On the supremum of partial sums of independent random variables

Péter Major

Mathematical Institute of the Hungarian Academy of Sciences.

The original problem: Let $\xi_1, \ldots, \xi_n$ be a sequence of i.i.d. random variables with some distribution $\mu$ on a space $(X, \mathcal{X})$.

Introduce the empirical distribution $\mu_n$

$$
\mu_n(A) = \frac{1}{n} \text{ the number of indices } j, 1 \leq j \leq n, \text{ for which } \xi_j \in A.
$$

of this sequence together with its normalization $\nu_n(A) = \sqrt{n}(\mu_n(A) - \mu(A))$.

Consider a function $f(x_1, \ldots, x_k)$ of $k$ variables on $(X^k, \mathcal{X}^k)$ and define its integral of (order $k$)
$$J_{n,k}(f) = \int' f(x_1, \ldots, x_k) \nu_n(dx_1) \ldots \nu_n(dx_k)$$

$$= n^{k/2} \int' f(x_1, \ldots, x_k)(\mu_n(dx_1) - \mu(dx_1)) \ldots (\mu_n(dx_k) - \mu(dx_k)).$$

with respect to the (random) signed measure $\nu_n$.

Remark: The notation $\int'$ in the definition of the random integral $J_{n,k}(f)$ means that the diagonals, i.e. (the points $x = (x_1, \ldots, x_k)$ such that $x_j = x_{j'}$ for some pairs of indices $j \neq j'$) are omitted from the domain of integration. (For some technical reasons this seems to be the right formulation of the above problems.)

Problem: Let a nice class $\mathcal{F}$ of functions of $k$ variables $f(x_1, \ldots, x_k)$ be given. Take the integral $J_{n,k}(f)$ for all functions $f \in \mathcal{F}$.
\( \mathcal{F} \), and give a good estimate on their supremum, i.e. on the probability

\[
P \left( \sup_{f \in \mathcal{F}} J_{n,k}(f) > u \right)
\]

for all numbers \( u > 0 \).

**A similar problem:** Introduce the \( U \)-statistic \( I_{n,k}(f) \)

\[
I_{n,k}(f) = \frac{1}{k!} \sum_{1 \leq j_s \leq n, s=1,\ldots,k, j_s \neq j_{s'}} f(\xi_{j_1}, \ldots, \xi_{j_k})
\]

(of order \( k \)) determined by the kernel function \( f(x_1, \ldots, x_k) \) and independent \( \mu \)-distributed independent random variables \( \xi_1, \ldots, \xi_n \).

**Problem’.** Give a good estimate on the probabilities

\[
P \left( n^{-k/2} \sup_{f \in \mathcal{F}} I_{n,k}(f) > u \right),
\]
where $F$ is a nice class of functions of $k$ variables.

To understand the problem better let us restrict our attention to the case $k = 1$ when we have to bound the tail distribution of the supremum of the normalized partial sums $\frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(\xi_j)$ with a nice class of functions $f \in F$. A crucial part of the problem: How to define nice classes of functions?

A similar problem investigated by Michel Talagrand: Give a good upper bound on

$$E \sup_{f \in F} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(\xi_j).$$

By a concentration inequality which says that

$$P \left( \sup_{f \in F} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(\xi_j) - E \sup_{f \in F} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(\xi_j) > x \right)$$
is small the two problems are equivalent. Moreover, Talagrand’s estimate on the expectation of the supremum contains an implicit estimate on its tail distribution.

The main difference between my result and that of Talagrand is the choice of the class of nice functions $\mathcal{F}$.

Partial sums of independent random variables behave similarly to Gaussian random variables. If $Ef(\xi_j) = 0$ for all $f \in \mathcal{F}$, then

$$E \left( \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(\xi_j) \cdot \frac{1}{\sqrt{n}} \sum_{j=1}^{n} g(\xi_j) \right)$$

$$= \int f(x)g(x)\mu(dx)$$

for all $f, g \in \mathcal{F}$, where $\mu$ is the distribution of $\xi_j$.

Solve first the analogous Gaussian problem.
Let $\eta_t$, $E\eta_t = 0$, $t \in T$, be a countable set of (jointly) Gaussian random variables. 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 $T$.

Our problem: Give a good estimate on $E \sup_{t \in T} \eta_t$ with the help of the function $d_2(s,t)$.

Results with the help of a classical and natural method, the the chaining argument.

A classical estimate due to R. M. Dudley.

Define

$$e(n) = \min_{T_n \subset T, \text{card } T_n \leq 2^{2n}} \{ \alpha, \text{ for all } t \in T \text{ there is some } \bar{t} \in T_n \text{ such that } d_2(t, \bar{t}) \leq \alpha \}$$

Theorem of Dudley:

$$E \sup_{t \in T} \eta_t \leq \sum_{n=0}^{\infty} 2^{n/2} e(n).$$
The content of the notion $e(n)$. Place uniformly $2^{2^n}$ points on the set $T$ in such a way that each point $t \in T$ is close to some point of this set. The number $e(n)$ is the smallest $\alpha$ for which each point is closer to this set than $\alpha$.

The idea of the proof:

Fix some $u > 0$, and sets $T_n \subset T$, $\text{card } T_n \leq 2^{2^n}$, $n = 1, 2, \ldots$, and put

$$Q(u) = P \left( \sup_{t \in T} \eta_t > u \sum_{n=0}^{\infty} 2^{n/2} e(n) \right)$$

and

$$Q_N(u) = P \left( \sup_{t \in T_1 \cup \cdots \cup T_N} \eta_t > u \sum_{n=0}^{N-1} 2^{n/2} e(n) \right), \quad N = 0, 1, 2, \ldots.$$

We can estimate well $Q(u)$ for all $u \geq 2$, by giving a good bound on $Q_N(u) - Q_{N-1}(u)$ for
all \( N = 1, 2, \ldots \). Here we exploit that for all \( t \in T_N \) there is some \( \bar{t} \in T_{N-1} \) which is close to it. We get good estimates which show that the main contribution to the expectation we want to bound comes from the event \( \sup_{t \in T} \eta_t \leq 2 \sum_{n=0}^{\infty} 2^{n/2} e(n) \).

Talagrand found a sharper estimate by introducing the right quantity \( \gamma_2(T, d) \) needed in the study of this problem.

To define it, let us first introduce the diameter \( \Delta(A) = \sup_{s,t \in A} d(s, t) \) of a set \( A \subset T \) in a metric space \( (T, d) \), and the notion of

**Admissible sequence of partitions.** A sequence of refining partitions \( A_0 \subset A_1 \subset A_2 \subset \cdots \) of the set of parameters \( T \) is an admissible sequence of partitions if \( \text{card} A_0 = 1 \), and \( \text{card} A_n \leq 2^{2^n}, n = 1, 2, \ldots \).
Given an admissible partition $A_0 \subset A_1 \subset A_2 \subset \cdots$ and a point $t \in T$ let $A_n(t)$ be that element $B$ of the partition $A_n$ for which $t \in B$.

Given a countable parameter set $T$ with a metric $d$ we define

$$\gamma_2(T, d) = \inf \sup \sum_{n=0}^{\infty} 2^{n/2} \Delta(A_n(t)),$$

where the infimum is taken for all admissible sequence of partitions of $T$.

The estimate of Talagrand.

Let $\eta_t$, $E\eta_t = 0$, $t \in T$, a sequence of Gaussian random variables, $d_2(s, t) = [E(\eta_t - \eta_s)^2]^{1/2}$. Then

$$E\sup_{t \in T} \eta_t \leq L\gamma_2(T, d_2).$$

Moreover

$$\frac{1}{L} \gamma_2(T, d_2) \leq E\sup_{t \in T} \eta_t \leq L\gamma_2(T, d_2).$$
with a universal constant $T$.

This is the main estimate about the expected value of the supremum of Gaussian random variables. Another hard problem. How to estimate $\gamma_2(T, d)$.

What can be said about the supremum of normalized partial sums of i.i.d. random variables.

The result (and proof) about the expected value of the supremum of Gaussian random variables $\eta_t, t \in T$, remains valid also in the non-Gaussian case if

$$P(|\eta_s - \eta_t| > u) \leq 2e^{-u^2/2d_2(s,t)}.$$

What can be told about normalized partial sums of i.i.d. random variables?

A classical result, Bernstein’s inequality says:
**Bernstein’s inequality.** Let \( \xi_1, \ldots, \xi_n \) be independent random variables,

\[
P(|\xi_j| \leq 1) = 1, \text{ and } E\xi_j = 0, \ 1 \leq j \leq n.
\]

Put \( \sigma_j^2 = E\xi_j^2, 1 \leq j \leq n \), \( S_n = \sum_{j=1}^{n} \xi_j \) and \( \text{Var}S_n = V_n^2 = \sum_{j=1}^{n} \sigma_j^2 \). Then

\[
P(S_n > u) \leq \exp \left\{ -\frac{u^2}{2V_n^2 \left( 1 + \frac{u}{3V_n^2} \right)} \right\}
\]

for all numbers \( u > 0 \).

If \( u \leq \text{const.} \ V_n^2 \) it supplies a Gaussian type estimate, but if \( u \gg V_n^2 \), it supplies a bad estimate. Only very weak improvement is possible which does not help if \( u \gg V_n^2 \).

For normalized partial sums \( S_n(f) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(\xi_j) \), \( f \in \mathcal{F} \), of i.i.d. random variables \( \xi_j \), \( Ef(\xi_1) = 0 \).
and \( \sup_{x} |f(x)| \leq 1 \) for all \( f \in \mathcal{F} \) the following Gaussian type estimate holds.

\[
P(|S_n(f) - S_n(g)| \geq u) \leq 2e^{-u^2/100d_2(f,g)},
\]

if \( u \leq 3d_2(f,g)^2\sqrt{n} \)

with \( d_2(f,g)^2 = \int (f(x) - g(x))^2 \mu(dx) \), where \( \mu \) is the distribution of \( \xi_j \).

If \( \sup_{x} |f(x)| < c \) with a small number \( c > 0 \), then such a good Gaussian estimate holds in a larger interval.

For the supremum of partial sums similar estimates hold as for the supremum of Gaussian random variables, but under more restrictive conditions.

Two such results.

To formulate the first result the following notion turned out to be useful.
Definition of $L_2$-dense classes of functions.
Let a measurable space $(X, \mathcal{X})$ be given together with a set $\mathcal{F}$ of $\mathcal{X}$ measurable real valued functions on this space. $\mathcal{F}$ is called an $L_2$-dense class of functions with parameter $D$ and exponent $L$ if for all numbers $1 \geq \varepsilon > 0$ and probability measure $\nu$ there exists a finite $\varepsilon$-dense subset $\mathcal{F}_\varepsilon = \{f_1, \ldots, f_m\} \subset \mathcal{F}$ in the space $L_2(X, \mathcal{X}, \nu)$ with $m \leq D\varepsilon^{-L}$ elements such that $\inf_{f_j \in \mathcal{F}_\varepsilon} \int |f - f_j|^2 \, d\nu < \varepsilon^2$ for all functions $f \in \mathcal{F}$.

Theorem A. (Estimate on the supremum of a class of partial sums). Let us consider a sequence of independent and identically distributed random variables $\xi_1, \ldots, \xi_n$, $n \geq 2$, with values in a measurable space $(X, \mathcal{X})$ and with some distribution $\mu$. Beside this, let a countable and $L_2$-dense class of functions $\mathcal{F}$ with some parameter $D \geq 1$ and exponent $L \geq$
be given on the space \((X, \mathcal{X})\) which satisfies
the conditions
\[
\|f\|_\infty = \sup_{x \in X} |f(x)| \leq 1, \quad \text{for all } f \in \mathcal{F} \tag{1}
\]
\[
\|f\|_2^2 = \int f^2(x) \mu(dx) \leq \sigma^2 \quad \text{for all } f \in \mathcal{F}
\tag{2}
\]
with some constant \(0 < \sigma \leq 1\), and
\[
\int f(x) \mu(dx) = 0 \quad \text{for all } f \in \mathcal{F}. \tag{3}
\]
Define the normalized partial sums \(S_n(f) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} f(\xi_k)\) for all \(f \in \mathcal{F}\).

There exist some universal constants \(C > 0\), \(\alpha > 0\) and \(M > 0\) such that the supremum of the normalized random sums \(S_n(f), f \in \mathcal{F}\), satisfies the inequality
\[
P \left( \sup_{f \in \mathcal{F}} |S_n(f)| \geq u \right) \leq C \exp \left\{ -\alpha \left( \frac{u}{\sigma} \right)^2 \right\}
\]
for those numbers \(u\) for which
\[
\sqrt{n}\sigma^2 \geq u \geq M\sigma \left( L^{3/4} \log^{1/2} \frac{2}{\sigma} + (\log D)^{3/4} \right)
\]
with the parameter $D$ and exponent $L$ of the $L_2$-dense class $\mathcal{F}$.

Under the conditions of this theorem the supremum of partial sums is not much greater than its largest (worst) term.

Next I formulate a result of Talagrand which says how his method about the estimate of the expected value of the supremum of Gaussian random variables can be adapted for the supremum of partial sums of i.i.d. random variables.

Choose a sequence of i.i.d. random variables $\xi_1, \ldots, \xi_n$ with values on a measurable space $(X, \mathcal{X})$ and distribution $\mu$ together with a countable set $\mathcal{F}$ of $\mathcal{X}$ measurable real valued functions on this space which satisfy (1) and (3). Put $T = \mathcal{F}$, and define on it the metrics $d_2(f, g) = \left[ \int (f - g)^2 \mu(dx) \right]^{1/2}$ and $d_\infty(f, g) = \sup_x |f(x) - g(x)|$, for $f, g \in \mathcal{F}$. 
Put (similarly to $\gamma_2(T, d)$)

$$\gamma_\alpha(\mathcal{F}, d) = \inf \sup \sum_{n=0}^{\infty} 2^{n/\alpha} \Delta(A_n(t)),$$

with arbitrary number Green $\alpha > 0$ and metric $d$ on $\mathcal{F}$, where the infimum is taken for all admissible sequences of partitions $A_n$, $n = 0, 1, 2, \ldots$, of $\mathcal{F}$, and $\Delta(A_n(t))$ is the diameter of $A_n(t)$ with respect to the metric $d$.

**Theorem B. (Another estimate on the supremum of a class of partial sums).** Put $S_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(\xi_j)$. Then

$$ES_n(\mathcal{F}) \leq L \left( \gamma_2(\mathcal{F}, d_2) + \frac{1}{\sqrt{n}} \gamma_1(\mathcal{F}, d_{\infty}) \right)$$

with an appropriate universal constant $L > 0$.

**Examples:**

If $X = [0, 1]$, $f(x)$ is a function on $[0, 1]$, $f_{s,t}(x) = f(x)$ if $s \leq x \leq t$, $f_{s,t}(x) = 0$ otherwise, then
\( \mathcal{F} = \{f_{s,t}(x), 0 \leq s < t \leq 1 \} \) is an \( L_2 \)-dense class, and Theorem A can be applied for it. More generally.

**Example.** Take a measurable functions \( f(x) \), on a measurable space \((X, \mathcal{X})\) with the property \( \sup_{x \in X} |f(x)| \leq 1 \). Let \( \mathcal{D} \) be a Vapnik-Červonenkis class consisting of measurable subsets of the set \( X \). Define for all \( D \in \mathcal{D} \) the function \( f_D(\cdot) \) as \( f_D(x) = f(x) \) if \( x \in D \), and \( f_D(x) = 0 \) if \( x \notin D \), i.e. \( f_D(\cdot) \) is the restriction of the function \( f(\cdot) \) to the set \( D \). Then the set of functions \( \mathcal{F} = \{f_D: D \in \mathcal{D}\} \) is \( L_2 \)-dense.

Here again Theorem A is applicable.

An example, when Theorem B is applicable.

\[
E \sup_{f \in \mathcal{C}} \frac{1}{\sqrt{n}} \left| \sum_{l=1}^{n} (f(X_l)) \right| \leq L \sqrt{\log n}
\]
with a universal constant $L$, where $X_1, \ldots, X_n$ is a sequence of independent random variables, uniformly distributed on $[0, 1] \times [0, 1]$, and $C$ is the class of Lipschitz 1 functions $f(x)$ on $[0, 1] \times [0, 1]$ such that $\int_{[0,1] \times [0,1]} f(x) \, dx = 0$.

This result is equivalent to a (famous) result of Ajtai–Komlós–Tusnády.