the following problem. Let us have a finite set $X = \{x_1, \dots, x_N\}$ with large cardinality N and a class $\mathcal{F}$ of subsets of X which is a Vapnik–Ĉervonenkis class with some parameter $B \ge 1$ and exponent $K \ge 1$, and such that all sets $A \in \mathcal{F}$ contain no more than ρN points, where ρ is a relatively small number, say $\rho \le \min(\frac{1}{2000}, (2K)^{-30})$. Let us choose n points from the set X randomly and independently from each other so that at each step we choose every point of X with the same probability $\frac{1}{N}$. Let $n \ge \alpha K$ with an appropriately chosen sufficiently large but fixed universal constant α. Give a good estimate on the probability that one of the sets $A \in \mathcal{F}$ contains each one of the n selected points.

This probability is very small, and it can be well bounded by means of Theorem 1A. Indeed, if we define the class of functions $\tilde{\mathcal{F}}$ which contains the indicator functions of the sets in $\mathcal{F}$, then $\tilde{\mathcal{F}}$ is a class of functions with

polynomially increasing covering numbers which satisfies the conditions of Theorem 1A, and the probability we want to estimate equals the probability of the event $B_p(\tilde{\mathcal{F}})$ defined in (4) with $p = n$. Hence Theorem 1A gives the estimate $C_1 B \rho^{C_2 n}$ with some universal constants $C_1 > 0$ and $C_2 > 0$ for the probability we want to estimate.

Theorem 1 can be useful in the solution of the following generalized version of the previous problem. Let us consider the same class of functions $\mathcal{F}$ as before, and choose randomly n points from the set X independently of each other by the uniform distribution, and let us assume this time that $\rho \le n^{-1/200}$. Let us bound the probability of the event that one of the sets $A \in \mathcal{F}$ contains at least αK elements from the randomly chosen n points, where α is a sufficiently large fixed number, and K is the exponent of the Vapnik–Červonenkis class $\mathcal{F}$. This probability can be estimated similarly to the previous case, only we have to apply Theorem 1 instead of Theorem 1A. We get the estimate $B\rho^{\alpha CK}$ with a universal constant C.

Similar problems can be solved with the help of results which give a good bound on the probability at the left-hand side of formula (2) if $\mathcal{F}$ is a class of functions with polynomially increasing kernel functions, and the inequality $\int f(x)\mu(\,dx) \le \rho$ holds for all functions $f \in \mathcal{F}$ with a small number ρ. (The measure μ in this formula is the distribution of the random variables which take part in the definition of the quantity $S_n(f)$ appearing in Theorem 1.) I know some results in the literature in this direction, but they do not provide a good estimate in the previous problems. In particular, they do not show that the above considered probabilities are very small even in such cases when we consider relatively short sequences of selected points in the first problem or we are looking for such a set $A \in \mathcal{F}$ in the second problem which contains only relatively few elements from the random sequence we have selected. (We may get good estimates if the sequence we consider has a length $n \ge \alpha K$ or we are looking for a set $A \in \mathcal{F}$ containing at least αK points of our sequence. Here $\alpha > 0$ is a universal constant, and K is the exponent of the Vapnik–Červonenkis class.) The earlier results I know about do not give a good estimate in these problems, because they provide a sharp estimate in formula (2) of Theorem 1 only if the parameter ρ in it is relatively large.

In this paper I gave a good estimate in Theorem 1 if the parameter ρ in it is very small, namely if the inequality $\rho \le n^{-200}$ holds. I am also interested in the question what can be told if this condition is dropped. This is the subject of my paper [5]. I also discuss some examples in that paper which show that its estimates are sharp, and I compare them with the results of

some earlier works. The proofs in [5] are based on Theorem 1 of this work. But since the arguments of the two papers are essentially different I decided to handle them separately.

I wrote that the parameter ρ in Theorem 1 must be small. But actually the condition $\rho \le n^{-200}$ imposed on it provides a relatively large bound. In the proof of the generalized version of Theorem 1 in paper [5] I shall adapt one of the main ideas in the Vapnik–Červonenkis theory. I try to reduce the estimation of the probability we are interested in to the estimation of the probability of relatively few events that can be simply bounded. Relatively few events means in this context that their number is only a polynomially, and not e.g. an exponentially increasing function of the sample size n. I shall be able to carry out such a program, since the parameter ρ in Theorem 1 is bounded by a (negative) power of the sample size n. This is the reason why it is important that we can apply Theorem 1 with relatively large parameters ρ.

I finish this section by some comments on the terminology of this paper. I applied the notion of classes of functions with polynomially increasing covering numbers. I introduced the same notion in my work [4] under the name 'L_1-dense classes of functions'. I changed this name to make reference to the closely related notion of uniform covering numbers which was introduced in the literature at several places, see e.g. [1], [6] or [8]. I used the terminology Vapnik–Červonenkis classes in a non-standard way. Usually one calls a class $\mathcal{C}$ of subsets of a set X a Vapnik–Červonenkis class of dimension d if d is the largest integer such that for all subsets D of X containing d elements its intersections with the sets of $\mathcal{C}$ contain all subsets of D. (See e.g. [1].) An important combinatorial result called the Sauer lemma implies that a Vapnik–Červonenkis class of dimension d is a Vapnik–Červonenkis class with exponent $K = d$ and parameter $B = (\frac{e}{d})^d$ by our terminology. (On the other hand, if a class of sets is a Vapnik–Červonenkis class with exponent K and parameter B by our terminology, then it is a Vapnik–Červonenkis class of dimension $d < \bar{d}$, if $\bar{d}$ is such a number for which $B\bar{d}^K < 2^{\bar{d}}$.) Hence we may also speak of Vapnik–Červonenkis classes with given exponent and parameter. I prefer this terminology, because it expresses the most important property of Vapnik–Červonenkis classes, and it indicates the similarity of this notion with the notion of classes of functions with polynomially increasing covering numbers.

3. The proof of Theorem 1A.

Theorem 1A will be proved by means of induction with respect to the parameter k (appearing in the definition of the size N of the set X). The first result of this section, Lemma 3.1, formulates a result similar to Theorem 1A in the special case when the set X, where the functions f are defined contains relatively few points. We need it to start our induction procedure.

Lemma 3.1. *Let us fix a number ρ, $0 < \rho < 1$, and a set $X = \{x_1, \dots, x_{N_0}\}$, with $N_0 \le \frac{1}{8}\rho^{-3/2}$ points together with a class of functions $\mathcal{F}$ defined on X which satisfies the following weakened version of the property having polynomially increasing covering numbers with some parameter $D \ge 1$ and exponent $L \ge 1$. $\mathcal{N}(\varepsilon, \mathcal{F}, L_1(\mu)) \le D\varepsilon^{-L}$ for all $0 < \varepsilon \le 1$, where μ is the uniform distribution on X, and $\mathcal{N}(\varepsilon, \mathcal{F}, L_1(\mu))$ was introduced in the definition of uniform covering numbers. Let us also assume that $\int f(x)\, d\mu(x) \le \rho$ and $f(x) \ge 0$ for all $f \in F$ and $x \in X$. Let us consider an integer $p \ge 2L$, and the set $B_p = B_p(\mathcal{F}) \subset X^p$ introduced in formula (4) together with the uniform measure μ_p on the p-fold product X^p of the space X. The inequality*

$$\mu_p(B_p) \le D\rho^{p/4} \tag{6}$$

holds.

Proof of Lemma 3.1. Let us choose a set of functions $f_1, \dots, f_s$, $f_j \in \mathcal{F}$ for all $1 \le j \le s$, with cardinality $s \le D(2N_0)^L$ such that for all $f \in \mathcal{F}$ there is a function f_j, $1 \le j \le s$, for which the inequality $\int |f(x) - f_j(x)|\mu(\, dx) \le \frac{1}{2N_0}$ holds. If $\int |f(x) - f_j(x)|\mu(\, dx) \le \frac{1}{2N_0}$, then $|f(x) - f_j(x)| \le \frac{1}{2}$ for all $x \in X$. This follows from the inequality $\frac{1}{N_0}|f(x) - f_j(x)| \le \int |f(x) - f_j(x)|\mu(\, dx) \le \frac{1}{2N_0}$ for all $x \in X$. As a consequence, $\{x\colon\ f(x) = 1\} \subset \{x\colon\ f_j(x) \ge \frac{1}{2}\}$ for such a pair of functions f and f_j, and

$$\begin{aligned} B_p &= B_p(\mathcal{F}) = \bigcup_{f \in \mathcal{F}} B_p(f) \\ &\subset \bigcup_{j=1}^{s} \left\{ (x_{t_1}, \dots, x_{t_p})\colon f_j(x_{t_k}) \ge \frac{1}{2} \quad \text{for all } 1 \le k \le p \right\}. \end{aligned}$$

Besides, we have for each j, $1 \le j \le s$,

$$\begin{aligned} &\mu_p \left\{ (x_{t_1}, \dots, x_{t_p})\colon f_j(x_{t_k}) \ge \frac{1}{2} \quad \text{for all } 1 \le k \le p \right\} \\ &\qquad = \left(\mu \left\{ x_t\colon f_j(x_t) \ge \frac{1}{2} \right\} \right)^p \le (2\rho)^p. \end{aligned}$$

Hence the relations $p \ge 2L$ and $N_0 \le \frac{1}{8}\rho^{-3/2}$ imply that

$$\mu_p(B_p) \le s(2\rho)^p \le D(2N_0)^{p/2}(2\rho)^p \le D\rho^{p/4}.$$

□

In our inductive proof we also need a result presented in Lemma 3.2. It is a version of the following heuristic statement.

Let us consider a class of functions $\mathcal{F}$ on a finite set X with polynomially increasing covering numbers which consists of non-negative functions bounded by 1, and take the supremum of the integrals $\int f(x)\mu(\,dx)$ of all functions $f \in \mathcal{F}$ with respect to the uniform distribution μ on X. Let the cardinality of the set X be $2N$, where the number N is of the form $N = A2^k$ with some positive integers A and k, and let the above supremum of integrals be bounded by a number ρ_{k+1}. Then there is a number ρ_k slightly larger than ρ_{k+1} with the following property. For most subsets $Y \subset X$ with cardinality N the supremum of the integrals of the restrictions of the functions $f \in \mathcal{F}$ to the set Y with respect to the uniform distribution on Y can be bounded by ρ_k.

Lemma 3.2. *Let us define two sequences of numbers*

$$N_k = 2^k N_0, \quad \text{and} \quad \rho_k = \rho \prod_{j=0}^{k-1} \left(1 + \frac{3}{N_j^{1/8}}\right)^{-1}, \qquad k = 1, 2, \dots, \quad \rho_0 = \rho, \tag{7}$$

with the help of some starting numbers N_0 and ρ which satisfy the relations $\rho \le \min(\frac{1}{1000}, L^{-20})$ and $\frac{1}{16}\rho^{-3/2} < N_0 \le \frac{1}{8}\rho^{-3/2}$. Let us fix an integer $k \ge 0$, and a set $X = \{x_1, \dots, x_{2N_k}\}$ with $N_{k+1} = 2N_k = N_0 2^{k+1}$ elements, and consider a class of functions $\mathcal{F}$ defined on the set X which has polynomially increasing covering numbers with parameter $D \ge 1$ and exponent $L \ge 1$ and satisfies the inequality $0 \le f(x) \le 1$ for all points $x \in X$ and functions $f \in \mathcal{F}$. Put $R_{k+1}(f) = \sum_{j=1}^{N_{k+1}} f(x_j)$, and assume that the class of functions $\mathcal{F}$ also satisfies the condition $R_{k+1}(f) \le N_{k+1}\rho_{k+1}$ for all $f \in \mathcal{F}$. Let us define the quantity $R_Y(f) = \sum_{x_j \in Y} f(x_j)$ for all functions $f \in \mathcal{F}$ and sets $Y \subset X$. The following Statement (a) holds.

(a) *The number of sets $Y \subset X$ such that $|Y| = N_k$, and $\sup_{f \in \mathcal{F}} R_Y(f) \ge N_k\rho_k$ is less than $\binom{2N_k}{N_k} D \exp\left\{-\frac{1}{100}2^{k/20}\rho^{-1/20}\right\}$.*

Proof of lemma 3.2. Let us fix a partition of $X = \{x_1, \dots, x_{2N_k}\}$ to two point subsets $\{x_{j_1}, x_{j_2}\}, \dots, \{x_{j_{2N_k-1}}, x_{j_{2N_k}}\}$ together with a sequence of i.i.d. random variables $\varepsilon_1, \dots, \varepsilon_{N_k}$ with distribution $P(\varepsilon_l = 1) = P(\varepsilon_l = -1) = \frac{1}{2}$ for all $1 \le l \le N_k$. Let us define with their help the 'randomized sum'

$$U_k(f) = \sum_{l=1}^{N_k} \varepsilon_l \left(f(x_{j_{2l-1}}) - f(x_{j_{2l}}) \right) \tag{8}$$

for all $f \in \mathcal{F}$.

Let us observe that for all $f \in \mathcal{F}$ the inequality

$$P(U_k(f) > 2z) \le \exp\left\{ -\frac{2z^2}{\sum\limits_{l=1}^{N_k} (f(x_{j_{2l-1}}) - f(x_{j_{2l}}))^2} \right\} \le e^{-z^2/2N_k\rho_{k+1}} \tag{9}$$

holds for all $z > 0$ by the Hoeffding inequality (see e.g. [7] Appendix B) and the inequality

$$\sum_{l=1}^{N_k} (f(x_{j_{2l-1}}) - f(x_{j_{2l}}))^2 \le 2 \sum_{j=1}^{2N_k} f(x_j)^2 \le 2R_{k+1}(f) \le 4N_k\rho_{k+1}. \tag{10}$$

(In formula (10) we exploit the condition $0 \le f(x) \le 1$ which implies that $f(x_j)^2 \le f(x_j)$.)

Define the (random) set

$$V_k = V_k(\varepsilon_1, \dots, \varepsilon_{N_k}) = \bigcup_{l:\, \varepsilon_l = 1} \{x_{j_{2l-1}}\} \cup \bigcup_{l:\, \varepsilon_l = -1} \{x_{j_{2l}}\}.$$

With such a notation we can write

$$\left\{ \omega\colon \sum_{s \in V_k(\varepsilon_1(\omega), \dots, \varepsilon_{N_k}(\omega))} f(x_s) > N_k\rho_{k+1} + z \right\}$$

$$\subset \left\{ \omega\colon \sum_{s \in V_k(\varepsilon_1(\omega), \dots, \varepsilon_{N_k}(\omega))} f(x_s) > \frac{R_{k+1}(f)}{2} + z \right\} = \{\omega\colon\, U_k(f)(\omega) > 2z\}.$$

Hence

$$P\left(\left\{\omega\colon \sum_{s\in V_k(\varepsilon_1(\omega),\dots,\varepsilon_{N_k}(\omega))} f(x_s) > N_k\rho_{k+1} + z\right\}\right) \le e^{-z^2/2N_k\rho_{k+1}} \qquad (11)$$

for all $z > 0$ by relation (9).

I claim that relation (11) implies the following Statement (b).

(b) For all $f \in \mathcal{F}$ and $z > 0$ the number of sets $V \subset X$ such that $|V| = N_k$, and $\sum_{x\in V} f(x) \ge N_k\rho_{k+1} + z$ is less than or equal to $\binom{2N_k}{N_k} e^{-z^2/2N_k\rho_{k+1}}$.

Indeed, it follows from relation (11) that for a fixed partition of the set X to two point subsets the number of those subsets $V \subset X$ which contain exactly one point from each element of this partition, (and as a consequence contain exactly N_k points), and $\sum_{s\in V} f(x_s) > N_k\rho_{k+1} + z$ is less than or equal to $2^{N_k} e^{-z^2/2N_k\rho_{k+1}}$. We get an upper bound for the quantity considered in statement (b) by summing up the number of sets V with these properties for all partitions of X to two point subsets, and taking into account how many times we counted each set V in this procedure. The number of the partitions of X to two point subsets equals $(2N_k-1)(2N_k-3)\cdots 3\cdot 1 = \frac{(2N_k)!}{2^{N_k}N_k!}$, and each partition provides at most $2^{N_k} e^{-z^2/2N\rho_{k+1}}$ sets V with the desired properties. All sets V were counted $N_k!$-times in this calculation. (The multiciplicity by which a set V, $|V| = N_k$, was counted in the above calculation agrees with the number of those partitions of X to two point subsets which have the property that all of their elements contain a fixed element of V.) These considerations imply Statement (b).

Given a number $0 \le u < 1$ there exist $s \le Du^{-L}$ functions $f_1, \dots, f_s$ in $\mathcal{F}$ such that for all $f \in \mathcal{F}$ and sets $Y \subset X$ one of the functions f_j, $1 \le j \le s$, satisfies the inequality $\sum_{x\in Y} |f_j(x) - f(x)| \le \sum_{x\in X} |f_j(x) - f(x)| \le uN_{k+1}$. We get this relation by exploiting that $\mathcal{F}$ is a class of functions which has polynomially increasing covering numbers with parameter D and exponent L if we apply inequality (1) to estimate $\mathcal{N}(\varepsilon, \mathcal{F}, L_1(\mu))$ with the uniform distribution μ on X instead of $\sup_\nu \mathcal{N}(\varepsilon, \mathcal{F}, L_1(\nu))$. This has the consequence that if $\sum_{x\in Y} f(x) \ge N_k\rho_{k+1} + z + 2uN_k$ for some $Y \subset X$ and $f \in \mathcal{F}$,

then there exists some index $1 \le j \le s$ such that $\sum_{x \in Y} f_j(x) \ge N_k \rho_{k+1} + z$ with the same set $Y \subset X$. Hence Statement (b) implies that the number of sets Y such that $|Y| = N_k$ and $\sum_{x \in Y} f(x) \ge N_k \rho_{k+1} + z + 2uN_k$ with some $f \in \mathcal{F}$ is less than or equal to $s \cdot e^{-z^2/2N_k\rho_{k+1}} \binom{2N_k}{N_k} = Du^{-L} e^{-z^2/2N_k\rho_{k+1}} \binom{2N_k}{N_k}$.

Put $z = N_k \rho_{k+1} \cdot N_k^{-1/8}$ and $u = \frac{z}{N_k}$. With such a choice we get that the number of sets $Y \subset X$ such that $|Y| = N_k$ and $\sup_{f \in \mathcal{F}} R_Y(f) \ge N_k \rho_{k+1}(1 + 3N_k^{-1/8}) = N_k \rho_k$ is less than

$$D \left(\frac{N_k^{1/8}}{\rho_{k+1}} \right)^L e^{-N_k^{3/4} \rho_{k+1}/2} \binom{2N_k}{N_k} = \binom{2N_k}{N_k} D \left(\frac{2^{k/8} N_0^{1/8}}{\rho_{k+1}} \right)^L e^{-2^{3k/4} N_0^{3/4} \rho_{k+1}/2}. \tag{12}$$

It follows from the definition of ρ_k that $\frac{1}{2}\rho \le \rho_{k+1} \le \rho$, and we also have $L \le \rho^{-1/20}$ because of the condition imposed on the number ρ. These relations together with the condition $\frac{1}{16}\rho^{-3/2} < N_0 \le \frac{1}{8}\rho^{-3/2}$ of Lemma 3.2 enable us to bound the expression in (12) from above by

$$\begin{aligned} &\binom{2N_k}{N_k} D \left(C_1 2^{k/8} \rho^{-19/16} \right)^{\rho^{-1/20}} e^{-C_2 2^{3k/4} \rho^{-1/8}} \\ &\quad \le \binom{2N_k}{N_k} D \exp\left\{ -C_3 2^{k/20} \rho^{-1/20} \right\} \end{aligned}$$

with appropriate constants C_1, C_2 and C_3. One can choose e.g. $C_3 = \frac{1}{100}$, and this implies Statement (a). (In the last step of this estimation we have exploited that for a small number $\rho > 0$ and all positive integers k the term $e^{-C_2 2^{3k/4} \rho^{-1/8}}$ is much smaller than the reciprocal of $\left(C_1 2^{k/8} \rho^{-19/16} \right)^{\rho^{-1/20}}$ which is of order $\exp\left\{ -\text{const.}\, \rho^{-1/20} (k + \log \frac{1}{\rho}) \right\}$.) □

Remark. It may be worth remarking that the most important part of Lemma 3.2, relation (9) or its consequence (11) can be considered as a weakened version of Lemma 3 in [3], and even its proof is based on the ideas worked out there. In formula (9) a random sum denoted by $U_k(f)$ was estimated by means of the Hoeffding inequality. To get this estimate we had to bound the variance of the random variable $U_k(f)$, and this was done in formula (10). In Lemma 3 of [3] a similar random sum was investigated, but in that case a good asymptotic formula and not only an upper bound was proved for the tail

distribution of the random sum. In the proof of that result a sharp version of the central limit theorem was applied instead of the Hoeffding inequality, and we needed a good asymptotic formula and not only a good upper bound for the variance of the random sum we investigated. The proof of the good asymptotic formula for this variance was the hardest part in the proof of Lemma 3 in [3].

Proof of Theorem 1A. Let us fix some numbers N_0, ρ and L which satisfy the conditions of Lemma 3.2. Take an integer $k \geq 0$, define the numbers N_k and ρ_k by formula (7), consider a space $X = \{x_1, \dots, x_{N_k}\}$ with N_k elements and a class of functions $\mathcal{F}$ on it which has polynomially increasing covering numbers with parameter $D \geq 1$ and exponent $L \geq 1$, and it has the properties that if $f \in \mathcal{F}$, then $0 \leq f(x) \leq 1$ for all $x \in X$, and $\int f(x)\mu(\, dx) \leq \rho_k$ with the uniform distribution μ on X. Fix an integer p such that $p \geq 2L$, $p \leq \rho^{-1/100}$, and let us also consider the sets $B_p(f)$, $f \in \mathcal{F}$, and $B_p = B_p(\mathcal{F})$ introduced in formulas (3) and (4). They consist of sequences $x^{(p)} = (x_{s_1}, \dots, x_{s_p}) \in X^p$ with some nice properties. Let $V(p, \rho, N_0, k) = V_{D,L}(p, \rho, N_0, k)$ denote the supremum of the cardinality of the sets $B_p(\mathcal{F})$ if the supremum is taken for all possible sets X and class of functions $\mathcal{F}$ with the above properties (with parameters N_k and ρ_k).

I claim that

$$V(p, \rho, N_0, k) \leq C_k D N_k^p \rho^{p/4} \quad \text{for all } k = 0, 1, 2, \dots \tag{13}$$

with

$$C_k = \prod_{j=0}^{k} (1 + 2^{-j}\rho). \tag{14}$$

Relation (13) will be proved by means of induction with respect to k. Its validity for $k = 0$ follows from Lemma 3.1. Let us assume that it holds for some k, take a set X with cardinality $N_{k+1} = 2N_k$ together with a class of functions $\mathcal{F}$ which satisfies the above conditions with the parameters D, L, p, ρ_{k+1} and N_{k+1}, and let us give a good bound on the cardinality of the set $B_p(\mathcal{F})$ defined in (3) and (4) in this case. To calculate the number of sequences $x^{(p)} = (x_{s_1}, \dots, x_{s_p}) \in X^p$ which belong to the set $B_p(\mathcal{F})$ let us take all sets $Y \subset X$ with cardinality $|Y| = N_k$, let us bound the number of those sequences $x^{(p)} \in B_p(\mathcal{F})$ for which also the property $x^{(p)} \in Y^p$ holds, and let us sum up these numbers for all sets $Y \subset X$ such that $|Y| = N_k$. Then take into account how many times we counted a sequence $x^{(p)}$ in this

summation. I claim that we get the following estimate in such a way:

$$|B_p(\mathcal{F})| \le DN_k^p \frac{\binom{2N_k}{N_k}}{\binom{2N_k-p}{N_k-p}} \left(C_k \rho^{p/4} + \exp\left\{ -\frac{1}{100} 2^{k/20} \rho^{-1/20} \right\} \right) \tag{15}$$

with the coefficient C_k defined in (14).

To prove relation (15) let us first observe that if $\mathcal{F}$ is a class of functions on the set X which has polynomially increasing covering numbers with parameter D and exponent L, and we restrict the domain where the functions of $\mathcal{F}$ are defined to a smaller set $Y \subset X$ then the class of functions we obtain in such a way remains a class of functions which has polynomially increasing covering numbers with the same parameter D and exponent L. Hence if we fix a set Y with cardinality $|Y| = N_k$ for which the property $\sup_{f \in \mathcal{F}} R_Y(f) \le N_k \rho_k$ holds (with the quantity $R_Y(f)$ introduced in the formulation of Lemma 3.2), then the number of those sequences $x^{(p)}$ for which $x^{(p)} \in B_p(\mathcal{F}) \cap Y^p$ can be bounded by our induction hypothesis by $C_k N_k^p D \rho^{p/4}$. We shall bound the number of the sequences $x^{(p)} \in \mathcal{B}_p(\mathcal{F}) \cap Y^p$ for the remaining sets Y with cardinality $|Y| = N_k$ by the trivial upper bound N_k^p, but the number of such sets Y is less than $\binom{2N_k}{N_k} D \exp\left\{ -\frac{1}{100} 2^{k/20} \rho^{-1/20} \right\}$ by Lemma 3.2. This yields the upper bound $C_k N_k^p D \rho^{p/4} \binom{2N_k}{N_k} + N_k^p \binom{2N_k}{N_k} D \exp\left\{ -\frac{1}{100} 2^{k/20} \rho^{-1/20} \right\}$ for the sum we get by summing up the number of sequences $x^{(p)} \in Y^p \cap B_p(\mathcal{F})$ for all subsets with $|Y| = N_k$ elements. To prove (15) we still have to take into account how many times we counted the sequences $x^{(p)} \in B_p(\mathcal{F})$ in this summation. If all coordinates of a sequence $x^{(p)} \in B_p(\mathcal{F})$ are different, then we counted it $\binom{2N_k-p}{N_k-p}$-times, because to find a set Y, $|Y| = N_k$, containing the elements of this sequence $x^{(p)}$ we have to extend these points with $N_k - p$ new points from the remaining $2N_k - p$ points of X. If some coordinates of a sequence $x^{(p)}$ may agree, then we might have counted this sequence with greater multiplicity. The above considerations imply (15).

To prove relation (13) with the help of (15) let us observe that under the conditions of Theorem 1A (in particular, we have $\frac{1}{N_0} \le 16\rho^{3/2}$, $p^2 \le \rho^{-1/50} \le \frac{1}{16}\rho^{-1/6}$, $2(N_k - p) \ge N_k = 2^k N_0$ for all $k = 0, 1, 2, \dots$ with a sufficiently small $\rho > 0$)

$$N_k^p \frac{\binom{2N_k}{N_k}}{\binom{2N_k-p}{N_k-p}} = N_k^p \frac{\binom{2N_k}{N_k}}{\binom{2N_k-p}{N_k}} = N_k^p \frac{2N_k(2N_k-1)\cdots(2N_k-p+1)}{N_k(N_k-1)\cdots(N_k-p+1)}$$

$$= N_{k+1}^p \left(1 + \frac{1}{2(N_k - 1)}\right)\left(1 + \frac{2}{2(N_k - 2)}\right)\cdots\left(1 + \frac{p-1}{2(N_k - p + 1)}\right)$$
$$\le N_{k+1}^p \exp\left\{\frac{p^2}{2^{k+1}N_0}\right\} \le N_{k+1}^p e^{2^{-(k+1)}\rho^{4/3}} \le N_{k+1}^p \left(1 + \frac{1}{3}2^{-(k+1)}\rho\right),$$

and

$$\exp\left\{-\frac{1}{100}2^{k/20}\rho^{-1/20}\right\} = \rho^{p/4}\exp\left\{-\frac{1}{100}2^{k/20}\rho^{-1/20} + \frac{p}{4}\log\frac{1}{\rho}\right\}$$
$$\le C_k\rho^{p/4}\cdot\frac{1}{3}2^{-(k+1)}\rho$$

with the coefficient C_k defined in (14). These estimates together with (15) imply (13) for parameter $k+1$.

It is not difficult to prove Theorem 1A with the help of relation (13). To do this let us observe that $\rho_k \ge \frac{\rho}{2}$ and $C_k \le 2$ for all $k = 0, 1, 2, \ldots$. Hence taking a class of functions $\mathcal{F}$ on a set X with cardinality N_k with some $k \ge 0$ which satisfies the conditions of Theorem 1A we can write (by exploiting that $\int f(x)\mu(\,dx) \le \frac{\rho}{2} \le \rho_k$) the estimate

$$\mu_p(B_p(\mathcal{F})) = N_k^{-p}|B_p(\mathcal{F})| \le N_k^{-p}V(\rho, p, N_0, k) \le 2D\rho^{p/4}$$

by relation (13). □

4. The proof of Theorem 1.

First we prove a special case of Theorem 1 in Lemma 4.1. Here we take a class of functions $\mathcal{F}$ on a finite set X of cardinality 2^k with some integer k, and μ is the uniform distribution on X.

Lemma 4.1. *Let us consider a finite set $X = \{x_1, \ldots, x_{2^k}\}$ with $N = 2^k$ elements together with a class of functions $\mathcal{F}$ on X which has monotone increasing covering numbers with some parameter $D \ge 1$ and exponent $L \ge 1$, and it contains functions $f \in \mathcal{F}$ with the properties $0 \le f(x) \le 1$ for all $x \in X$ and $\int f(x)\mu(\,dx) \le \rho$ with some $0 < \rho < 1$. Here μ denotes the uniform distribution on X. Let us take the n-fold direct product X^n of X with some number $n \ge 2$, and define the function $S_n(f)(x_{s_1}, \ldots, x_{s_n}) = \sum_{j=1}^{n} f(x_{s_j})$ for all $(x_{s_1}, \ldots, x_{s_n}) \in X^n$ and $f \in \mathcal{F}$. Let us assume that $\rho \le 2n^{-200}$, and*

$N = 2^k \geq \rho^{-2}$. *Then the set* $B_n(u) \subset X^n$ *defined as*

$$B_n(u) = \left\{ (x_{s_1}, \dots, x_{s_n}) \colon \sup_{f \in \mathcal{F}} S_n(f)(x_{s_1}, \dots, x_{s_n}) > u \right\} \tag{16}$$

satisfies the inequality

$$\mu_n(B_n(u)) \leq 2D\rho^{u/25} \quad \textit{for all } u \geq 40L, \tag{17}$$

where μ_n *denotes the uniform distribution on* X^n.

Proof of Lemma 4.1 Let us define for all functions $f \in \mathcal{F}$ and integers j, $1 \leq j \leq R$, where R is defined by the relation $n < 2^R \leq 2n$, the functions $f^{(j)}(x) = \min(2^{-j}, f(x))$ and $\bar{f}^{(j)}(x) = 2^j f^{(j)}(x)$, $x \in X$. Put $\mathcal{F}_j = \{f^{(j)} \colon f \in \mathcal{F}\}$ and $\bar{\mathcal{F}}_j = \{\bar{f}^{(j)} \colon f \in \mathcal{F}\}$. One can simply check that if $\mathcal{F}$ is a class of functions which has polynomially increasing covering numbers with parameter D and exponent L, then $\mathcal{F}_j$ is a class of functions which has polynomially increasing covering numbers with parameter D and exponent L, while $\bar{\mathcal{F}}_j$ is a class of functions which has polynomially increasing covering numbers with parameter $D2^{jL}$ and exponent L. We can also state that $\int f^{(j)}(x)\mu(\,dx) \leq \rho$, and $\int \bar{f}^{(j)}(x)\mu(\,dx) \leq 2^j\rho$ for all $f \in \mathcal{F}$.

Let us define for all $f \in \mathcal{F}$ and $1 \leq j \leq R$ the following function $H_j(f)$ on X^n:

$$H_j(f)(x_{s_1}, \dots, x_{s_n}) = \text{the number of such indices } l \text{ for which } \bar{f}^{(j)}(x_{s_l}) = 1.$$

We can write

$$S_n(f)(x_{s_1}, \dots, x_{s_n}) \leq \sum_{j=1}^{R} 2^{1-j} H_j(f)(x_{s_1}, \dots, x_{s_n}) + 1$$

for all $f \in \mathcal{F}$. This formula implies the inequality

$$\sup_{f \in \mathcal{F}} S_n(f)(x_{s_1}, \dots, x_{s_n}) \leq \sum_{j=1}^{R} 2^{1-j} \sup_{f \in \mathcal{F}} H_j(f)(x_{s_1}, \dots, x_{s_n}) + 1,$$

and the relation

$$\left\{ (x_{s_1}, \dots, x_{s_n}) \colon \sup_{f \in \mathcal{F}} S_n(f)(x_{s_1}, \dots, x_{s_n}) > u \right\}$$
$$\subset \bigcup_{j=1}^{R} \left\{ (x_{s_1}, \dots, x_{s_n}) \colon \frac{\sup_{f \in \mathcal{F}} H_j(f)(x_{s_1}, \dots, x_{s_n})}{2^{j-1}} > \frac{(u-1)}{\sqrt{2}-1} 2^{-j/2} \right\}.$$

Hence

$$\mu_n(B_n(u)) \le \sum_{j=1}^{R} \mu_n(D_n(u,j)) \tag{18}$$

for the sets $B_n(u)$ defined in (16) and

$$D_n(u,j) = \left\{ (x_{s_1},\dots,x_{s_n})\colon \sup_{f\in\mathcal{F}} H_j(f)(x_{s_1},\dots,x_{s_n}) > \frac{\sqrt{2}-1}{2}(u-1)2^{j/2} \right\},$$
$$, \qquad 1 \le j \le R.$$

We can prove Lemma 4.1 with the help of relation (18) if we give a good estimate on the measures $\mu_n(D_n(u,j))$. This can be done with the help of Theorem 1A.

Indeed, the set $D_n(u,j)$ consists of such sequences $(x_{s_1},\dots,x_{s_n}) \in X^n$ which have a subsequence $(x_{s_{p_1}},\dots,x_{s_{p_t}})$ with $t = t(j) = [\frac{\sqrt{2}-1}{2}(u-1)2^{j/2}]+1$ elements, where $[\cdot]$ denotes integer part, with the property that there is a function $f \in \mathcal{F}$ such that the function $\bar{f}_j(\cdot)$ defined with its help equals 1 in all coordinates of this subsequence. More explicitly,

$$D_n(u,j) = \bigcup_{\{l_1,\dots,l_t\}\subset\{1,\dots,n\}} \left(\bigcup_{f\in\mathcal{F}} \{(x_{s_1},\dots,x_{s_n})\colon \bar{f}^{(j)}(x_{s_{l_1}}) = 1,\dots,\bar{f}^{(j)}(x_{s_{l_t}}) = 1\} \right) \tag{19}$$

with $t = t(j) = [\frac{\sqrt{2}-1}{2}(u-1)2^{j/2}] + 1$.

The outside union in (19) consists of $\binom{n}{t(j)} \le n^{t(j)}$ terms, and the cardinality of the sequences $(x_1,\dots,x_n)$ in the inner union can be bounded by means of Theorem 1A for each term if it is applied with $p = t(j)$, in the space X consisting of $N = 2^k = N_0 2^{\bar{k}}$ points, for the class of functions $\bar{\mathcal{F}}_j$ which is a class of functions which has polynomially increasing covering numbers with parameter $D2^{jL}$ and exponent L. Moreover, the functions $\bar{f}^{(j)} \in \bar{\mathcal{F}}_j$ satisfy the inequality $\int \bar{f}^{(j)}(x)\mu(\,dx) \le 2^j\rho$. This means that under the conditions of Lemma 4.1 we can apply Theorem 1A for the class of functions $\bar{\mathcal{F}}_j$ with parameter $\bar{\rho} = \bar{\rho}_j = 2^{j+1}\rho$ instead of ρ. (We have to check that all conditions of Theorem 1A hold with $\rho = \bar{\rho}$ and $p = t(j)$. In particular, we can state that $\bar{\rho} = 2^{j+1}\rho \le L^{-20}$, since $\rho \le 2n^{-200}$, $2^j \le 2n$, and since we estimate the probability in formula (17) only under the condition $u \ge 40L$, and this probability is zero if $u > n$. Hence we may assume that $40L \le u \le n$. We chose the term N_0 in the application of Theorem 1A as $N_0 = 2^{k_0}$ with k_0

defined by the relation $\frac{1}{16}\bar{\rho}^{-3/2} < 2^{k_0} \le \frac{1}{8}\bar{\rho}^{-3/2}$, and $\bar{k} = k - k_0$. Observe that $2^{k_0} \le \frac{1}{8}\bar{\rho}^{-3/2} \le \rho^{-2} \le N$.)

We will prove with the help of the above relations the inequality

$$\begin{aligned} \mu_n(D_n(u,j)) &= \frac{|D_n(u,j)|}{N^n} \le 2n^{t(j)} D 2^{jL} (2^{j+1}\rho)^{t(j)/4} \\ &\le 2D(8n^5\rho)^{t(j)/4} \le 2D\rho^{t(j)/5} \le D\rho^{ju/25}. \end{aligned} \tag{20}$$

To get the first estimate in the second line of formula (20) observe that under the condition of Lemma 4.1 $\frac{\sqrt{2}-1}{2}(u-1) \ge 4L$, hence $2^{jL} \le 2^{j2^{-j/2}t(j)/4} \le 2^{t(j)/4}$, and by the definition of the number R we have

$$(2^{j+1})^{t(j)/4} \le (2^{R+1})^{t(j)/4} \le (4n)^{t(j)/4}.$$

We imposed the condition $n \le (\frac{2}{\rho})^{1/200}$, and this implies the second inequality. Finally $t(j) \ge \frac{ju}{5}$. (In the last inequality a $j = 1$ parameter is the worst case.) Relation (17) follows from (18) and (20). □

Now we turn to the proof of the main result of this paper.

Proof of Theorem 1. We may assume that all functions $f \in \mathcal{F}$ are non-negative, i.e. $0 \le f(x) \le 1$ for all $f \in \mathcal{F}$ and $x \in X$, because we can replace the function f by its absolute value $|f|$, and apply the result for this new class of functions which also satisfies the conditions of Theorem 1. Next I show that we also may assume that the class of functions $\mathcal{F}$ contains only finitely many functions, and it satisfies the same conditions as the original class of function $\mathcal{F}$ with the only difference that we assume that $\mathcal{F}$ is a class of functions which has polynomially increasing covering numbers with the same exponent L but with parameter $2^L D$ instead of D.

Indeed, if a number is an upper bound for the probability $P\left(\sup_{f\in\mathcal{F}'} S_n(f) > u\right)$ for all finite subsets $\mathcal{F}' \subset \mathcal{F}$, then it is also an upper bound for $P\left(\sup_{f\in\mathcal{F}} S_n(f) > u\right)$. Besides, the conditions of Theorem 1 remain valid if $\mathcal{F}$ is replaced by an arbitrary class of functions $\mathcal{F}' \subset \mathcal{F}$ with a small modification. Namely, we can state that $\mathcal{F}'$ is a class of functions which has polynomially increasing covering numbers with exponent L but with a possibly different parameter $\bar{D} = 2^L D$. (We had to change the parameter D of a class of functions $\mathcal{F}' \subset \mathcal{F}$ with polynomially increasing covering numbers, because although we can choose a sequence of functions $f_1, \dots, f_m$

with $m \le D\varepsilon^{-L}$ elements which is an ε-dense set in $\mathcal{F}$ with respect to the $L_1(\nu)$ norm with a probability measure ν, and $f_j \in \mathcal{F}$, but these functions may be not contained in $\mathcal{F}'$. To overcome this difficulty we choose a sequence $f_1, \dots, f_m$ with $m \le 2^L D\varepsilon^{-L}$ elements such that $\min\limits_{1 \le j \le m} \int |f - f_j|\, d\nu \le \frac{\varepsilon}{2}$ for all functions $f \in \mathcal{F}'$. We may also assume that for all these functions f_j there is a function $f \in \mathcal{F}'$ whose distance from f_j in the $L_1(\nu)$ norm is less than or equal to $\frac{\varepsilon}{2}$, since we can drop those functions f_j which do not have this property. Then we replace those functions f_j for which $f_j \notin \mathcal{F}'$ by a function $f \in \mathcal{F}'$ such that $\int |f - f_j|\, d\nu \le \frac{\varepsilon}{2}$. In such a way we get an ε-dense subclass of $\mathcal{F}'$ with $m \le 2^L D\varepsilon^{-L}$ elements in the $L_1(\nu)$ norm.)

In the next step I show that we may restrict our attention to the case when the functions of the class of functions $\mathcal{F}$ (consisting of finitely many functions) take only finitely many values. For this goal first I split up the interval $[0,1]$ into n subintervals of the following form: $B_j = (\frac{j-1}{n}, \frac{j}{n}]$, $2 \le j \le n$, and $B_1 = [0, \frac{1}{n}]$. (We defined the function B_1 in a slightly different way in order to guarantee that the point zero is also contained in some set B_j.) Then given a class of function $\mathcal{F}$ on a set X that contains finitely many functions $f_1, \dots, f_R$ we define the following sets $A(s(1), \dots, s(R)) \subset X$ (depending on $\mathcal{F}$):

$$A(s(1), \dots, s(R)) = \{x\colon\ f_j(x) \in B_{s(j)}, \quad \text{for all } 1 \le j \le R\},$$

where $1 \le s(j) \le n$ for all $1 \le j \le R$.

In such a way the sets $A(s(1), \dots, s(R))$ make up a partition of the set X. Actually, for the sake of a simpler argument we shall diminish a bit the set X, by defining it as the union of those sets $A(s(1), \dots, s(R))$ for which $\mu(A(s(1), \dots, s(R))) > 0$ with the measure μ appearing in Theorem 1. This restriction will cause no problem in our later considerations.

We shall define new functions $\tilde{f}_j(x)$, $1 \le j \le R$, by means of the partition of X to the sets $A(s(1), \dots, s(R))$ by the formula

$$\tilde{f}_j(x) = \frac{\int_{A(s(1),\dots,s(R))} f_j(x)\mu(\,dx)}{\mu(A(s(1), \dots, s(R)))}, \quad 1 \le j \le R, \quad \text{if } x \in A(s(1), \dots, s(R)).$$

We have $|f_j(x) - \tilde{f}_j(x)| \le \frac{1}{n}$ for all $1 \le j \le n$ and $x \in X$. Hence

$$\left| \sup_{1 \le j \le R} (S_n(f_j) - S_n(\tilde{f}_j)) \right| \le 1,$$

for almost all sequences $\xi_1(\omega), \dots, \xi_n(\omega)$, and as a consequence

$$P\left(\sup_{1\le j\le R} S_n(f_j) > u+1\right) \le P\left(\sup_{1\le j\le R} S_n(\tilde{f}_j) > u\right) \tag{21}$$

Let us also observe that the class of functions $\tilde{\mathcal{F}} = \{\tilde{f}_j,\ 1 \le j \le R\}$ also satisfies the conditions of Theorem 1, i.e. $\int \tilde{f}_j(x)\mu(\,dx) \le \rho$ for all $1 \le j \le R$, and $\tilde{\mathcal{F}}$ is a class of functions which has polynomially increasing covering numbers with parameter $\bar{D} = 2^L D$ and exponent L. (The conditions on the numbers n and ρ clearly remain valid.)

The first relation follows from the identity $\int \tilde{f}_j(x)\mu(\,dx) = \int f_j(x)\mu(\,dx)$ which holds because of the identities

$$\int_{A(s(1),\dots,s(R))} \tilde{f}_j(x)\mu(\,dx) = \int_{A(s(1),\dots,s(R))} f_j(x)\mu(\,dx)$$

for all sets $A(s(1), \dots, s(R))$.

To prove that $\tilde{\mathcal{F}}$ has polynomially increasing covering numbers with parameter $2^L D$ and exponent L let us introduce for all probability measures ν the probability measure $\tilde{\nu} = \tilde{\nu}(\nu)$ which is defined by the property that for all (measurable) sets $A(s(1), \dots, s(R))$ and $B \subset A(s(1), \dots, s(R))$ the identity $\tilde{\nu}(B) = \mu(B)\frac{\nu(A(s(1),\dots,s(R)))}{\mu(A(s(1),\dots,s(R)))}$ holds. Because of the special form of the functions $\tilde{f}_j$ if a set of functions $\tilde{\mathcal{F}}_{\varepsilon,\tilde{\nu}} \subset \tilde{\mathcal{F}}$ is an ε-dense subset of $\tilde{\mathcal{F}}$ in the $L_1(\tilde{\nu})$ norm, then it is also a ε-dense subset of $\tilde{\mathcal{F}}$ in the $L_1(\nu)$ norm. (In the proof of this statement we exploit that

$$\tilde{\nu}(A(s(1), \dots, s(R))) = \nu(A(s(1), \dots, s(R)))$$

for all sets $A(s(1), \dots, s(R))$, and it depends only on the value of a measure ν on the sets $A(s(1), \dots, s(R))$ whether a set of functions $\{f_1, \dots, f_m\} \subset \tilde{\mathcal{F}}$ is an ε-dense subset of $\tilde{\mathcal{F}}$ in the $L_1(\nu)$ norm.)

Hence it is enough to prove the existence of an ε-dense subset $\tilde{\mathcal{F}}_{\varepsilon,\tilde{\nu}}$ of $\tilde{\mathcal{F}}$ in the $L_1(\tilde{\nu})$ norm with cardinality bounded by $\bar{D}\varepsilon^{-L}$ only with respect to such measures $\tilde{\nu}$ which can be written in the form $\tilde{\nu} = \tilde{\nu}(\nu)$ with some probability measure ν. In this case the relation we want to check follows from the fact that the original class of functions $\mathcal{F}$ has polynomially increasing covering numbers with parameter $2^L D$ and exponent L, (we apply this property for the measure $\tilde{\nu}(\nu)$) and the inequality

$$\int |\tilde{f}_j - \tilde{f}_{j'}| d\tilde{\nu}(\nu) \le \int |f_j - f_{j'}| d\tilde{\nu}(\nu)$$

holds for all pairs $f_j, f_{j'} \in \mathcal{F}$ and probability measure ν. This inequality holds, since

$$\int_{A(s(1),\dots,s(R))} |\tilde{f}_j(x) - \tilde{f}_{j'}(x)|\, d\tilde{\nu}(\nu)(x) \le \int_{A(s(1),\dots,s(R))} |f_j(x) - f_{j'}(x)|\, d\tilde{\nu}(\nu(x))$$

for all sets $A(s(1), \dots, s(R))$.

I claim that for all $k \ge 1$ we can define such a 'discretized' probability measure $\bar{\mu}_k$ on the σ-algebra $\mathcal{X}_k$ with atoms $A(s(1), \dots, s(R))$ in the space X for which

$$|\bar{\mu}_k(A(s(1), \dots, s(R))) - \mu(A(s(1), \dots, s(R)))| \le 2^{-k}, \tag{22}$$

and the probability $\bar{\mu}_k$ has the property that for all sets $A(s(1), \dots, s(R))$

$$\bar{\mu}_k(A(s(1), \dots, s(R))) = \alpha(A(s(1), \dots, s(R)))2^{-k} \tag{23}$$

with a non-negative integer $\alpha(A(s(1), \dots, s(R)))$. (To find such a probability measure $\bar{\mu}_k$ let us list the sets $A(s(1), \dots, s(R))$ as $B_1, \dots, B_Q$, and define the measure $\bar{\mu}_k$ by the relation $\sum_{l=1}^{s} \bar{\mu}_k(B_l) = \beta_s 2^{-k}$ if $(\beta_s - 1)2^k < \sum_{l=1}^{s} \mu_k(B_l) \le \beta_s 2^{-k}$ with a positive integer β_s. We assume this relation for all $1 \le s \le Q$.)

Clearly,

$$P\left(\sup_{1\le j\le R} S_n(\tilde{f}_j) > u\right) = \lim_{k\to\infty} P_{\bar{\mu}_k}\left(\sup_{1\le j\le R} S_n(\tilde{f}_j) > u\right) \tag{24}$$

for all $u > 0$, where $P_{\bar{\mu}_k}$ means that we consider the probability of the same event as at the left-hand side of the identity, but this time we take i.i.d. random variables $\xi_1, \dots, \xi_n$ with distribution $\bar{\mu}_k$ (on the σ-algebra generated by the atoms $A(s(1), \dots, s(R))$) in the definition of the random variables $S_n(\tilde{f}_j)$.

We shall bound the probabilities at the right-hand side in formula (24) for all large indices k by means of Lemma 4.1. This will be done with the help of the following construction. Take a space $\hat{X} = \hat{X}_k = \{x_1, x_2, \dots, x_{2^k}\}$ with 2^k elements and with the uniform distribution $\mu = \mu^{(k)}$ on its points. Let us fix a partition of $\hat{X}$ consisting of some sets $\hat{A}(s(1), \dots, s(R))$ with $\alpha(A(s(1), \dots, s(R)))$ elements. The number $\alpha(\cdot)$ was introduced in (23). Let us define the functions $\hat{f}_j(x) = \hat{f}_j^{(k)}(x)$, $1 \le j \le R$, $x \in \hat{X}$, by the formula $\hat{f}_j(x) = \frac{s(j)}{n}$, $1 \le j \le R$, if $x \in \hat{A}(s(1), \dots, s(R))$. Take the n-fold

direct product $\hat{X}^n$ of $\hat{X}$ together with the uniform distribution $\mu_n = \mu_n^{(k)}$ on it and the functions $S_n(\hat{f}_j)(x_{t_1}, \dots, x_{t_n}) = \sum_{l=1}^{n} \hat{f}_j(x_{t_l})$, $1 \le j \le R$, if $(x_{t_1}, \dots, x_{t_n}) \in \hat{X}^n$ on the space $\hat{X}^n$. I claim that

$$P_{\bar{\mu}_k}\left(\sup_{1\le j\le R} S_n(\tilde{f}_j) > u\right) \tag{25}$$
$$= \mu_n^{(k)}\left(\left\{(x_{t_1}, \dots, x_{t_n}): \sup_{1\le j\le R} S_n(\hat{f}_j^{(k)})(x_{t_1}, \dots, x_{t_n}) > u\right\}\right) \le 2\bar{D}\rho^{u/25}$$

with $\bar{D} = 2^L D$ if $k \ge k_0(n, R)$ with some index $k_0(n, R)$ and $u > 40L$.

The identity in formula (25) holds, since the joint distribution of the random vectors $S_n(f_j)(\xi_1, \dots, \xi_n)$, $1 \le j \le R$, where $\xi_1, \dots, \xi_n$ are independent random variables with distribution $\bar{\mu}_k$ and of the random vectors $S_n(\hat{f}_j^{(k)})(x_{t_1}, \dots, x_{t_n})$, $1 \le j \le R$, where the distribution of $(x_{t_1}, \dots, x_{t_n}) \in \hat{X}^n$ is $\mu_n^{(k)}$, agree. To prove the last inequality of (25) it is enough to check that for all sufficiently large numbers k the class of functions $\hat{\mathcal{F}} = \{\hat{f}_1, \dots, \hat{f}_R\}$ on the space $\hat{X} = \hat{X}_k$ satisfies the conditions of Lemma 4.1. Namely, $\hat{\mathcal{F}}$ has polynomially increasing covering numbers with parameter $\bar{D} = 2^L D$ and exponent L, and $\int \hat{f}_j^{(k)}(x)\mu(\,dx) \le \bar{\rho}$ with a number $\bar{\rho} \le 2n^{-200}$ for all $\hat{f}_j \in \hat{\mathcal{F}}$.

We have to explain why $\hat{\mathcal{F}}$ is a class of functions on $\hat{X}$ with polynomially increasing covering numbers. Let us observe that since the functions $\hat{f}_j$ are constant on the sets $\hat{A}(s(1), \dots, s(R))$ we have to take into consideration only the values $\hat{\nu}(\hat{A}(s(1), \dots, s(R))$ of the measures $\hat{\nu}$ on the space $\hat{X}$ when we want to show that $\hat{\mathcal{F}}$ has polynomially increasing covering numbers with parameter $\bar{D} = 2^L D$ and exponent L.

To do it let us correspond to all measures $\hat{\nu}$ on $\hat{X}$ the measure $\tilde{\nu}$ on X defined by the relation $\tilde{\nu}(A(s(1), \dots, s(R))) = \hat{\nu}(\hat{A}(s(1), \dots, s(R)))$ for all sets $A(s(1), \dots, s(R))$. Then we get that if a class of functions $\tilde{\mathcal{F}}_{\varepsilon,\tilde{\nu}} = \{\tilde{f}_{l_1}, \dots, \tilde{f}_{l_s}\}$ is an is an ε-dense class of $\tilde{\mathcal{F}}$ with respect to the $L_1(\tilde{\nu})$ norm, then the class of function $\hat{\mathcal{F}}_{\varepsilon,\hat{\nu}} = \{\hat{f}_{l_1}, \dots, \hat{f}_{l_s}\}$ is an ε-dense class with respect to the $L_1(\hat{\nu})$ norm. This implies that $\hat{\mathcal{F}}$ has such polynomially covering numbers as we claimed.

On the other hand, $\int \hat{f}_j^{(k)}(x)\mu(\,dx) = \int \tilde{f}_j(x)\bar{\mu}_k(\,dx)$, and

$$\lim_{k\to\infty} \int \tilde{f}_j(x)\bar{\mu}_k(\,dx) = \int \tilde{f}_j(x)\mu(\,dx) \le \rho \le n^{-200}$$

because of formula (22), and this implies that $\int \hat{f}_j^{(k)}(x)\mu(\,dx) \le (1+\frac{\log 2}{n})\rho$ for large parameters k. Since under the conditions of Theorem 1 $\bar{\rho} = (1+\frac{\log 2}{n})\rho \le 2n^{-200}$ for large enough parameters k, hence we can apply Lemma 4.1 with this $\bar{\rho}$ in the estimation of the probabilities in formula (25). Since we can restrict our attention to the case $u \le n$ we have $\bar{\rho}^{u/25} \le 2\rho^{u/25}$ for large enough k and we get the inequality part of formula (25) from Lemma 4.1.

Under the conditions of Theorem 1 relations (25), (24), (21), and the reduction of the investigation of the supremum to the case when the set of test functions $\mathcal{F}$ is replaced by its finite subsets imply that

$$P\left(\sup_{f\in\mathcal{F}} S_n(f) > u+1\right) \le 2^{L+1}D\rho^{u/25} \quad \text{if } u \ge 40L.$$

Theorem 1 follows from this relation.

Indeed, we can rewrite what this inequality means if the number $u+1$ is replaced by the number u on its left-hand side. In such a way we get that under the conditions of Theorem 1 the probability of the left-hand side in formula (2) can be bounded by $2^{L+1}D\rho^{(u-1)/25} = \bar{C}D\rho^{u/50}$ with $\bar{C} = 2^{L+1}\rho^{-1/25+u/50}$. (We replaced the condition $u \ge 40L$ by $u \ge 41L$ in (2). This enabled us to work with u instead of $u+1$ in this estimate.) To complete the proof of Theorem 1, it is enough to show that $\bar{C} \le 1$ under its conditions. This inequality holds, since $\bar{C} \le n^{L+1}\rho^{u/100} \le n^{2L}\rho^{u/100}$, if $n \ge 2$, $u \ge 4$, and $L \ge 1$. If also the relations $\rho \le n^{-200}$, and $u \ge L$ hold, then $\bar{C} \le n^{2(L-u)} \le 1$. □

[1] R. M. Dudley, *Uniform Central Limit Theorems. Second Edition.* Cambridge University Press, Cambridge (2014)

[2] D. Haussler, Sphere packing numbers of subsets of Boolean n–cube with bounded Vapnik–Chervonenkis dimension. Journal of Combinatorial Theory **A 69**, 217–232 (1995)

[3] J. Komlós, P. Major, G. Tusnády, An approximation of partial sums of independent rv.'s and the sample DF. II Z. Wahrscheinlichkeitstheorie verw. Gebiete 34, 33–58 (1976)

[4] P. Major, On the estimation of multiple random integrals and U-statistics *Lecture Notes in Mathematics* vol 2079 (Springer) Heidelberg New York Dordrecht London (2013)

[5] P. Major, On the tail behaviour of the distribution function of the supremum of a class of sums of i.i.d. random variables. Submitted to Stochastic Processes and their Applications.

[6] D. Pollard, A central limit theorem for empirical processes. Austral Math. Soc. (Series A). 235–248 **33**, (1982)

[7] D. Pollard, *Convergence of Stochastic Processes.* (Springer, New York, (1984)

[8] A. W. van der Vaart, J. A. Wellner, *Weak Convergence and Empirical Processes.* Springer, New York (1996)