Sharp tail distribution estimates for the supremum of a class of sums of i.i.d. random variables.

Péter Major

2015. október 22.

The problem we are interested in

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 $\mathcal{F}$ consisting of countably many functions be given on the space $(X, \mathcal{X})$ with the properties $\int f(x)\mu(\,dx) = 0$, $\sup_{x \in X} |f(x)| \le 1$ and $\int f(x)^2 \mu(\,dx) \le \sigma^2$ with some $0 < \sigma \le 1$ for all elements $f \in \mathcal{F}$.

Let $\mathcal{F}$ be a class of functions with polynomially increasing covering numbers with exponent $L \ge 1$ and parameter $D \ge 1$. (I recall the definition of this notion later.)

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

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

of the supremum of these sums. This estimate may depend on σ, L and D.

An additional 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 tail distribution we are investigating is small only if v is larger than this concentration point. In our estimation we want to find a good level above which this tail distribution begins to decrease radically. This is a hard and important part of our problem.

I recall the definition of classes of functions with polynomially incerasing covering numbers together with the exponent L and parameter D of these classes. I do it in two steps.
This notion is a useful version of the Vapnik–Červonenkis classes, when we are working with classes of functions instead of classes of sets.

Definition of a class of functions with polynomially increasing covering numbers

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.

First I discuss an example which indicates what kind of results we can expect in our problem. We are mainly interested in the case when the exponent L and the parameter D are bounded (by a number not depending on σ^2), and σ^2 may be very small.

A useful example

Example. Take a sequence of independent, uniformly distributed random variables $\xi_1, \dots, \xi_n$ on the unit interval $[0,1]$, fix a number $0 \le \sigma^2 \le 1$, and define a class of functions $\mathcal{F}_\sigma$ and $\bar{\mathcal{F}}_\sigma$ as set of functions defined on the unit interval $[0,1]$ in the following way. $\mathcal{F}_\sigma = \{f_1, \dots, f_k\}$, and $\bar{\mathcal{F}} = \{\bar{f}_1, \dots, \bar{f}_k\}$ with $k = k(\sigma) = [\frac{1}{\sigma^2}]$, where $[\cdot]$ denotes integer part, and $\bar{f}_j(x) = \bar{f}_j(x|\sigma) = 1$ if $x \in [(j-1)\sigma^2, j\sigma^2)$, $\bar{f}_j(x) = \bar{f}_j(x|\sigma) = 0$ if $x \notin [(j-1)\sigma^2, j\sigma^2)$, $1 \le j \le k$, and $f_j(x) = f_j(x|\sigma) = \bar{f}_j(x) - \sigma^2$, $1 \le j \le n$. Give a good estimate on $P_n(v) = P(\sup_j S_n(f_j) > v)$.

$\mathcal{F}$ satisfies our conditions. It is a class of functions with polynomially increasing covering numbers with exponent L and parameter D which do not depend on σ^2, and the parameter σ^2 introduced in the model is an upper bound for all $\int f_j(x)^2 \mu(\,dx)$.

Our first question: For which numbers is $P_n(v)$ much smaller than 1? Answer to this question:

An estimate on the function $P_n(v)$ in the models of the above example. A number $\bar C>0$ can be chosen in such a way that for all $\delta>0$ there is an index $n_0(\delta)$ such that for all sample sizes $n\ge n_0(\delta)$ and numbers $0\le\sigma\le 1$ the inequality

$$P_n(\hat u(\sigma))=P\left(\sup_{f\in\mathcal{F}_\sigma}|S_n(f)|\ge\hat u(\sigma)\right)\ge 1-\delta,$$

holds with

1.) $\hat u(\sigma)=\frac{\bar C}{\sqrt{n}}$ if $\sigma^2\le n^{-400}$,

2.) $\hat u(\sigma)=\frac{\bar C}{\sqrt{n}}\frac{\log n}{\log(\frac{\log n}{n\sigma^2})}$ if $n^{-400}<\sigma^2\le\frac{\log n}{8n}$, and

3.) $\hat u(\sigma)=\bar C\sigma\log^{1/2}\frac{2}{\sigma}$ if $\frac{\log n}{8n}\le\sigma^2\le 1$.

This result says that we cannot get a good estimate on the probability we are interested for $v\le\hat u(\sigma)$. First I explain this result, then I discuss what we can say if $v>\hat u(\sigma)$.

In case 3.) of this example σ^2 is relatively large. In this case the $S_n(f)$ behaves similarly to the Gaussian case, (like a functional of a Brownian bridge), and similar estimates hold for the tail distribution of $\sup_{f\in\mathcal{F}} S_n(f)$ as in the corresponding Gaussian model. But to get such a good estimate we need this condition. K. S. Alexander also observed this fact in his research.

In case 2.) $S_n(f)$ does not have a good Gaussian, but has a good Poissonian approximation. This provides a slightly weaker estimate than in case 1.), since the Poissonian tail distribution tends to zero slower at ∞ than the Gaussian one. Here we explained what we get in this case.

In case 1.) we considered the case when σ^2 is very small. Here we exploited the trivial fact that if we take an arbitrary partition of the probability space a sample point gets into one of the elements of the partition. In this case this observation provides the right estimate.

The next Theorem (the main result of this paper) states that in the general case we get an estimate suggested by the above example. Actually the situation is somewhat more complex, since we also consider the case when the parameters L and D may be large.

We can get a good estimate on $P(\sup_{f\in\mathcal{F}}|S_n(f)|>v)$ only if $v>\hat u(\sigma)$. We also want to find the tail distribution in this case. Bernstein's and Bennett's inequality suggest the upper bound $e^{-C\sqrt{n}\log(v/\sqrt{n}\sigma^2}$ if $v\ge \text{const.}\,\sqrt{n}\sigma^2$ and e^{-Cv^2/σ^2} if $v\le \text{const.}\,\sqrt{n}\sigma^2$. (See my lecture note On the estimation of multiple random integrals an U-statistics).

In cases 1.) and 2.) $\hat u(\sigma)\ge \text{const.}\sqrt{n}\sigma^2$, and the Theorem gives the estimate we expect. Case 3.) is more complex. In the Theorem we give the estimate we expect if $v\ge \text{const.}\,\sqrt{n}\sigma^2$. (The situation is somewhat more difficult, because we also deal with the case when the parameters L and D are large.) In Case 3.) it is possible that $\hat u(\sigma)<v\le \text{const.}\,\sqrt{n}\sigma^2$. We prove the (Gaussian) estimate we expect in this case in an Extension of the Theorem.

Theorem. *Let a sequence of i.i.d. random variables $\xi_1,\dots,\xi_n$, $n\ge 2$, with values in $(X,\mathcal{X})$ with some distribution μ and a countable class of functions $\mathcal{F}$ on the same space $(X,\mathcal{X})$ with polynomially increasing covering numbers with exponent $L\ge 1$ and parameter $D\ge 1$ be given. Let the functions $f\in\mathcal{F}$ satisfy the relations $\sup_{x\in X}|f(x)|\le 1$, $\int f(x)\mu(\,dx)=0$, and $\int f^2(x)\mu(\,dx)\le\sigma^2$ with some number $0\le\sigma^2\le 1$ for all $f\in\mathcal{F}$. The normalized sums $S_n(f)$, $f\in\mathcal{F}$, satisfy the inequality*

$$P\left(\sup_{f\in\mathcal{F}}|S_n(f)|\ge v\right)\le C_1 e^{-C_2\sqrt{n}v\log(v/\sqrt{n}\sigma^2}\quad \textit{for all } v\ge u(\sigma)$$

with some universal constants $C_j>0$, $1\le j\le 5$, if one of the following conditions is satisfied.

1.) $\sigma^2\le\frac{1}{n^{400}}$, *and* $u(\sigma)=\frac{C_3}{\sqrt{n}}(L+\frac{\log D}{\log n})$,

2.) $\frac{1}{n^{400}}<\sigma^2\le\frac{\log n}{8n}$, *and* $u(\sigma)=\frac{C_4}{\sqrt{n}}\left(L\frac{\log n}{\log(\frac{\log n}{n\sigma^2})}+\log D\right)$,

3.) $\frac{\log n}{8n}<\sigma^2\le 1$, *and* $u(\sigma)=\frac{C_5}{\sqrt{n}}(n\sigma^2+L\log n+\log D)$.

Next we consider the case $\sigma^2 \ge \frac{\log n}{8n}$ and $\sqrt{n}\sigma^2 \ge v \ge \bar{u}(\sigma)$ with some $\bar{u}(\sigma)$ which has the same order of magnitude as $\hat{u}(\sigma)$. Actually this result was proved earlier.

Extension of the Theorem. *Let us consider, similarly to the Theorem, a sequence of i.i.d. random variables $\xi_1, \dots, \xi_n$, $n \ge 2$, with values in a space $(X, \mathcal{X})$ with some distribution μ which satisfies the conditions of the Theorem. In the case $\frac{\log n}{8n} < \sigma^2 \le 1$ the supremum of the normalized sums $S_n(f)$, $f \in \mathcal{F}$, satisfies the inequality*

$$P\left(\sup_{f\in\mathcal{F}} |S_n(f)| \ge v\right) \le Ce^{-\alpha v^2/\sigma^2}$$

with appropriate (universal) constants $\alpha > 0$, $C > 0$ and $C_6 > 0$ if $\sqrt{n}\sigma^2 \ge v \ge \bar{u}(\sigma)$, where $\bar{u}(\sigma)$ is defined as $\bar{u}(\sigma) = C_6\sigma(L^{3/4}\log^{1/2}\frac{2}{\sigma} + (\log D)^{1/2})$.

The idea of the proof

We choose an appropriate number $\delta > 0$, and choose by exploiting that $\mathcal{F}$ is a class of functions with polynomially increasing covering numbers $m = D\delta^{-L}$ functions $f_j \in \mathcal{F}$, $1 \le j \le m$, and set of functions $\mathcal{D}_j \subset \mathcal{F}$ in such a way that $\int |g - f_j| \, d\mu \le \delta$, if $g \in \mathcal{D}_j$ and $\bigcup_{j=1}^m \mathcal{D}_j = \mathcal{F}$.
We can write

$$
\begin{aligned}
& P\left(\sup_{f\in\mathcal{F}} |S_n(f)| \ge v\right) && (1)\\
& \quad \le P\left(\sup_{1\le j\le m} |S_n(f_j)| \ge \frac{v}{2}\right) + \sum_{j=1}^m P\left(\sup_{f\in\mathcal{D}_j} |S_n(f-f_j)| \ge \frac{v}{2}\right).
\end{aligned}
$$

We choose $\delta > 0$ in an appropriate way. Then we can give a good estimate on the second term of the sum at the right-hand side of (1) by means of my paper Sharp estimate on the supremum of a class of sums of small i.i.d. random variables.
The first term can be estimated by means of the inequality

$$P\left(\sup_{1\le j\le m}|S_n(f_j)|\ge\frac{v}{2}\right)\le\sum_{j=1}^{m}P\left(|S_n(f_j)|\ge\frac{v}{2}\right)$$

and Bennett's inequality. The theorem can be proved in such a way.

The Extension of the Theorem can be proved similarly. Only in this case the first term at the right-hand side of (1) must be estimated in a different way. We exploit the properties of the class of functions $\mathcal{G}=\{f_1,\dots,f_m\}$, and give a good estimate with the help of the chaining argument. (Observe that $\mathcal{G}$ is a class of functions with polynomially increasing covering numbers).

Some remarks on the problem

In a previous paper I gave a good estimate on the probability $P(\sup_{f\in\mathcal{F}} S_n(f) > v)$ if $\mathcal{F}$ is a class of functions with polynomially increasing covering numbers consisting of functions bounded by 1, and $\int |f(x)|\mu(\,dx) \le \rho$ with a sufficiently small ρ. More precisely, $\rho \le n^{-\alpha}$ with an appropriate $\alpha > 1$. Here μ denotes the distribution of the random variables we are working with.

This result played a crucial role in our investigation. It enabled us to reduce the problem to the case where we take the supremum for an appropriate finite subset of $\mathcal{F}$, because it made possible to control the small contribution of the disregarded terms to the supremum we are investigating.

This approach is similar to the truncation technique applied in the proof of limit theorems, by which the small but irregular effect of the large terms is disregarded. Here a similar method is applied.

The chaining argument or a more refined version of it worked out by Talagrand enables us to handle the regular effects in similar problems. But the control of the small irregularities demands a different method. In earlier works the irregularities were controlled by means of a method called the symmetrization argument. In the present paper I could find a more powerful method that works under more general conditions.

The control of the irregular effects is a more general, open problem. Here we exploited that the class of functions we are working with has polynomially increasing covering numbers. Other models have other good properties, and we have to find the method to exploit them.

On the other hand, I consider a method to control the irregularities good only if I see models where it it gives new results. I met some generalizations of the symmetrization argument which demanded new complicated notions and arguments. But as I saw no real application of them I do not know whether they are useful.