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

Péter Major

2015. október 7.

Formulation of the problem, motivation, and methods to solve it

The problem:

Let $\xi_1,\dots,\xi_n$ be a sequence of i.i.d. random variables with some distribution μ on a measurable space $(X,\mathcal{X})$.

Let a class of functions $\mathcal{F}$ be given on the space $(X,\mathcal{X})$ with some nice properties, such that $\int f(x)\mu(\,dx)=0$ and $\sup_{x\in X}|f(x)|\le 1$ for all elements $f\in\mathcal{F}$.

Define the normalized sums $S_n(f)=\frac{1}{\sqrt{n}}\sum_{j=1}^n f(\xi_j)$ for all $f\in\mathcal{F}$, and give a good estimate on the tail distribution

$$P\left(\sup_{f\in\mathcal{F}} S_n(f)>x\right) \quad \text{for all numbers } x>0$$

of the supremum of these sums.

An important remark: By an important result, called the concentration inequality, the distribution of this supremum is concentrated in a small neighbourhood of a concentration point. As a consequence, the above probability is small only if x is larger than this concentration point. An important and hard part of the problem is to find a good level above which this probability begins to decrease radically. This means a good estimate on value of the concentration point. This is the hardest part of the problem.

(We want to give an estimate on the concentration point with the accuracy of a universal multiplying constant.)

Our motivations to study this problem

Motivation 1.: Dudley's theory of uniform central limit theorem. Given a nice class of functions $\mathcal{F}$ and a sequence of i.i.d. random variables $\xi_1, \dots, \xi_n$ prove that the class of normalized sums $S_n(f) = \frac{1}{\sqrt{n}} \sum_{j=1}^n f(\xi_j)$, $f \in \mathcal{F}$, satisfy some sort of functional limit theorem.

Crucial point of the proof: Show that $\sup_{f,f' \in \mathcal{F}'} S_n(f - f')$ is small if $\mathcal{F}' \subset \mathcal{F}$, and $\mathcal{F}'$ has the property that $E[f(\xi_1) - f'(\xi_1)]^2 < \delta$ with a small $\delta > 0$ for all $f, f' \in \mathcal{F}'$. We have formulated a natural generalization of this problem, where we may consider $\delta_n \to 0$ as $n \to \infty$ instead of a fixed $\delta > 0$.

Motivation 2.: To get good limit theorems for so-called non-parametric maximum likelihood estimates it is useful to prove sharp estimates on the tail distribution of the supremum of multiple integrals with respect to normalized empirical distibution functions, i.e. of expressions of the form:

$$\sup_{f\in\mathcal{F}}\int\cdots\int f(x_1,\dots,x_k)$$
$$\sqrt{n}(dF_n(x_1)-dF(x_1))\dots\sqrt{n}(dF_n(x_k)-dF(x_k)),$$

where F is a distribution function F_n is the empirical distribution function of an F distributed sample, and $\mathcal{F}$ is a nice class of functions of k variables. (See my lecture note On the estimation of multiple random integrals and U-statistics.) Here we investigate this problem for $k=1$.

The Gaussian version of the problem

The sums $S_n(f) = \frac{1}{\sqrt{n}} \sum\limits_{j=1}^{n} f(\xi_j)$ satisfy the central limit theorem.

Hence it is natural to consider a Gaussian version of our problem. We want to understand what kind of results and methods suggests the result of this Gaussian version.

The Gaussian problem: Let η_t, $E\eta_t = 0$, $t \in T$, be a countable set of random variables with (jointly) Gaussian distribution. Put $d_2(s,t) = \left[E(\eta_s - \eta_t)^2\right]^{1/2}$, $s,t \in T$. Then $d_2(s,t)$ is a metric on the parameter set T. Give a good estimate on the probability $P\left(\sup_{t \in T} \eta_t > x\right)$ for all numbers $x > 0$ with the help of the function $d_2(s,t)$.

The result about the Gaussian problem

There is a simple and natural method, called the chaining argument to study this problem. It yields the following result.

Theorem. *Let η_t, $t \in T$, be a set of Gaussian random variables indexed by a countable set T. Assume that $E\eta_t = 0$, $E\eta_t^2 \le \sigma^2$ with some $0 \le \sigma \le 1$ for all $t \in T$, and the metric $d_2(s,t) = \left[E(\eta_s - \eta_t)^2\right]^{1/2}$, $s, t \in T$, has the following property. There exist some constants $L \ge 1$ and $D \ge 1$ such that for all $0 \le \varepsilon \le 1$ a subset $\{t_1, \dots, t_P\} \subset T$ can be found with cardinality $P \le D\varepsilon^{-L}$ for which $\min_{1 \le j \le P} d_2(t, t_j) \le \varepsilon$ for all $t \in T$. Then the inequality*

$$P\left(\sup_{T \in T} |\eta_t| \ge u\right) \le C(D+1)\exp\left\{-\frac{1}{256}\left(\frac{u}{\sigma}\right)^2\right\}$$

$$\textit{if } u \ge ML^{1/2}\sigma \log^{1/2}\frac{2}{\sigma}$$

holds with some universal constants $C > 0$ and $M > 0$.

A comparison of our problem with the above result

In our problem $\mathcal{F}$ plays the role of the parameter set T, and $d_2(f,g)^2 = E(S_n(f-g)^2) = \int [f(x)-g(x)]^2 \mu(\,dx)$, $f,g \in \mathcal{F}$.
The previous theorem also holds if the random variables η_t, $t \in T$, are non-Gaussian, but they satisfy the Gaussian type inequality

$$P(|\eta_t - \eta_s| > u) \le C_1 e^{-C_2 u^2/d_2(s,t)^2} \quad \text{for all } s,t \in T \text{ and } u > 0$$

with some constants $C_1 > 0$ and $C_2 > 0$.
What is the case in our problem?
Some classical results (e.g. Bernstein's inequality) provide the above inequality under some restriction, (e.g. if $u \le \mathrm{const.}\sqrt{n} d_2^2(s,t)$), but it may not hold without such a restriction. (See my lecture note On the estimation of multiple random integrals and U-statistics, Chapter 3.)

To get results similar to the Gaussian case for sums of bounded i.i.d. random variables some additional restriction has to be imposed, and the proof requires new ideas. I know of two approaches. Approach 1. Due to Talagrand. Define the metric $d_\infty(f,g)=\sup_{x\in X}|f(x)-g)x)|$, $f,g\in\mathcal{F}$, and introduce the additional condition that there exist some constants $D\ge1$ and $L\ge1$ such that for all $0\le\varepsilon\le1$ $P\le D(n^{-1/2}\varepsilon)^{-L}$ functions $f_j\in\mathcal{F}$, $1\le j\le P$, can be found for which $\min_{1\le j\le P} d_\infty(f,f_j)\le\varepsilon$ for all $f\in\mathcal{F}$. This condition together with the condition on the metric $d_2(\cdot,\cdot)$ in the theorem about the Gaussian version imply a similar estimate. Talagrand proved a stronger result formulated with a different terminology in his book The generic chaining.

The proof is based on an appropriate version of the chaining argument. In the proof we exploit that if the terms of a sum of independent random variables have a very small bound in the supremum norm, then the tail distribution of the sum has a good Gaussian upper bound even at high levels.

Talagrand found interesting applications of his result, but there are important models where it does not work.

Example where Talagrand's result cannot be applied. Let $(X, \mathcal{X})$ be the unit interval $[0,1]$ with the Borel σ-algebra, let $\xi_1, \dots, \xi_n$ be i.i.d. random variables on $[0,1]$ with uniform distribution. Put $\mathcal{F} = \{f_{a,b}(x)\}$, where $0 \le a < b \le a + \sigma^2$ with a small number $\sigma^2 > 0$, a, b are rational numbers, and $f_{a,b}(x) = I_{a,b}(x) - (b - a)$, where $I_{a,b}(\cdot)$ is the indicator function of the interval (a, b).

This example satisfies the conditions we imposed for our models, and the condition imposed on the metric $d_2(\cdot, \cdot)$ with parameters L and D which have an upper bound not depending on σ^2, but the $d_\infty(\cdot, \cdot)$ metric behaves badly. But all functions $f_{a,b}$ are far from each other in the supremum norm.

To handle such models we introduce a different method.

Approach 2. (Based on a Vapnik–Červonenkis type argument.) We formulate an additional condition on $\mathcal{F}$ with the help of a notion called the class of functions with polynomially increasing covering numbers instead of the condition about the behaviour of the d_∞ metric. We prove results with its help.

We give the definition of a class of functions with polynomially increasing covering numbers in two steps.
First step of the definition.

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 $f\in\mathcal{F}$.

Second step of the definition.

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

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

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.

There is a good estimate for the probability of $P(\sup_{f\in\mathcal{F}} |S_n(f) > x)$ if the class of functions $\mathcal{F}$ has polynomially increasing covering numbers with some parameter $D \ge 1$ and $L \ge 1$. See On the estimation of multiple random integrals and U-statistics, Theorem 4.1. This result describes fairly well when we can get such a good estimate in our problem as in its Gaussian counterpart.

The proof is based on some ideas of K. S. Alexander, and applies the so-called symmetrization argument. It says that under some not too restrictive conditions $\sup_{f\in\mathcal{F}}\sum_{j=1}^n f(\xi_j)$ has a similar tail distibution as its randomized version $\sup_{f\in\mathcal{F}}\sum_{j=1}^n \varepsilon_j f(\xi_j)$, where $\varepsilon_1,\dots,\varepsilon_n$ are i.i.d. random variables, independent also of the random variables ξ_j, and $P(\varepsilon_1=1)=P(\varepsilon_1=-1)=\frac12$.
The symmetrized version $\sup_{f\in\mathcal{F}}\sum_{j=1}^n \varepsilon_j f(\xi_j)$ can be better handled than the original expression, and this is exploited in the symmetrization argument..

Nevertheless, this technique gives a weak estimate if $Ef(\xi_1)^2\le\sigma^2$ for all $f\in\mathcal{F}$ with a very small σ^2. In this case the above mentioned Theorem 4.1 in my lecture note does not give a sharp estimate. Our goal is to give a sharp estimate in all cases. It turned out that the next result, the main result of the paper under discussion, plays a crucial role in achieving this goal.
Here is this result.

Main theorem. *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 $D \geq 1$ and exponent $L \geq 1$, and $\sup_{x \in X} |f(x)| \leq 1$ for all $f \in \mathcal{F}$. Let $\xi_1, \dots, \xi_n$, $n \geq 2$, be a sequence of i.i.d. random variables with values in the space $(X, \mathcal{X})$ with a distribution μ, and assume that the inequality $\int |f(x)|\mu(\,dx) \leq \rho$ holds for all $f \in \mathcal{F}$ with a number $0 < \rho \leq n^{-200}$. Put $\bar{S}_n(f) = \bar{S}_n(f)(\xi_1, \dots, \xi_n) = \sum_{j=1}^n f(\xi_j)$ for all $f \in \mathcal{F}$. The inequality*

$$P\left(\sup_{f \in \mathcal{F}} |\bar{S}_n(f)| \geq u\right) \leq D\rho^{Cu} \quad \textit{for all } u > 41L$$

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

(The condition $\int f(x)\mu(\,dx) = 0$ for all $f \in \mathcal{F}$ is not needed in this result.)

An example that may help to understand the content of the Main theorem

Let $(X,\mathcal{X})$ be a finite set with N elements, μ the uniform distribution on it, ($\mu(A)=\frac{1}{N}\cdot\text{the number of elements in } A$.) Fix a number $d\ge 1$ and let $\mathcal{F}$ be the set of indicator functions of the subsets of X containing at most d elements. Let $\xi_1,\dots,\xi_n$ be i.i.d. random variables, and define $\bar S_n(f)=\sum_{1\le j\le n} f(\xi_j)$ with their help.

Question: What can we say about $P(\sup_{f\in\mathcal{F}}\bar S_n(f)>u)$?

We are interested in the case when $\frac{n}{N}$ is small. It is clear that $P(\sup_{f\in\mathcal{F}}\bar S_n(f)\ge u)=1$ if $u\le d$. If $u>d$, then this probability is a fast decreasing function of u. (See Section 1 of the paper for details.) On the other hand the indicator functions $\{x\colon\ f(x)=1\}$, $f\in\mathcal{F}$, constitute a (classical) example of Vapnik–Červonenkis classes.

The notion of class of functions with polynomially increasing covering numbers can be considered as a version of Vapnik–Červonenkis classes for classes of functions, and these two notions behave similarly. In particular, by some results the class of functions in the above example has polynomially increasing covering numbers with exponenent $L = (1+\varepsilon)d$ for all $\varepsilon > 0$ and appropriate $D = D(\varepsilon)$.
Some considerations show that the main theorem can be interpreted in the following way:

For a class of functions $\mathcal{F}$ with polynomially increasing covering numbers and the property that $\sup_{f\in\mathcal{F}} \int |f(x)|\mu(\,dx)$ is very small the tail distribution of $|\sup_{f\in\mathcal{F}} \bar{S}_n(f)|$ satisfies such an inequality that the above (simple) example suggests.

The real role of the Main theorem

We are interested in the Main theorem not for itself. It is interesting for us, because it yields a better application of the Vapnik–Červonenkis argument than the symmetrization method.

The condition $\int |f(x)|\mu(\,dx) \le \rho$ for all $f \in \mathcal{F}$ with $0 < \rho \le n^{-200}$ is not very restrictive. This (together with the existence of polynomially increasing covering numbers in our models) enables us to split up the class of functions $\mathcal{F}$ into the union of relatively few subsets $\mathcal{F}_j$, $1 \le j \le M$, ($M \le n^K$ with some $K > 0$) in such a way that fixing some $f_j \in \mathcal{F}_j$ the tail distribution of $\sup_{f \in \mathcal{F}_j} |\bar{S}_n(f - f_j)|$ can be well bounded by means of the Main theorem.

Since we have to work only with polynomially many subsets $\mathcal{F}_j \subset \mathcal{F}$ we can give a better, more complete solution of our problem. This is the topic of my paper Sharp tail distribution estimates for the supremum of a class of sums of i.i.d. random variables.

The main ideas of the proof

Let X be a finite set with 2^k elements with a large integer k, μ be the uniform distribution on X. Consider a model satisfying the conditions of the Main theorem with such an X and μ, and prove the estimate of the Main theorem for $P(\sup_{f\in\mathcal{F}} \bar{S}_n(f)| \ge n)$. We may assume that $f(x) \ge 0$ for all $x \in X$ and $f \in \mathcal{F}$.

We prove first this special result in Theorem 1A and then the Main theorem with its help. We prove Theorem 1A with the help of an induction procedure for k with a large starting number k_0. The starting step of the induction is done in Lemma 3.1.

To carry out our induction procedure we need a result formulated in Lemma 3.2. This states that if a class of functions $\mathcal{F}$ with polynomially increasing covering numbers on a set X of cardinality 2^k with a large number k has the property $\int f(x)\mu(\,dx) \le \rho$ for all $f \in \mathcal{F}$ with the uniform distribution μ on X, then this property is preserved for most subsets of X with cardinality 2^{k-1} if ρ is replaced by a slightly larger $\bar{\rho}$.

To prove Theorem 1A observe that $\sup_{f\in\mathcal{F}} \bar{S}_n(f) \ge n$ if and only if there is some $f\in\mathcal{F}$ for which $f(\xi_j)=1$ for all $1\le j\le n$. Hence to prove Theorem 1A we have to give a good estimate on the number of sequences $(x_{l_1},\dots,x_{l_n})\in X^n$ such that $x_{l_j}\in B_f=\{x\colon\ f(x)=1\}$ for all $1\le j\le n$, and some $f\in\mathcal{F}$. We prove such an estimate by means of induction with respect to k with the help of Lemma 3.2.

We assume that an appropriate estimate holds on the number of such sequences for classes of functions $\mathcal{F}$ with polynomially increasing covering numbers with exponent $L\ge1$ and parameter $D\ge1$ on a set X of cardinality 2^{k-1} if also the condition $\int|f(x)|\mu(\,dx)\le\rho_{k-1}$ holds for all $f\in\mathcal{F}$ with an appropriately chosen ρ_{k-1}. Then we prove the analogous result for parameter k with an appropriately chosen ρ_k by means of Lemma 3.2.
We define for all functions $f\in\mathcal{F}$ and sets $B\subset X$ of cardinality 2^{k-1} the function f_B as the restriction of f to B, and put $\mathcal{F}_B=\{f_B\colon\ f\in\mathcal{F}\}$.

Then by applying the inductional hypothesis for $k-1$ for the sets $B \subset X$ of cardinality 2^{k-1} with the classes of functions $\mathcal{F}_B$ and taking an average for all sets B of cardinality 2^{k-1} we get an estimate on the number of sequences $(x_{l_1}, \dots, x_{l_n}) \in X^n$ with the requested properties. In this calculation we apply Lemma 3.2 and choose the constants ρ_k in an appropriate way. If we do this carefully, then we can carry out the induction procedure, and by letting $k \to \infty$ we get the proof of Theorem 1A.

The crucial part of the problem is the proof of Lemma 3.2. Its main step is to show that $\int (f_B - f_{X \setminus B})\, d\mu$ is small with probability almost 1 for any $f \in \mathcal{F}$, if we choose the set B randomly among the sets $B \subset X$ of cardinality 2^{k-1}, and f_B denotes the restriction of f to the set B. This implies that $\int f_B \, d\mu_B \le \bar\rho$ with a number slightly larger than ρ with probability almost 1, where μ_B is the uniform distribution on B. With an appropriate choice of $\bar\rho$ we can achieve that even $\sup_{f \in \mathcal{F}} \int f_B \, d\mu_B \le \bar\rho$ with probability almost 1. Here we exploit that $\mathcal{F}$ has polynomially increasing covering numbers. The last inequality implies Lemma 3.2.

We can estimate the tail distribution of $\int (f_B - f_{X\setminus B})\,d\mu = 2^{-k}(\sum_{x\in B} f(x) - \sum_{x\notin B} f(x))$, where B is a randomly chosen subset of X with cardinality 2^{k-1} by means of a method that appeared also in the proof of Lemma 3 of the paper Komlós–Major–Tusnády: An approximation of partial sums of independent rv's and the sample DF.

The method: Take a pairing $(x_{l_1}, x_{l_2}), \dots (x_{l_{2^k-1}}, x_{l_{2^k}})$ of the set X, and define a random set B with 2^{k-1} elements by putting in each pair one randomly chosen element into the set B and the other one into the complementary set. We can well estimate the tail distribution of $2^{-k}(\sum_{x\in B} f(x) - \sum_{x\notin B} f(x))$ if we choose only sets B obtained in such a way e.g. with the help of Hoeffding's inequality. Then averaging for all possible pairings of X we can get the estimate we need to complete the proof of Lemma 2.2.

Finally I briefly explain how to prove the Main theorem with the help of Theorem 1A.

First I give a good estimate on $P(\sup_{f\in\mathcal{F}} \bar{S}_n(f) > u)$ for all $u > 0$ under the conditions of Theorem 1A in Lemma 4.1. To prove this estimate I show that the event $\sup_{f\in\mathcal{F}} \bar{S}_n(f) > u$ may hold only if for some index j there is a relatively long sequence $\{l_1,\dots,l_s\}\subset\{1,\dots,n\}$ such that $f(\xi_{l_1})\ge 2^{-j},\dots,$ $f(\xi_{l_s})\ge 2^{-j}$ for some $f\in\mathcal{F}$. The probability of such an event can be well estimated with the help of Theorem 1A. Then a careful calculation provides the proof of Lemma 4.1.

The Main theorem is proved by means of Lemma 4.1. We can reduce the proof to the case when $\mathcal{F}$ is a finite set. If its cardinality is R, then we can approximate $\mathcal{F}=\{f_1,\dots,f_R\}$ with a class of functions $\mathcal{G}=\{g_1,\dots,g_R\}$ whose elements take only finitely many values, the μ-distribution of all events $\{g_1(x)=u_1,\dots,g_R(x)=u_R\}$ is an integer multiplied by 2^{-k} with some number k, and the functions g_j are so close to the functions f_j that it is enough to prove the Main Theorem for $\mathcal{G}$ instead of $\mathcal{F}$. On the other hand, this can be done with the help of Lemma 4.1.