ESTIMATION OF MULTIPLE RANDOM INTEGRALS AND U-STATISTICS.

Péter Major
Alfréd Rényi Alfréd Mathematical Institute
of the Hungarian Academy of Sciences

This text is an enlarged version of my inauguration talk to the Hungarian Academy of Sciences. First I formulate the problems discussed in this talk and briefly indicate the questions which inspired me to work on this subject. Then I write down the main results, discuss their background together with some pictures and mathematical ideas which explain them better.

1. Introduction. Formulation of the problems.

To formulate the problems first I introduce some notations.

Let $\xi_1,\dots,\xi_n$ be a sequence of independent and identically distributed random variables with some probability distribution μ on a measurable space $(X,\mathcal{X})$ and let μ_n,

$$\mu_n(A)=\frac{1}{n}\#\{j\colon\ \xi_j\in A,\ 1\le j\le n\},\quad A\in\mathcal{X},$$

denote its empirical distribution. Let a measurable function $f(x_1,\dots,x_k)$ of k-variables be given on the product space $(X^k,\mathcal{X}^k)$. Take the k-fold direct product of the normalized version $\sqrt{n}(\mu_n-\mu)$ of this empirical measure μ_n and consider the integral of the function f with respect to it. More explicitly, the (random) integral

$$J_{n,k}(f)=\frac{n^{k/2}}{k!}\int' f(x_1,\dots,x_k)(\mu_n(\,dx_1)-\mu(\,dx_1))\dots(\mu_n(\,dx_k)-\mu(\,dx_k)),$$

where the prime in $\int'$ means that the diagonals $x_j=x_l,\ 1\le j<l\le k$, are omitted from the domain of integration

(1.1)

will be considered.

The following two problems will be studied in this paper:

Problem A). Give a good estimate on the probabilities $P(J_{n,k}(f)>u)$ under appropriate conditions for the function f.

(The omission of the diagonals $x_j=x_l$, $j\neq l$, from the domain of integration turned out to be natural in possible applications.)

The second, more general problem is the following one.

Problem B). Let a nice class $\mathcal{F}$ of functions $f(x_1,\dots,x_k)$ be given on the space $(X^k,\mathcal{X}^k)$. Give a good estimate on the probabilities $P\left(\sup\limits_{f\in\mathcal{F}}J_{n,k}(f)>u\right)$, where $J_{n,k}(f)$ denotes the integral of the function f defined in formula (1.1).

It turned out useful to study these two problems together with their U-statistic analogues. To formulate them first I recall the definition of U-statistics.

The definition of U-statistics. *Let a sequence $\xi_1,\dots,\xi_n$ of independent and identically distributed random variables be given with values on some measurable space $(X,\mathcal{X})$ together with a function $f(x_1,\dots,x_k)$ on the k-fold product space $(X^k,\mathcal{X}^k)$ with some $k\le n$. The expression*

$$I_{n,k}(f)=\frac{1}{k!}\sum_{\substack{1\le j_s\le n,\ s=1,\dots,k\\ j_s\ne j_{s'} \text{ if } s\ne s'}} f\left(\xi_{j_1},\dots,\xi_{j_k}\right) \tag{1.2}$$

is called a U-statistic of order k with kernel function f.

I formulated the following versions of the above two problems.

Problem A′). Give a good estimate on the probabilities $P(n^{-k/2}I_{n,k}(f)>u)$ under appropriate conditions for the function f.

Problem B′). Let a nice class $\mathcal{F}$ of functions $f(x_1,\dots,x_k)$ be given on a (product) space $(X^k,\mathcal{X}^k)$ together with a sequence of independent and identically distributed random variables $\xi_1,\dots,\xi_n$ with values in $(X,\mathcal{X})$. Give a good estimate on the probabilities $P\left(\sup\limits_{f\in\mathcal{F}} n^{-k/2}I_{n,k}(f)>u\right)$, where $I_{n,k}(f)$ denotes the U-statistic of order k with kernel function f defined in formula (1.2).

It may be useful to remark that a U-statistic of order k with the kernel function f can be rewritten as

$$I_{n,k}(f)=\frac{n^k}{k!}\int' f(x_1,\dots,x_k)\mu_n(\,dx_1)\dots\mu_n(\,dx_k),$$

where μ_n is the empirical distribution of the sequence $\xi_1,\dots,\xi_n$. This shows that the essential difference between the random integrals introduced in formula (1.1) and the U-statistics is that in the random integrals $J_{n,k}(f)$ integration is taken with respect to the 'normalized' measures $\mu_n-\mu$, while in the integral representation of the U-statistics $I_{n,k}(f)$ with respect to the 'non-normalized' measures μ_n.

I met the above problems when tried to adapt a simple method applied in the study of the asymptotic behaviour of maximum likelihood estimates to the investigation of harder problems. In the study of maximum likelihood estimates the root of the so-called maximum likelihood equation has to be well estimated. This can be done by means of a good approximation of the function in the maximum likelihood equation which is obtained with the help of its Taylor expansion if the high order terms of this expansion are omitted. It has to be shown that such an approximation causes only a negligibly small error. But this can be proved relatively simply.

I tried to apply a similar method in the study of some so-called non-parametric maximum likelihood estimation problems. Such a problem arises for instance if we want to estimate an unknown distribution function by means of some partial information. In the study of the error of such an estimate in a fixed point a version of the Taylor expansion can be applied together with the omission of the higher order terms. But in this case it is much harder to show that such an approximation causes only a negligibly small error. To prove this a good solution of Problem A is needed. If we want to bound the error of the estimation in all points simultaneously, then we need a good solution of Problem B.

In the investigation of general non-parametric estimation problems several additional difficulties have to be overcome, but the solution of Problems A and B is especially important. Beside this, these problems are related to a better understanding of some fundamental probabilistic phenomena. Hence I found useful a detailed study of the above questions.

2. An overview of the problems. The study of the one-variate case.

Before a detailed discussion it is worth thinking over what kind of results can be expected. Let us observe that the normalized signed measures $\sqrt{n}(\mu_n-\mu)$ converge to a Gaussian field as $n\to\infty$. Hence it is natural to expect that under very general conditions such results hold both in the solution of Problem A and Problem B which their Gaussian counterparts suggest. But we have to understand the answer to the following two questions.

1.) What kind of estimates do the Gaussian counterparts of these problems suggest?
2.) What does the expression 'under very general conditions' mean?

To clarify the above questions it is useful to study first Problem A in the case $k=1$ when the distribution of sums of independent random variables has to be bounded. Such a bound is given in the following classical result called Bernstein's inequality.

Bernstein's inequality. *Let $\xi_1,\dots,\xi_n$ be independent random variables which satisfy the relations $P(|\xi_j|\le 1)=1$ and $E\xi_j=0$, $1\le j\le n$. Let us introduce the notation $\sigma_j^2=E\xi_j^2$, $1\le j\le n$, $S_n=\sum\limits_{j=1}^n\xi_j$ and $V_n^2=\operatorname{Var}S_n=\sum\limits_{j=1}^n\sigma_j^2$. The inequality*

$$P(S_n>u)\le\exp\left\{-\frac{u^2}{2V_n^2\left(1+\frac{u}{3V_n^2}\right)}\right\} \tag{2.1}$$

holds for all numbers $u>0$.

Bernstein's inequality yields an estimate on the distribution of sums of independent random variables suggested by the central limit theorem, although the coefficient $1+\frac{u}{3V_n^2}$ in the denominator of the upper bound slightly modifies the picture. In the next remark the effect of this factor is considered in different cases.

a) If $u \le \varepsilon V_n^2$ with some small number $\varepsilon > 0$, then $P(S_n > u) \le e^{-(1-\varepsilon)u^2/2V_n^2}$. This is almost such a good estimate as the estimate obtained by a formal application of the central limit theorem.

b) If $u \le 3V_n^2$, then $P(S_n > u) \le e^{-\text{const.}\, u^2/2V_n^2}$. This is a bound similar to that suggested by the central limit theorem, only it has a worse constant in the exponent.

c) If $u \gg V_n^2$, then

$$P(S_n > u) \le e^{-u}. \tag{2.2}$$

This is a very bad bound. In particular, it does not depend on the variance of the sum.

The question arises whether Bernstein's inequality can be improved in the 'bad case' $u \gg V_n^2$. To this question a positive answer can be given. Bennett's inequality formulated below yields a slight improvement of Bernstein's inequality in this case.

Bennett's inequality. *Let $\xi_1,\dots,\xi_n$ be a sequence of independent random variables which satisfy the relations $P(|\xi_j| \le 1) = 1$ and $EX_j = 0$, $1 \le j \le n$. Put $\sigma_j^2 = E\xi_j^2$, $1 \le j \le n$, $S_n = \sum\limits_{j=1}^n \xi_j$ and $V_n^2 = \operatorname{Var} S_n = \sum\limits_{j=1}^n \sigma_j^2$. Then the inequality*

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

holds for all numbers $u > 0$.

Hence there exists a constant $B = B(\varepsilon) > 0$ for all $\varepsilon > 0$ such that

$$P(S_n > u) \le \exp\left\{-(1-\varepsilon)u\log\frac{u}{V_n^2}\right\} \quad \textit{if } u > BV_n^2,$$

and there exists a number $K > 0$ such that

$$P(S_n > u) \le \exp\left\{-Ku\log\frac{u}{V_n^2}\right\} \quad \textit{if } u \ge 3V_n^2. \tag{2.3}$$

Formula (2.3) yields a slight improvement of formula (2.2), but even this bound is very far from the estimate suggested by the central limit theorem. On the other hand, as the next example shows, this estimate cannot be improved.

Lower bound for the distribution of sums of independent random variables in an appropriate example. *Let us fix a positive integer n together with two positive numbers u and σ^2 which satisfy the relations $0 < \sigma^2 \le \frac{1}{8}$, $n > 3u \ge 6$*

and $u > 3n\sigma^2$. Let us introduce the quantity $V_n^2 = n\sigma^2$, and consider a sequence of independent and identically distributed random variables $\xi_1, \dots, \xi_n$ such that $P(\xi_j = 1) = P(\xi_j = -1) = \frac{\sigma^2}{2}$ and $P(\xi_j = 0) = 1 - \sigma^2$. Put $S_n = \sum\limits_{j=1}^{n} \xi_j$. In this example $ES_n = 0$, $\operatorname{Var} S_n = V_n^2$, and

$$P(S_n \ge u) > \exp\left\{-Ku \log \frac{u}{V_n^2}\right\} \tag{2.4}$$

with some appropriate number $K > 0$.

In formula (2.4) that probability is bounded from below in a special case which is bounded from above in (2.3). These two estimates are very similar. The only difference between them is that they may contain a different constant $K > 0$. The above results can be summarized in the following way.

For small numbers $u > 0$ the probability $P(S_n > u)$ satisfies a good estimate suggested by the central limit theorem. Such a situation holds if $u \le \varepsilon V_n^2$. This probability satisfies a slightly weaker estimate for not too large numbers u (if $\varepsilon V_n^2 \le u \le C V_n^2$ with some fixed number $C > 0$), and it satisfies only very weak estimates for large numbers u (if $u \gg V_n^2$).

3. Some results useful in the study of the general case.

In the solution of Problem A) in the general case $k \ge 1$ similar results hold as in the special case $k = 1$ discussed before. To understand their similarity better it is useful to study first the following two questions.

Question 1. In the case $k = 1$ the sum of independent random variables with *zero expectation* was considered. What kind of normalization corresponds to this zero expectation in the case $k \ge 2$?

Question 2. In the case $k = 1$ the central limit theorem and the behaviour of the normal distribution function were in the background of the estimates. What kind of limit theorem and estimation take their part in the case $k \ge 2$?

Discussion of the first question.

It is useful to consider first the second moment of the expressions we are investigating. In the case $k = 1$ independent random variables of expectation zero are summed up. In this case the identity

$$\operatorname{Var}\left(\sum_{k=1}^{n} \xi_k\right) = \sum_{k=1}^{n} \operatorname{Var} \xi_k,$$

holds because of the identity $E\xi_i\xi_j = 0$ for all pairs $i \ne j$.

The multivariate version of this identity (in the case of U-statistics) would be the identity

$$\begin{aligned} I_{n,k}(f) &= \operatorname{Var}\left(\frac{1}{k!} \sum_{\substack{1\le j_s\le n,\ s=1,\dots,k,\\ j_s\ne j_{s'} \text{ if } s\ne s'}} f(\xi_{j_1},\dots,\xi_{j_k})\right) \\ &= \frac{1}{k!} \sum_{\substack{1\le j_s\le n,\ s=1,\dots,k,\\ j_s\ne j_{s'} \text{ if } s\ne s'}} \operatorname{Var} f(\xi_{j_1},\dots,\xi_{j_k}) \end{aligned}$$

This identity holds if

$$Ef(\xi_{j_1},\dots,\xi_{j_k})f(\xi_{j'_1},\dots,\xi_{j'_k})=0$$

for all pairs of k-tuples such that $\{j_1,\dots,j_k\}\ne\{j'_1,\dots,j'_k\}$. The above relation holds for the degenerate U-statistics introduced below.

Definition of degenerate U-statistics. *Take a U-statistic $I_{n,k}(f)$ determined by a sequence of independent and identically distributed random variables $\xi_1,\dots,\xi_n$ with distribution μ and a kernel function $f(x_1,\dots,x_k)$. This U-statistic is degenerate if*

$$E(f(\xi_1,\dots,\xi_k)|\xi_1=x_1,\dots,\xi_{j-1}=x_{j-1},\xi_{j+1}=x_{j+1},\dots,\xi_k=x_k)=0$$
$$\text{for all indices } 1\le j\le k \text{ and values } x_s\in X,\ s\in\{1,\dots,k\}\setminus\{j\}.$$

A U-statistic is degenerate if its kernel function is canonical, i.e. it satisfies the following property.

Definition of canonical functions. *A function $f(x_1,\dots,x_k)$ defined on the k-fold direct product $(X^k,\mathcal{X}^k)$ is canonical with respect to a probability measure μ on the space $(X,\mathcal{X})$ if*

$$\int f(x_1,\dots,x_{j-1},u,x_{j+1},\dots,x_k)\mu(\,du)=0$$
$$\text{for all indices } \ 1\le j\le k \text{ and values } x_s\in X,\ s\in\{1,\dots,k\}\setminus\{j\}.$$

The notion of degenerate U-statistics is useful, because in some sense such U-statistics behave so as sums of independent random variables with *expectation zero.* Beside this, the study of general U-statistics can be reduced to the study of degenerate U-statistics by means of the following Hoeffding-decomposition.

Hoeffding decomposition of general U-statistics. *All U-statistics $I_{n,k}(f)$ of order k can be written in the form of linear combination*

$$I_{n,k}(f)=\sum_{j=0}^{k} n^{k-j} I_{n,j}(f_j) \tag{3.1}$$

of degenerate U-statistics $I_{n,j}(f_j)$. *The (canonical) kernel functions f_j (of j variables) of the degenerate U-statistics $I_{n,j}(f_j)$, $0\le j\le k$, can be calculated explicitly. It can be shown that they satisfy the inequality*

$$\int f_j^2(x_1,\dots,x_j)\mu(\,dx_1)\dots\mu(\,dx_j)\le\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)$$

for all indices $0\le j\le k$.

The problems about the behaviour of the multiple random integrals $J_{n,k}(f)$ defined in formula (1.1) can also be reduced to problems about the behaviour of degenerate U-statistics by means of their appropriate decomposition. Such expressions can be written as the linear combination

$$J_{n,k}(f)=\sum_{j=0}^k c(n,j)n^{-j/2}I_{n,j}(f_j) \tag{3.2}$$

of degenerate U-statistics with the same kernel functions f_j which appear in formula (3.1) and with some appropriate coefficients $c(n,j)$ such that $c(n,j)<K(j)$ with some universal constant $K(j)$.

In the definition of the random integral $J_{n,k}(f)$ integration is taken with respect to the signed measure $\mu_n-\mu$, and this ‘normalization’ diminishes the value of the integral. This diminishing effect is reflected in the relatively small value of the coefficients $c(n,j)n^{-j/2}$ in formula (3.2).

Discussion of the second problem.

The previous results suggest that in the study of U-statistics and multiple random integrals the limit distributions of appropriately normalized degenerate U-statistics take the role of normal distributions. The limit theorems about degenerate U-statistics are known, and they can be formulated by means of multiple Wiener-Itô integrals with respect to a white noise. To formulate them first I recall the definition of white noise.

The notion of white noise. *Let a measure μ be given on some measurable space $(X,\mathcal{X})$. A system of jointly Gaussian random variables indexed by the measurable sets $A\subset X$ such that $\mu(A)<\infty$ is a white noise with reference measure μ if*

$$E\mu_W(A)\mu_W(B)=\mu(A\cap B)\quad\textit{and}\quad E\mu_W(A)=0$$

for all measurable sets $A,B\subset X$ such that $\mu(A)<\infty$ and $\mu(B)<\infty$.

If a white noise μ_W is given with some reference measure μ together with a function $f(x_1,\dots,x_k)$ square integrable with respect to the k-fold product μ^k of the measure μ, then the k-fold Wiener–Itô integral

$$Z_{\mu,k}(f)=\frac{1}{k!}\int f(x_1,\dots,x_k)\mu_W(\,dx_1)\dots\mu_W(\,dx_k) \tag{3.3}$$

of this function f with respect to the white noise μ_W can be defined in a natural way. (First this integral is defined for simple so-called step functions which take a constant value on finitely many rectangles, and disappear outside them. Then the integral can be extended to general functions by means of an appropriate L_2-isomorphism.)

The following result holds.

Limit distribution theorem for degenerate U-statistics. *Let us consider such a sequence $I_{n,k}(f)$, $n = k, k+1, \dots,$ of degenerate U-statistics which is determined by a sequence of independent and identically distributed random variables $\xi_1, \xi_2, \dots,$ on a measurable space $(X, \mathcal{X})$ with distribution μ and a (canonical) function $f(x_1, \dots, x_k)$ square integrable with respect to the measure μ^k. The normalized degenerate U-statistics $n^{-k/2} I_{n,k}(f)$ converge in distribution to the k-fold Wiener–Itô integral*

$$Z_{\mu,k}(f) = \frac{1}{k!} \int f(x_1, \dots, x_k) \mu_W(\,dx_1) \dots \mu_W(\,dx_k)$$

of the function f with respect to a white noise μ_W with reference measure μ if $n \to \infty$.

A heuristic explanation of the previous result.

If $I_{n,k}(f)$ is a degenerate U-statistics, then the identity

$$\begin{aligned} n^{-k/2} I_{n,k}(f) &= \frac{n^{k/2}}{k!} \int' f(x_1, \dots, x_k) \mu_n(\,dx_1) \dots \mu_n(\,dx_k) \\ &= \frac{n^{k/2}}{k!} \int' f(x_1, \dots, x_k) (\mu_n(\,dx_1) - \mu(\,dx_1)) \dots (\mu_n(\,dx_k) - \mu(\,dx_1)) \end{aligned}$$

holds, where μ_n denotes the empirical distribution of the sequence $\xi_1, \dots, \xi_n$. Beside this, the normalized empirical distributions $\sqrt{n}(\mu_n(\cdot) - \mu(\cdot))$ have a Gaussian limit as $n \to \infty$. This suggests a limiting procedure, and as more detailed considerations show (see e.g. [1]) the normalized empirical measures $\sqrt{n}(\mu_n(\cdot) - \mu(\cdot))$ can be replaced by the white noise $\mu_W(\cdot)$ in the limit process. The proof consists of the justification of this heuristic argument.

It is natural to extend the problems formulated in the introduction with their appropriate counterpart about Wiener–Itô integrals. Their solution indicates what kind of results can be expected in the original problems.

Let us consider the Wiener–Itô integral $Z_{\mu,k}(f)$ of a function $f(x_1, \dots, x_k)$ of k variables with respect to a white noise μ_W with reference measure μ introduced in formula (3.3) and study the following problems.

Problem A′′). Let us give a good estimate on the probability $P(Z_{\mu,k}(f) > u)$ for all numbers $u > 0$.

Problem B″). Let a nice class $\mathcal{F}$ of functions $f(x_1,\dots,x_k)$ of k variables be given. Take the Wiener–Itô integral $Z_{\mu,k}(f)$ of all functions $f\in\mathcal{F}$ with respect to a white noise μ_W. Give a good estimate on the distribution of the supremum of these random integrals, i.e. on the probability $P\left(\sup\limits_{f\in\mathcal{F}} Z_{\mu,k}(f)>u\right)$ for all numbers $u>0$.

4. Results about the distribution of random integrals and U-statistics.

It is worth considering first Problem A'') about the estimation of Wiener–Itô integrals. I present a result in this direction.

Estimation about the tail distribution of Wiener–Itô integrals. *Let a white noise μ_W be given with reference measure μ together with a function $f(x_1,\dots,x_k)$ of k variables on a measurable space $(X,\mathcal{X})$ such that*

$$\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\le\sigma^2$$

with some number $\sigma^2<\infty$. The Wiener–Itô integral

$$Z_{\mu,k}(f)=\frac{1}{k!}\int f(x_1,\dots,x_k)\mu_W(\,dx_1)\dots\mu_W(\,dx_k)$$

introduced in formula (3.3) satisfies the inequality

$$P(k!|Z_{\mu,k}(f)|>u)\le C\exp\left\{-\frac{1}{2}\left(\frac{u}{\sigma}\right)^{2/k}\right\}$$

for all numbers $u>0$ with some constant $C=C(k)>0$ depending only on the multiplicity k of the integral.

The next example shows that the above estimate is sharp.

Lower bound on the tail distribution of a special Wiener–Itô integral. *Let a σ-finite measure μ be given on a measurable space $(X,\mathcal{X})$ together with a white noise μ_W on $(X,\mathcal{X})$ with this reference measure μ. Let $f_0(x)$ be a real valued function on the space $(X,\mathcal{X})$ such that $\int f_0(x)^2\mu(\,dx)=1$. Let us introduce the function $f(x_1,\dots,x_k)=\sigma f_0(x_1)\cdots f_0(x_k)$ with some number $\sigma>0$, and consider the Wiener–Itô integral $Z_{\mu,k}(f)$ introduced in formula (3.3). Then the identity*

$$\int f(x_1,\dots,x_k)^2\,\mu(\,dx_1)\dots\,\mu(\,dx_k)=\sigma^2$$

holds, and the Wiener–Itô integral $Z_{\mu,k}(f)$ satisfies the inequality

$$P(k!|Z_{\mu,k}(f)|>u)\ge\frac{\bar{C}}{\left(\frac{u}{\sigma}\right)^{1/k}+1}\exp\left\{-\frac{1}{2}\left(\frac{u}{\sigma}\right)^{2/k}\right\}$$

for all numbers $u>0$ with some appropriate constant $\bar C>0$.

The integral $\sigma^2=\int f(x_1,\dots,x_k)^2\,\mu(\,dx_1)\dots\mu(\,dx_k)$ in the above results agrees with the variance of the random integral $(k!)^{-1/2}Z_{\mu,k}(f)$. Hence these results can be interpreted so that

$$P(Z_{k,\mu}(f)>u)\le \text{const.}\,P(\sigma\eta^k>u)$$

for all numbers $u>0$, where η is a standard normal random variable, and $\sigma^2=(k!)^{-1/2}EZ_{\mu,k}(f)^2$. Furthermore, this estimate is sharp.

Similar, but slightly weaker estimates hold for degenerate U-statistics and multiple random integrals with respect to normalized empirical distributions.

Estimate on the tail distribution of degenerate U-statistics. *Let $\xi_1,\dots,\xi_n$ be a sequence of independent and identically distributed random variables on a measurable space $(X,\mathcal X)$ with distribution μ. Take a function $f(x_1,\dots,x_k)$ on the space $(X^k,\mathcal X^k)$ canonical with respect to the measure μ which satisfies the conditions*

$$\|f\|_\infty=\sup_{x_j\in X,\,1\le j\le k}|f(x_1,\dots,x_k)|\le 1 \tag{4.1}$$

$$\|f\|_2^2=\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\le\sigma^2 \tag{4.2}$$

with some number $0<\sigma^2\le 1$, and consider the (degenerate) U-statistic defined in formula (1.2) with the help of these quantities. Then there exist some constants $A=A(k)>0$ and $B=B(k)>0$ depending only on the order k of the U-statistic such that the inequality

$$P(k!n^{-k/2}|I_{n,k}(f)|>u)\le A\exp\left\{-\frac{u^{2/k}}{2\sigma^{2/k}\left(1+B\left(un^{-k/2}\sigma^{-(k+1)}\right)^{1/k}\right)}\right\} \tag{4.3}$$

holds for all numbers $0\le u\le n^{k/2}\sigma^{k+1}$.

The above estimate can be considered as a multivariate generalization of Bernstein's inequality. For multiple integrals with respect to normalized empirical distributions the following similar estimate holds.

Estimate about the tail distribution of random integrals with respect to normalized empirical distributions. *Let a sequence $\xi_1,\dots,\xi_n$ of independent and identically distributed random variables be given with distribution μ which take their values on a measurable space $(X,\mathcal X)$ together with a function $f(x_1,\dots,x_k)$ on the k-fold product space $(X^k,\mathcal X^k)$ which satisfy relations (4.1) and (4.2) with some constant $0<\sigma\le 1$. Then there exist some constants $C=C_k>0$ and*

$\alpha = \alpha_k > 0$ *depending only on the multiplicity k of the integral $J_{n,k}(f)$ defined in formula (1.1) such that the inequality*

$$P\left(|J_{n,k}(f)| > u\right) \le C \exp\left\{-\alpha \left(\frac{u}{\sigma}\right)^{2/k}\right\} \quad \textit{for all numbers } 0 < u \le n^{k/2}\sigma^{k+1}$$

holds.

In the case $k = 1$ we have seen that the tail distribution $P(n^{-1/2}S_n > u)$ of the normalized sum $n^{-1/2}S_n$ of n independent, identically distributed, bounded random variables with expectation zero satisfies only a very weak estimate if $u \gg n^{1/2}\sigma^2$, an estimate which is very far from the bound suggested by the central limit theorem.

Similarly, in the case $k \ge 2$ the tail distribution $P(k!n^{-k/2}I_{n,k}(f) > u)$ of degenerate U-statistics satisfies a much weaker estimate than the bound suggested by the behaviour of Wiener–Itô integrals if $u \gg n^{k/2}\sigma^{k+1}$. This means that the previous estimates about the tail distribution of degenerate U-statistics and integrals with respect to normalized empirical distributions are sharp also in that sense that they give the domain where the sharp estimate suggested by the behaviour of Wiener–Itô integrals holds.

For the sake of completeness I present in the case $k = 2$ such a degenerate U-statistic whose tail-distribution satisfies only a much weaker estimate than formula (4.3) if $u \gg n\sigma^3$.

Lower bound for the tail distribution of a special degenerate U-statistic in the case $k = 2$. *Let $\xi_1, \dots, \xi_n$ be a sequence of independent, identically distributed random variables with values on the two-dimensional Euclidean space. Let $\xi_j = (\eta_{j,1}, \eta_{j,2})$, $1 \le j \le n$, where $\eta_{j,1}$ and $\eta_{j,2}$ are independent random variables, $P(\eta_{j,1} = 1) = P(\eta_{j,1} = -1) = \frac{\sigma^2}{8}$, and $P(\eta_{j,1} = 0) = 1 - \frac{\sigma^2}{4}$, $P(\eta_{j,2} = 1) = P(\eta_{j,2} = -1) = \frac{1}{2}$ for all indices $1 \le j \le n$. Let us introduce the function $f(x,y) = f((x_1,x_2),(y_1,y_2)) = x_1y_2 + x_2y_1$, $x = (x_1,x_2) \in R^2$, $y = (y_1,y_2) \in R^2$, and define the U-statistic of order 2*

$$I_{n,2}(f) = \sum_{1\le j,k\le n,\, j\ne k} (\eta_{j,1}\eta_{k,2} + \eta_{k,1}\eta_{j,2})$$

with this kernel function f and the independent random variables $\xi_1, \dots, \xi_n$. The expression $I_{n,2}(f)$ is a degenerate U-statistic. Furthermore, if $u \ge B_1 n\sigma^3$ with some appropriate constant $B_1 > 0$, $B_2^{-1}n \ge u \ge B_2 n^{-2}$ with some sufficiently large number $B_2 > 0$, and $\frac{1}{n} \le \sigma \le 1$, then the inequality

$$\begin{aligned} P(n^{-1}I_{n,2}(f) > u) &\ge \exp\left\{-Bn^{1/3}u^{2/3}\log\left(\frac{u}{n\sigma^3}\right)\right\} \\ &= \exp\left\{-B\frac{u}{\sigma}\left(\frac{n\sigma^3}{u}\right)^{1/3}\log\left(\frac{u}{n\sigma^3}\right)\right\} \end{aligned}$$

holds with some constant $B > 0$, which depends neither on the number n nor on the parameter σ.

5. A brief explanation of the results.

It is worth showing that the high even order moments $EI_{n,k}(f)^{2M}$ of a degenerate U-statistic $I_{n,k}(f)$ of order k satisfy such estimates as the moments $E\eta^{2kM}$ of a Gaussian random variable η with expectation zero and appropriate variance. Such estimates (together with the Markov inequality) imply the inequalities we want prove, and beside this there is a method which enables us to bound such moments.

Such moments can be estimated by means of the so-called diagram formula about random integrals. This formula makes possible to express the moments we are interested in as the sum of certain integrals defined with the help of some diagrams. To give a good estimate on the moments we want to bound it has to be shown that the "diagrams corresponding to the Gaussian effect" yield the main contribution to them. In such a way we can get an explanation why the tail distribution of degenerate U-statistics and random integrals satisfy such an estimate which the behaviour of Wiener–Itô integrals (i.e. the Gaussian case) suggests.

The explanation of the details in the estimation of multiple random integrals or degenerate U-statistics of order $k\ge 2$ demands the application of rather complicated notations. This requires much work which cannot be done in a short summary paper. Hence I omit its discussion. On the other hand, I consider a special case of this problem, the estimation of the moments of sums of independent random variables. This may explain very much also about the general case.

Let $\xi_1,\dots,\xi_n$ be a sequence of independent and identically distributed random variables such that $E\xi_1=0$, $\operatorname{Var}\xi_1=\sigma^2$, and let us estimate the even moments of the sum $S_n=\sum\limits_{j=1}^n \xi_j$. The identity

$$ES_n^{2M}=\sum_{\substack{(j_1,\dots,j_s,l_1,\dots,l_s)\\ j_1+\cdots+j_s=2M,\, j_u\ge2, \text{ for all indices } 1\le u\le s\\ l_u\ne l_{u'} \text{ if } u\ne u'}} E\xi_{l_1}^{j_1}\cdots E\xi_{l_s}^{j_s} \tag{5.1}$$

holds.

Simple combinatorial considerations show that in the sum at the right-hand side of the identity (5.1) most terms are indexed with such a vector

$$(j_1,\dots,j_M,l_1,\dots,l_M)$$

for which $j_u=2$ for all numbers $1\le u\le M$. The number of such terms equals $\binom{n}{M}\frac{(2M)!}{2^M}\sim n^M\frac{(2M)!}{2^M M!}$. Hence it is natural to expect that in typical cases $ES_n^{2M}\sim\left(n\sigma^2\right)^M\frac{(2M)!}{2^M M!}$. This consideration suggests the estimate

$$\sum_{1\le l_1<l_2<\cdots<l_M\le n} E\xi_{l_1}^2\cdots E\xi_{l_M}^2=\binom{n}{M}\frac{(2M)!}{2^M}\sigma^{2M}\sim\frac{(2M)!}{2^M M!}(n\sigma^2)^M=E\eta^{2M}$$

for the quantity ES_n^{2M}, where η is a Gaussian random variable with expectation zero and variance $\operatorname{Var} S_n$.

The above heuristic argument shows why we can expect such estimates for the moments of sums of independent random variables as in the Gaussian case. In nice cases this argument yields the right estimate.

But at working out the details some finer considerations have to be applied. It is not enough to give a good estimate on the number of the terms of different type, their magnitude has also to be taken into consideration. It is possible that relatively few summands with a large value yields the main contribution in the sum at the right-hand side of formula (5.1).

Let us consider for instance the following example. Let the terms of the sum be of the following form: $P(\xi_1 = 1) = P(\xi_1 = -1) = \frac{\sigma^2}{2}$, $P(\xi_1 = 0) = 1 - \sigma^2$. If σ^2 is very small and the number M is large, then

$$\sum_{j=1}^n E\xi_j^{2M} = n\sigma^2 \gg \sum_{1\le l_1<l_2<\cdots<l_M\le n} E\xi_{l_1}^2 \cdots E\xi_{l_M}^2 \sim \frac{(2M)!}{2^M M!} n^M \sigma^{2M}.$$

It can be seen by working out the details of the above example in the general case that we can get a good estimate for high moments of sums of independent random variables only if the variance of the summands is not too small, and not too high moments are bounded. The restriction we have to impose to give good moment estimates is related to the fact that the tail-distribution of degenerate U-statistics and random integrals satisfy good estimate only at not too high levels.

6. Estimation of the supremum of random integrals and U-statistics.

Let us consider the random integrals of a class of functions with respect to a normalized empirical distribution or a system of degenerate U-statistics defined with the help of a class of kernel functions and a sequence of independent and identically distributed random variables. We want to give a good estimate on the supremum of such integrals or U-statistics. It can be expected that in nice cases almost such a good estimate holds for this supremum as in the case when all but one random integrals or U-statistics are omitted from this class, and only the 'worst' integral or U-statistic is preserved, that one for which the weakest estimate holds. In the subsequent discussion such results will be given. To get them first the definition of good classes of functions has to be found for which good and substantial results hold.

It is useful to look for such classes of functions from which such a subclass with relatively few functions can be selected which is dense in the original class in an appropriate sense. The introduction of the following two notions proved to be useful.

Definition of L_2-dense classes of functions with respect to some measure. *Let a measurable space be $(Y, \mathcal{Y})$ be given together with a σ-finite measure ν and a class $\mathcal{G}$ of $\mathcal{Y}$-measurable, real valued functions on this space. This class of functions*

$\mathcal{G}$ is called an L_2-dense class with respect to ν with parameter D and exponent L if for all numbers $1 \ge \varepsilon > 0$ there exists a subclass $\mathcal{G}_\varepsilon = \{g_1, \dots, g_m\} \subset \mathcal{G}$ in the space $L_2(Y, \mathcal{Y}, \nu)$ consisting of $m \le D\varepsilon^{-L}$ elements such that $\inf\limits_{g_j \in \mathcal{G}_\varepsilon} \int |g - g_j|^2 \, d\nu < \varepsilon^2$ for all functions $g \in \mathcal{G}$.

The other useful notion is the following one.

Definition of L_2-dense classes of functions. *Let us have a measurable space $(Y, \mathcal{Y})$ and a set $\mathcal{G}$ of $\mathcal{Y}$-measurable real valued functions on this space. We call $\mathcal{G}$ an L_2-dense class of functions with parameter D and exponent L if it is L_2-dense with parameter D and exponent L with respect to all probability measures ν on $(Y, \mathcal{Y})$.*

It is useful to consider first Problem B$''$) about the supremum of Wiener–Itô integrals, then to describe the results on Problems B) and B$'$) about the supremum of random integrals with respect to normalized empirical distribution and degenerate U-statistics and to compare these results.

Estimate about the tail distribution of the supremum of Wiener–Itô integrals. *Let us consider a measurable space $(X, \mathcal{X})$ together with a σ-finite non-atomic measure μ on it, and let μ_W be a white noise with reference measure μ on $(X, \mathcal{X})$. Let $\mathcal{F}$ be a countable and L_2-dense class of functions $f(x_1, \dots, x_k)$ on $(X^k, \mathcal{X}^k)$ with some parameter D and exponent L with respect to the product measure μ^k such that*

$$\int f^2(x_1, \dots, x_k)\mu(\,dx_1)\dots\mu(\,dx_k) \le \sigma^2 \quad \text{with some } 0 < \sigma \le 1 \text{ for all } f \in \mathcal{F}.$$

Let us consider the multiple Wiener integrals $Z_{\mu,k}(f)$ introduced in formula (3.3) for all $f \in \mathcal{F}$. The inequality

$$P\left(\sup_{f \in \mathcal{F}} |Z_{\mu,k}(f)| > u\right) \le C(D+1)\exp\left\{-\alpha\left(\frac{u}{\sigma}\right)^{2/k}\right\}$$

holds with some universal constants $C = C(k) > 0$ and $\alpha = \alpha(k) > 0$ if

$$\left(\frac{u}{\sigma}\right)^{2/k} \ge ML\log\frac{2}{\sigma} \tag{6.1}$$

with some appropriate constant $M = M(k) > 0$.

In the above result — disregarding the value of the universal constants appearing in it — the same estimate is obtained about the tail distribution of Wiener–Itô integrals (under appropriate conditions) as in the estimate about the tail distribution of a single Wiener–Itô integral. The only essential difference between these two results is that in the present case an additional condition formulated in formula (6.1) had to be imposed. It is not difficult to present such an example which shows that such a condition is really needed. But here I omit its description.

The next result is an estimate on the tail-distribution of the supremum of random integrals $J_{n,k}(f)$ defined in formula (1.1).

Estimate on the tail distribution of the supremum of multiple integrals with respect to a normalized empirical distribution. *Let us have a probability measure μ on a measurable space $(X,\mathcal{X})$ together with a countable and L_2-dense class $\mathcal{F}$ of functions $f=f(x_1,\dots,x_k)$ of k variables with some parameter D and exponent L, $L\ge 1$, on the product space $(X^k,\mathcal{X}^k)$ such that*

$$\|f\|_\infty = \sup_{x_j\in X,\ 1\le j\le k} |f(x_1,\dots,x_k)| \le 1,$$

and

$$\|f\|_2^2 = Ef^2(\xi_1,\dots,\xi_k) = \int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k) \le \sigma^2$$

for all functions $f\in\mathcal{F}$ with some constant $0<\sigma\le 1$. Then there exist some constants $C=C(k)>0$, $\alpha=\alpha(k)>0$ and $M=M(k)>0$ depending only on the parameter k such that the supremum of the random integrals $J_{n,k}(f)$, $f\in\mathcal{F}$, defined by formula (1.1) satisfies the inequality

$$P\left(\sup_{f\in\mathcal{F}} |J_{n,k}(f)| \ge u\right) \le CD\exp\left\{-\alpha\left(\frac{u}{\sigma}\right)^{2/k}\right\},$$

provided that

$$n\sigma^2 \ge \left(\frac{u}{\sigma}\right)^{2/k} \ge M(L+\beta)^{3/2}\log\frac{2}{\sigma},$$

where $\beta=\max\left(\frac{\log D}{\log n},0\right)$ and the numbers D and L agree with the parameter and exponent of the L_2-dense class $\mathcal{F}$.

A similar estimate holds for the supremum of degenerate U-statistics $I_{n,k}(f)$, $f\in\mathcal{F}$. The only difference in comparison with the above result that in the case of the supremum of U-statistics the additional condition has to be imposed that the U-statistics $I_{n,k}(f)$ must be degenerate.

An essential difference between the results about the estimation of the supremum of Wiener–Itô integrals $Z_{\mu,k}(f)$ and integrals with respect to normalized empirical distribution $J_{n,k}(f)$ is that in the first case the class of functions $\mathcal{F}$ had to be L_2-dense with respect to the product measure μ^k, while in the second case a more restrictive condition was imposed. In the case of supremum of integrals with respect to a normalized empirical distribution the class of functions $\mathcal{F}$ had to satisfy the L_2-property, i.e. it had to be L_2-dense with respect to all probability measures. The question arises what the cause of this difference is.

The supremum of Wiener–Itô integrals can be bounded by means of a simple and natural method, the so-called ‘chaining argument’. In the case of the random integrals $J_{n,k}(f)$ this method is not strong enough to solve the problem, it only yields some partial results. To get a complete solution some additional methods have to be applied, and their application demands some additional restrictions.

The elaboration of the details would demand much work and the application of methods essentially different from previous ones. Hence I shall present only a brief sketch of the main ideas. The main emphasize will be put on the explanation of the main problems and methods. First I briefly explain the 'chaining argument'.

The 'chaining argument', and the boundary of this method.

Let us apply the notation worked out in the formulation of the results about the supremum of Wiener–Itô integrals. Let us take for all indices $N = 1, 2, \dots$ such a system of increasing subsets $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_N \subset \cdots \subset \mathcal{F}$ of the class of functions $\mathcal{F}$ with relatively small cardinalities which satisfy the relation

$$\inf_{g \in \mathcal{F}_N} \int (f(x_1, \dots, x_k) - g(x_1, \dots, x_k))^2 \, \mu(\,dx_1) \dots \mu(\,dx_k) \le 2^{-2N} \sigma^2.$$

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

The probabilities

$$P\left(\sup_{g \in \mathcal{F}_N} Z_{\mu,k}(g) > u\left(1 - 2^{-N}\right)\right)$$

can be well estimated by means of a recursion for $N = 1, 2, \dots,$ since for all functions $g \in \mathcal{F}_{N+1}$ there exists a function $g' \in \mathcal{F}_N$ (close to it) for which

$$\int (g(x_1, \dots, x_k) - g'(x_1, \dots, x_k))^2 \, \mu(\,dx_1) \dots \mu(\,dx_k) \le 2^{-2N} \sigma^2.$$

Hence the probability

$$P(|Z_{\mu,k}(g) - Z_{\mu,k}(g')| > 2^{-N} u) = P(|Z_{\mu,k}(g - g')| > 2^{-N} u)$$

can be well estimated by means of the previous results about the tail distribution Wiener–Itô integrals. By working out the details the result about the tail distribution of the supremum of Wiener–Itô integrals can be relatively simply proved.

The above considered 'chaining argument' is not strong enough to estimate the supremum of integrals with normalized empirical distribution or of degenerate U-statistics. It only makes possible to reduce the problem to the case when the expressions $\sigma^2(f) = \int f^2(x_1, \dots, x_k) \mu(\,dx_1) \dots \mu(\,dx_k)$ are small for all functions $f \in \mathcal{F}$.

The 'chaining argument' is a weak method in the study of this problem for the following reason.

There are only very weak estimates on probabilities of the form

$$P(I_{n,k}(f) > u) \quad \text{or} \quad P(J_{\mu,k}(f) > u)$$

if $\sigma^2(f)$ is very small, and the number u is relatively large. This is the consequence of the previously discussed fact that there cannot be given such a good estimate for the tail distribution of degenerate U-statistics or random integrals with respect to a normalized empirical distribution function with small variance as the Gaussian comparison would suggest.

This difficulty can be overcome by means of a different method, by means of a symmetrization argument. This method consists of reducing the estimation of a probability of the type

$$P\left(\frac{1}{k!}\sup_{f\in\mathcal{F}}\sum_{\substack{1\le j_s\le n,\ s=1,\dots,k,\\ j_s\ne j_{s'} \text{ if } s\ne s'}} f(\xi_{j_1},\dots,\xi_{j_k})>u\right)$$

to a probability of the type

$$P\left(\frac{1}{k!}\sup_{f\in\mathcal{F}}\sum_{\substack{1\le j_s\le n,\ s=1,\dots,k,\\ j_s\ne j_{s'} \text{ if } s\ne s'}} \varepsilon_{j_1}\dots\varepsilon_{j_k}f(\xi_{j_1},\dots,\xi_{j_k})>u\right), \tag{6.2}$$

where $\varepsilon_1,\dots,\varepsilon_n$ are independent random variables with distribution $P(\varepsilon_j=1)=P(\varepsilon_j=-1)=\frac{1}{2}$ for all indices $1\le j\le n$. Beside this, they are also independent of the random variables $\xi_1,\dots,\xi_n$.

The probabilities in formula (6.2) can be well estimated by means of a 'conditioning argument'. In the application of this method the conditional probability of the investigated event has to be bounded under the condition $\xi_1=x_1,\dots,\xi_n=x_n$ for all possible values $x_1,\dots,x_n$. There are good methods to estimate such type of conditional probabilities, but they are not discussed here. On the other hand, these methods work only if the class of functions $\mathcal{F}$ is L_2-dense. This is the reason why this property appears in this problem.

There is a point which should be emphasized even in this sketchy discussion of the problems. In the study of the supremum of random integrals with respect to a normalized empirical distribution or of degenerate U-statistics a different method was applied in the estimation of random variables with relatively large and small variance. In the case of relatively large variance the 'chaining argument' works, while in the case of small variance an appropriate symmetrization argument was applied. Behind the different approaches in these two cases there is a deeper reason.

The 'chaining argument' works well only in the study of the supremum of degenerate U-statistics or random integrals with respect to a normalized empirical measures with not too small variance; in the case when the U-statistics and random integrals satisfy such estimates which their 'Gaussian type limits' suggest. There can be defined some 'irregular events' whose appearance implies that the

U-statistics or random integrals take extremely large values. But in the case of random variables with not too small variances the probability of such irregular events is very small, and their effect can be disregarded. The case of U-statistics or random integrals with a small variance is different. In this case the probability of these irregularities (compared to the events of regular events) is relatively large, and in the estimation of the probabilities we are interested in their effect is dominant.

The ‘chaining argument’ works well in the case when the effect of the irregularities is negligible, and nice ‘Gaussian type estimates’ hold. On the other hand, if the effect of the irregularities is non-negligible, then it is useful to apply symmetrization type arguments.

In this work I tried to describe briefly the result of an important subject together with the heuristic picture behind the results. A more detailed discussion of this subject can be found in my work [1]. Beside this, I plan to publish a Lecture Note which also contains a complete discussion of the technical details. For the time being this Lecture Note [2] can be found only on my homepage. Both works [1] and [2] contain a more detailed list of references.

References

1.) Péter Major: Tail behaviour of multiple random integrals and U-statistics. *Probability Reviews.* 448–505, (2005)

2.) Péter Major: On the tail behavior of multiple random integrals and degenerate U-statistics. http://www.renyi.hu/˜major