ON THE TAIL BEHAVIOUR OF MULTIPLE RANDOM INTEGRALS AND DEGENERATE U-STATISTICS

Péter Major

Alfréd Rényi Mathematical Institute of the Hungarian Academy of Sciences

1. Introduction

In this work the following problem will be investigated: Fix a positive integer n, consider n independent, identically distributed random variables $\xi_1, \dots, \xi_n$ on a measurable space $(X, \mathcal{X})$ with some distribution μ and define their empirical distribution μ_n together with its normalization $\sqrt{n}(\mu_n - \mu)$. Take a function $f(x_1, \dots, x_k)$ of k variables on the k-fold product $(X^k, \mathcal{X}^k)$ of the space $(X, \mathcal{X})$, introduce also the k-th power of the normalized empirical measure $\sqrt{n}(\mu_n - \mu)$ on the space $(X^k, \mathcal{X}^k)$ and define the integral of the function f with respect to this signed product measure. This integral is a random variable, and for all $u > 0$ we want to give a good estimate on the probability that it is larger than u. More precisely, we take the integrals not on the whole space, we omit the diagonals $x_s = x_{s'}$, $1 \le s, s' \le k$, $s \ne s'$, of the space X^k from the domain of integration. Such a modification of the integral seems to be natural.

We shall also be interested in the following generalized version of the above problem. Let us have a nice class of functions $\mathcal{F}$ of k variables on the product space $(X^k, \mathcal{X}^k)$ and consider the integral of all functions of this class with respect to the k-fold direct product of our normalized empirical measure. Give a good estimate on the probability that the supremum of these integrals is larger than some number $u > 0$.

The reader may ask why the above problems deserve a closer study. I found them important, because they may help to solve some important problems in probability theory and mathematical statistics. I met such problems when tried to adapt the method of proof about the Gaussian limit behaviour of the maximum likelihood estimate to some other problems. In the original problem the asymptotic behaviour of the solution of the so-called maximum likelihood equation has to be investigated. The study of this equation is hard in its original form. But by making an appropriate Taylor expansion of the function whose root we are looking for and throwing away its higher order terms we get an approximation whose behaviour can be simply understood. So to describe the limit behaviour of the maximum likelihood estimate it suffices to show that this approximation causes only a negligible error.

One would try to apply a similar procedure in more difficult situations. I met some non-parametric maximum likelihood problems, for instance the description of the limit behaviour of the so-called Kaplan–Meyer product limit estimate when such an approach could be applied. But in those problems it was harder to justify that the simplifying approximation causes only a negligible error. To show this the solution of the above mentioned problems were needed. In the non-parametric maximum likelihood estimate problems I met the estimation of multiple (random) integrals played a role similar to the estimation of the coefficients in the Taylor expansion in the study of the maximum likelihood estimate. Although I could apply this approach only in some special cases, I believe that it works in very general situations. But it demands some further work to show this.

The problem suggested in this work is interesting and non-trivial even in the special case $k=1$. The solution of the problem in this case leads to some interesting, non-trivial generalization of the fundamental theorem of the mathematical statistics about the difference of the empirical and real distribution of a large sample.

The above mentioned problems have a natural counterpart about the behaviour of so-called U-statistics, a fairly popular subject in probability theory. The investigation of multiple random integrals and U-statistics are closely related, and it turned out that it is useful to consider them simultaneously. Hence both subjects will be discussed in this work.

Let us try to get some feeling what kind of results we can expect. It is useful to observe that for large sample size n the normalized empirical measure $\sqrt{n}(\mu_n-\mu)$ behaves similarly to a Gaussian random measure. This suggests that in the problems we are interested in similar results should hold as in the case of multiple Gaussian integrals. Hence we may expect that the tail behaviour of the distribution of a k-fold random integral with respect to a normalized empirical measure is similar to that of the k-th power of a Gaussian random variable with expectation zero and an appropriate variance. Moreover, a similar estimate should hold for the supremum of random integrals of a class of functions under not too restrictive conditions. We may also hope that the methods of the theory of multiple Gaussian integrals can be adapted to the investigation of our problems.

The above belief is essentially correct, but there is an essential difference between the behaviour of multiple Gaussian integrals and multiple integrals with respect to a normalized empirical measure. If the variance of a multiple integral with respect to a normalized empirical measure is small, what turns out to be equivalent to the small L_2-norm of the function we are integrating, then the behaviour of this integral is different from that of multiple Gaussian integrals with the same variance. In this case the effect of some irregularities of the normalized empirical distribution turns out to be non-negligible, and no good Gaussian approximation holds any longer. Hence some new methods have to be worked out and the hardest problems in our study appear at this point.

The precise formulation of the results will be contained in the main part of the work. Besides their proof I also try to explain the main ideas behind them and the notions introduced in their investigation. This work contains some new results, and also the proof of some already rather classical theorems is presented. To make the picture behind the problems more understandable I also discuss their Gaussian counterpart.

The proofs apply results from different parts of the probability theory. Papers investigating similar results refer to works dealing with quite different subjects, and this makes their reading rather hard. To overcome this difficulty I tried to work out the details and to present a self-contained discussion even at the price of a longer text. Thus I wrote down (in the main text or in the Appendix) the proof of many interesting and basic results, like results about Vapnik–Červonenkis classes, about U-statistics and their decomposition to sums of so-called degenerate U-statistics, logarithmic Sobolev inequalities, Borell's inequality about homogeneous polynomials of Rademacher functions, etc.

I tried to give such an exposition where different parts of the problem are explained as independently of other as possible, and they can be understood in themselves.

This work was explained at the probability seminar of the University Debrecen (Hungary).

2. Motivation of the investigation. Discussion of some problems

Here I try to show by means of some examples why the solution of the problems mentioned in the introduction may be useful in the study of some important probabilistic problems. I try to give a good picture about the main ideas but do not work out all details. Actually, the elaboration of some details omitted would demand hard work. But as the discussion of this section is quite independent of the rest of the paper, these omissions cause no problem in understanding the subsequent part.

I start with a short discussion of the maximum likelihood estimate in the simplest case. We study the following problem. Let us have a class of density functions $f(x,\vartheta)$ on the real line depending on a parameter $\vartheta\in R^1$ and observe a sequence of independent random variables $\xi_1(\omega),\dots,\xi_n(\omega)$ with a density function $f(x,\vartheta_0)$, where ϑ_0 is an unknown parameter we want to estimate with the help of the above sequence of random variables.

We can carry out this estimation with the help of the maximum likelihood method. It suggests to choose the estimate $\hat\vartheta_n=\hat\vartheta_n(\xi_1,\dots,\xi_n)$ of the parameter ϑ_0 as the number where the density function of the random vector $(\xi_1,\dots,\xi_n)$, i.e. the product

$$\prod_{k=1}^n f(\xi_k,\vartheta)=\exp\left\{\sum_{k=1}^n \log f(\xi_k,\vartheta)\right\}$$

takes its maximum. This point can be found as the solution of the so-called maximum likelihood equation

$$\sum_{k=1}^n \frac{\partial}{\partial\vartheta}\log f(\xi_k,\vartheta)=0. \tag{2.1}$$

We are interested in the asymptotic behaviour of the random variable $\hat\vartheta_n-\vartheta_0$, where $\hat\vartheta_n$ is the (appropriate) solution of the equation (2.1).

The direct study of this equation is rather hard, but a Taylor expansion of the expression at the left-hand side of (2.1) around the (unknown) point ϑ_0 helps to give a good and simple approximation of $\hat\vartheta_n$, and it enables us to describe the asymptotic behaviour of $\hat\vartheta_n-\vartheta_0$.

This Taylor expansion yields that

$$\begin{aligned}\sum_{k=1}^n \frac{\partial}{\partial\vartheta}\log f(\xi_k,\hat\vartheta_n)&=\sum_{k=1}^n\frac{\frac{\partial}{\partial\vartheta}f(\xi_k,\vartheta_0)}{f(\xi_k,\vartheta_0)}\\ &\quad+(\hat\vartheta_n-\vartheta_0)\left(\sum_{k=1}^n\left(\frac{\frac{\partial^2}{\partial\vartheta^2}f(\xi_k,\vartheta_0)}{f(\xi_k,\vartheta_0)}-\frac{\left(\frac{\partial}{\partial\vartheta}f(\xi_k,\vartheta_0)\right)^2}{f^2(\xi_k,\bar\vartheta_0)}\right)\right)+O\left(n(\hat\vartheta_n-\vartheta_0)^2\right)\end{aligned}$$

$$= \quad \sum_{k=1}^{n} \left(\eta_k + \zeta_k(\hat\vartheta_n - \vartheta_0)\right) + O\left(n(\hat\vartheta_n - \vartheta_0)^2\right), \tag{2.2}$$

where

$$\eta_k = \frac{\frac{\partial}{\partial\vartheta} f(\xi_k, \vartheta_0)}{f(\xi_k, \vartheta_0)} \quad \text{and} \quad \zeta_k = \frac{\frac{\partial^2}{\partial\vartheta^2} f(\xi_k, \vartheta_0)}{f(\xi_k, \vartheta_0)} - \frac{\left(\frac{\partial}{\partial\vartheta} f(\xi_k, \vartheta_0)\right)^2}{f^2(\xi_k, \bar\vartheta_0)}$$

for $k = 1, \dots, n$. We want to understand the asymptotic behaviour of the (random) expression on the right-hand side of (2.2). The relation

$$E\eta_k = \int \frac{\frac{\partial}{\partial\vartheta} f(x, \vartheta_0)}{f(x, \vartheta_0)} f(x, \vartheta_0)\, dx = \frac{\partial}{\partial\vartheta} \int f(x, \vartheta_0)\, dx = 0$$

holds, since $\int f(x, \vartheta)\, dx = 1$ for all ϑ, and differentiating this relation we get the last identity. Similarly, $E\eta_k^2 = -E\zeta_k = \int \frac{\left(\frac{\partial}{\partial\vartheta} f(x,\vartheta_0)\right)^2}{f(x,\vartheta_0)}\, dx > 0$, $k = 1, \dots, n$. Hence by the central limit theorem $\chi_n = \frac{1}{\sqrt{n}} \sum\limits_{k=1}^{n} \eta_k$ is asymptotically normal with expectation zero and variance $I^2 = \int \frac{\left(\frac{\partial}{\partial\vartheta} f(x,\vartheta_0)\right)^2}{f(x,\vartheta_0)}\, dx > 0$. In the statistics literature this number I is called the Fisher information. By the laws of large numbers $\frac{1}{n} \sum\limits_{k=1}^{n} \zeta_k \sim -I^2$.

Thus relation (2.2) suggests the approximation $\tilde\vartheta_n = -\frac{\sum\limits_{k=1}^{n} \eta_k}{\sum\limits_{k=1}^{n} \zeta_k}$ of the maximum-likelihood estimate $\hat\vartheta_n$, and $\sqrt{n}(\tilde\vartheta_n - \vartheta_0)$ is asymptotically normal with expectation zero and variance I^2. The random variable $\tilde\vartheta_n$ is not a solution of the equation (2.1), the value of the expression at the left-hand side is of order $O(n(\tilde\vartheta_n - \vartheta_0)^2) = O(1)$ in this point. On the other hand, the derivative of the function at the left-hand side is large in this point, it is greater than const. n with some const. > 0. This implies that the maximum-likelihood equation has a solution $\hat\vartheta_n$ such that $\hat\vartheta_n - \tilde\vartheta_n = O\left(\frac{1}{n}\right)$. This has the consequence that $\sqrt{n}(\hat\vartheta_n - \vartheta_0)$ and $\sqrt{n}(\tilde\vartheta_n - \vartheta_0)$ have the same asymptotic limit behaviour.

The previous method can be summarized in the following way: Take a simpler linearized version of the expression we want to estimate by means of an appropriate Taylor expansion, describe the limit distribution of this linearized version and show that the linearization causes only a negligible error.

We want to show that such a method also works in more difficult situations. But in some cases it is harder to show that the error we have committed by replacing the original expression by a simpler linearized version is negligible, and to do this we need the solution of the problems mentioned in the introduction. We shall present such an example by studying a fairly popular model of the mathematical statistics, the so-called Kaplan–Meyer method for the estimation of the empirical distribution function with the help of censored data.

The following problem is considered. Let (X_i, Z_i), $i = 1, \dots, n$, be a sequence of independent, identically distributed random vectors such that the components X_i and Z_i are also independent with distribution functions $F(x)$ and $G(x)$. We want to estimate the distribution function F of the random variables X_i, but we cannot observe the variables X_i, only the random variables $Y_i = \min(X_i, Z_i)$ and $\delta_i = I(X_i \le Z_i)$. In other words, we want to solve the following problem. There are certain objects whose lifetime X_i are independent and F distributed. But we cannot observe this lifetime X_i, because after a time Z_i the observation must be stopped. We also know whether the real lifetime X_i or the censoring variable Z_i was observed. We make n independent experiments and want to estimate with their help the distribution function F.

Kaplan and Meyer, on the basis of some maximum-likelihood estimation type considerations, proposed the following so-called product limit estimator $S_n(u)$ to estimate the unknown survival function $S = 1 - F$:

$$1 - F_n(u) = S_n(u) = \begin{cases} \prod_{i=1}^{n} \left(\dfrac{N(Y_i)}{N(Y_i)+1} \right)^{I(Y_i \le u, \delta_i = 1)} & \text{if } u \le \max(Y_1, \dots, Y_n) \\ 0 & \text{if } u \ge \max(Y_1, \dots, Y_n),\ \delta_n = 1, \\ \text{undefined} & \text{if } u \ge \max(Y_1, \dots, Y_n),\ \delta_n = 0, \end{cases} \tag{2.3}$$

where

$$N(t) = \#\{Y_i,\ \ Y_i > t,\ 1 \le i \le n\} = \sum_{i=1}^{n} I(Y_i > t).$$

We want to show that the above estimate (2.3) is really good. For this goal we shall approximate the random variables $S_n(u)$ by some appropriate random variables. To do this first we introduce some notations.

Put

$$\begin{aligned} &H(u) = P(Y_i \le u) = 1 - \bar{H}(u), \\ &\tilde{H}(u) = P(Y_i \le u,\, \delta_i = 1), \quad \tilde{\tilde{H}}(u) = P(Y_i \le u,\, \delta_i = 0) \end{aligned} \tag{2.4}$$

and

$$\begin{aligned} &H_n(u) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \le u) \\ &\tilde{H}_n(u) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \le u,\, \delta_i = 1), \quad \tilde{\tilde{H}}_n(u) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \le u,\, \delta_i = 0). \end{aligned} \tag{2.5}$$

Clearly $H(u) = \tilde{H}(u) + \tilde{\tilde{H}}(u)$ and $H_n(u) = \tilde{H}_n(u) + \tilde{\tilde{H}}_n(u)$. We shall estimate $F_n(u) - F(u)$ for $u \in (-\infty, T]$ if

$$1 - H(T) > \delta \quad \text{with some fixed } \delta > 0. \tag{2.6}$$

Condition (2.6) implies that there are more than $\frac{\delta}{2} n$ sample points Y_j larger than T with probability almost 1. It has exponentially small probability that this is not the case. This observation helps to show in the subsequent calculations that some events have negligibly small probability.

We introduce the so-called cumulative hazard function and its empirical version

$$\Lambda(u) = -\log(1 - F(u)), \quad \Lambda_n(u) = -\log(1 - F_n(u)). \tag{2.7}$$

Since $F_n(u) - F(u) = \exp(-\Lambda(u))\,(1 - \exp(\Lambda(u) - \Lambda_n(u)))$ a simple Taylor expansion yields

$$F_n(u) - F(u) = (1 - F(u))\,(\Lambda_n(u) - \Lambda(u)) + R_1(u), \tag{2.8}$$

and it is easy to see that $R_1(u) = O\left(\Lambda(u) - \Lambda_n(u))^2\right)$. It follows from the subsequent estimations that $\Lambda(u) - \Lambda_n(u) = O(n^{-1/2})$, thus $nR_1(u) = O(1)$. Hence it is enough to investigate the term $\Lambda_n(u)$. We shall show that $\Lambda_n(u)$ has an expansion with $\Lambda(u)$ as the main term plus $n^{-1/2}$ a term which is a linear functional of an appropriate normalized empirical distribution function plus an error term of order $O(n^{-1})$.

From (2.3) it is obvious that

$$\Lambda_n(u) = -\sum_{i=1}^{n} I(Y_i \le u,\, \delta_i = 1) \log\left(1 - \frac{1}{1 + N(Y_i)}\right).$$

We can get rid of the unpleasant logarithmic function in this formula by means of the relation $-\log(1 - x) = x + O(x^2)$ for small x which yields that

$$\Lambda_n(u) = \sum_{i=1}^{n} \frac{I(Y_i \le u,\, \delta_i = 1)}{N(Y_i)} + R_2(u) = \tilde{\Lambda}_n(u) + R_2(u), \tag{2.9}$$

and the error term $nR_2(u)$ is exponentially small.

The expression $\tilde{\Lambda}_n(u)$ is still inappropriate for our purposes. Since the denominators $N(Y_i) = \sum\limits_{j=1}^{n} I(Y_j > Y_i)$ are dependent for different indices i we cannot see directly the limit behaviour of $\tilde{\Lambda}_n(u)$.

We try to approximate $\tilde{\Lambda}_n(u)$ by a simpler expression. A natural approach would be to approximate the terms $N(Y_i)$ in it by their conditional expectation $(n-1)\bar{H}(Y_i) = (n-1)(1 - H(Y_i)) = E(N(Y_i)|Y_i)$. This is a too rough 'first order' approximation, but the following 'second order approximation' will be sufficient for our goals. Put

$$N(Y_i) = \sum_{j=1}^{n} I(Y_j > Y_i) = n\bar{H}(Y_i)\left(1 + \frac{\sum\limits_{j=1}^{n} I(Y_j > Y_i) - n\bar{H}(Y_i)}{n\bar{H}(Y_i)}\right)$$

and express the terms $\frac{1}{N(Y_i)}$ in the sum defining $\tilde{\Lambda}_n$ by means of the relation $\frac{1}{1+z} = \sum\limits_{k=0}^{\infty} (-1)^k z^k = 1 - z + \varepsilon(z)$ with the choice $z = \frac{\sum_{j=1}^{n} I(Y_j > Y_i) - n\bar{H}(Y_i)}{n\bar{H}(Y_i)}$. As $|\varepsilon(z)| < 2z^2$ for

$|z| < \frac{1}{2}$ we get that

$$
\begin{aligned}
\tilde{\Lambda}_n(u) &= \sum_{i=1}^n \frac{I(Y_i \le u,\, \delta_i = 1)}{n\bar{H}(Y_i)} \left(1 + \sum_{k=1}^\infty \left(- \frac{\sum\limits_{j=1}^n I(Y_j > Y_i) - n\bar{H}(Y_i)}{n\bar{H}(Y_i)} \right)^k \right) \\
&= \sum_{i=1}^n \frac{I(Y_i \le u,\, \delta_i = 1)}{n\bar{H}(Y_i)} \left(1 - \frac{\sum\limits_{j=1}^n I(Y_j > Y_i) - n\bar{H}(Y_i)}{n\bar{H}(Y_i)} \right) + R_3(u) \\
&= 2A(u) - B(u) + R_3(u),
\end{aligned} \tag{2.10}
$$

where

$$
A(u) = A(n,u) = \sum_{i=1}^n \frac{I(Y_i \le u,\, \delta_i = 1)}{n\bar{H}(Y_i)}
$$

and

$$
B(u) = B(n,u) = \sum_{i=1}^n \sum_{j=1}^n \frac{I(Y_i \le u,\, \delta_i = 1) I(Y_j > Y_i)}{n^2 \bar{H}^2(Y_i)}.
$$

It can be proved by means of standard methods that $nR_3(u)$ is exponentially small. Thus from (2.9) and (2.10) we get that

$$
\Lambda_n(u) = 2A(u) - B(u) + \text{negligible error}. \tag{2.11}
$$

This means that to solve our problem we have to describe the asymptotic behaviour of the random variables $A(u)$ and $B(u)$. We can get a better insight into their behaviour by rewriting the sum $A(u)$ as an integral with respect to an empirical measure and the double sum $B(u)$ as a two-fold integral with respect empirical measures. These integrals can be rewritten as sums of random integrals with respect to normalized empirical measures and deterministic measures. In such a way we get a representation of $\Lambda_n(u)$ in the form of a sum whose terms can be well understood.

Let us write

$$
\begin{aligned}
A(u) &= \int_{-\infty}^{+\infty} \frac{I(y \le u)}{1 - H(y)}\, d\tilde{H}_n(y), \\
B(u) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{I(y \le u) I(x > y)}{(1 - H(y))^2}\, dH_n(x) d\tilde{H}_n(y).
\end{aligned}
$$

To rewrite the term $B(u)$ in a form better for our purposes observe that

$$
\begin{aligned}
H_n(x)\tilde{H}_n(y) = H(x)\tilde{H}(y) &+ H(x)(\tilde{H}_n(y) - \tilde{H}(y)) + (H_n(x) - H(x))\tilde{H}(y) \\
&+ (H_n(x) - H(x))(\tilde{H}_n(y) - \tilde{H}(y)).
\end{aligned}
$$

Hence it can be written in the form $B(u) = B_1(u) + B_2(u) + B_3(u) + B_4(u)$, where

$$B_1(u) = \int_{-\infty}^{u} \int_{-\infty}^{+\infty} \frac{I(x > y)}{(1 - H(y))^2} \, dH(x) \, d\tilde{H}(y) \;,$$

$$B_2(u) = \frac{1}{\sqrt{n}} \int_{-\infty}^{u} \int_{-\infty}^{+\infty} \frac{I(x > y)}{(1 - H(y))^2} \, dH(x) \, d\left(\sqrt{n}(\tilde{H}_n(y) - \tilde{H}(y))\right) ,$$

$$B_3(u) = \frac{1}{\sqrt{n}} \int_{-\infty}^{u} \int_{-\infty}^{+\infty} \frac{I(x > y)}{(1 - H(y))^2} \, d\left(\sqrt{n}\,(H_n(x) - H(x))\right) \, d\tilde{H}(y) \;,$$

$$B_4(u) = \frac{1}{n} \int_{-\infty}^{u} \int_{-\infty}^{+\infty} \frac{I(x > y)}{(1 - H(y))^2} \, d\left(\sqrt{n}\,(H_n(x) - H(x))\right) \, d\left(\sqrt{n}(\tilde{H}_n(y) - \tilde{H}(y))\right) .$$

In the above decomposition of $B(u)$ the term B_1 is a deterministic function, B_2, B_3 are linear functionals of empirical processes and B_4 is a nonlinear functional of empirical processes. The deterministic term $B_1(u)$ can be calculated explicitly. Indeed,

$$B_1(u) = \int_{-\infty}^{u} \int_{-\infty}^{+\infty} \frac{I(x > y)}{(1 - H(y))^2} \, dH(x) d\tilde{H}(y) = \int_{-\infty}^{u} \frac{d\tilde{H}(y)}{1 - H(y)}.$$

Then the relations $\tilde{H}(u) = \int_{-\infty}^{u} (1 - G(t)) \, dF(t)$ and $1 - H = (1 - F)(1 - G)$ imply that

$$B_1(u) = \int_{-\infty}^{u} \frac{dF(y)}{1 - F(y)} = - \log(1 - F(u)) = \Lambda(u). \tag{2.12}$$

Observe that

$$\begin{aligned} A(u) &= \int_{-\infty}^{u} \frac{d\,\tilde{H}_n(y)}{1 - H(y)} \\ &= \int_{-\infty}^{u} \frac{d\tilde{H}(y)}{1 - H(y)} + \frac{1}{\sqrt{n}} \int_{-\infty}^{u} \frac{d\left(\sqrt{n}(\tilde{H}_n(y) - \tilde{H}(y))\right)}{1 - H(y)} \\ &= B_1(u) + B_2(u). \end{aligned} \tag{2.13}$$

From relation (2.11) using (2.12) and (2.13) it follows that

$$\Lambda_n(u) - \Lambda(u) = B_2(u) - B_3(u) - B_4(u) + \text{negligible error}. \tag{2.14}$$

Integrating B_2 and B_3 in the variable x and then integrating by parts B_2 we get that

$$\begin{aligned} B_2(u) &= \frac{1}{\sqrt{n}} \int_{-\infty}^{u} \frac{d\left(\sqrt{n}(\tilde{H}_n(y) - \tilde{H}(y))\right)}{1 - H(y)} \\ &= \frac{\sqrt{n}\left(\tilde{H}_n(u) - \tilde{H}(u)\right)}{\sqrt{n}(1 - H(u))} - \frac{1}{\sqrt{n}} \int_{-\infty}^{u} \frac{\sqrt{n}(\tilde{H}_n(y) - \tilde{H}(y))}{(1 - H(y))^2} \, dH(y) \\ B_3(u) &= \frac{1}{\sqrt{n}} \int_{-\infty}^{u} \frac{\sqrt{n}\,(H_n(y) - H(y))}{(1 - H(y))^2} \, d\tilde{H}(y). \end{aligned}$$

Using the above forms of B_2 and B_3, (2.12) we can write

$$\begin{aligned}\sqrt{n}\,(\Lambda_n(u)-\Lambda(u)) &= \frac{\sqrt{n}\left(\tilde H_n(u)-\tilde H(u)\right)}{1-H(u)} - \int_{-\infty}^{u} \frac{\sqrt{n}(\tilde H_n(y)-\tilde H(y))}{(1-H(y))^2}\,dH(y) \\ &\quad + \int_{-\infty}^{u} \frac{\sqrt{n}\,(H_n(y)-H(y))}{(1-H(y))^2}\,d\tilde H(y) \\ &\quad - \sqrt{n}B_4(u) + \text{negligible error.}\end{aligned} \tag{2.15}$$

Formula (2.15) almost agrees with the statement we wanted to prove. Here we expressed the normalized error $\sqrt{n}\,(\Lambda_n(u)-\Lambda(u))$ as a sum of linear functionals of normalized empirical measures plus some negligible error terms and the error term $\sqrt{n}B_4(u)$. So to get a complete proof it is enough to show that $\sqrt{n}B_4(u)$ also yields a negligible error. But $B_4(u)$ is a double integral of a bounded function (here we apply again formula (2.6)) with respect to a normalized empirical measure. Hence to bound this term we need a good estimate of multiple stochastic integrals (with multiplicity 2) and this is just the problem formulated in the introduction. The estimate we need here follows from Theorem 8.1 of the present work. Let us remark that the problem discussed here corresponds to the estimation of the coefficient of the second term in the Taylor expansion considered in the study of the maximum likelihood estimation. One may worry a little bit how to bound $B_4(u)$ with the help of estimations of double stochastic integrals, since in the definition of $B_4(u)$ we integrate by different normalized empirical processes in the two coordinates. But this is a not too difficult technical problem, it can be simply overcome for instance by rewriting the integral as a double integral with respect to the empirical process $\left(\sqrt{n}\,(H_n(x)-H(x)),\sqrt{n}\left(\tilde H_n(y)-\tilde H(y)\right)\right)$ in the space R^2.

By working out the details of the above calculation we get that the linear functional $B_2(u)-B_3(u)$ of normalized empirical processes yields a good estimate on the expression $\sqrt{n}(\Lambda_n(u)-\Lambda(u))$ for a fixed parameter u. But we want to prove somewhat more, we want to get an estimate uniform in the parameter u, i.e. to show that even the random variable $\sup\limits_{u\le T}|\sqrt{n}(\Lambda_n(u)-\Lambda(u))-B_2(u)+B_3(u)|$ is small. This can be done by making estimates uniform in the parameter u in all steps of the above calculation. There appears only one difficulty when trying to carry out this program. Namely, we need an estimate on $\sup\limits_u|B_4(u)|$, i.e. we have to bound the supremum of multiple random integrals with respect to a normalized random measure for a nice class of kernel functions. This can be done, but at this point the second problem mentioned in the introduction appears. This difficulty can be overcome by means of Theorem 8.2 of this work.

Thus we can find the limit behaviour of the Kaplan–Meyer estimate by means of an appropriate expansion. The steps of this investigation are fairly standard, the only hard part is the solution of the problems mentioned in the introduction. We expect that such a method also works in much more general situation. This may justify a detailed study of the problems considered in this work.

I finish this section with a remark of Richard Gill he made in a personal conversation after my talk on this subject at a conference. He told that this approach had given a complete proof about the limit behaviour of this estimate, but it had exploited the explicit formula given in the Kaplan–Meyer estimate. He missed the application of an argument based on the non-parametric maximum likelihood character of this estimate. This was a completely justified remark, since if we do not restrict our attention to this problem, but try to generalize it to general non-parametric maximum likelihood estimates, then we have to understand how the maximum likelihood character can be exploited. I believe that this can be done, but it demands further studies.

3. Some estimates about sums of independent random variables

We need some results about the distribution of sums of independent random variables bounded by a constant with probability one. Later only the results about sums of independent and identically distributed variables will be interesting for us, but since these results can be generalized without any effort to sums of not necessarily identically distributed random variables here we shall drop the condition about the identical distribution of the summands. We are interested in the question when these estimates give such a good bound as the central limit theorem suggests, and what can be told if this is not the case. More explicitly, we consider the following problem: Let $X_1,\dots,X_n$ be independent random variables $EX_j=0$, $\operatorname{Var} X_j=\sigma_j^2$, $1\le j\le n$, and take the random sum $S_n=\sum\limits_{j=1}^n X_j$ and its variance $\operatorname{Var} S_n=V_n^2=\sum\limits_{j=1}^n\sigma_j^2$. We want to get a good bound on the probability $P(S_n>xV_n)$. The central limit theorem would suggest that under general conditions an upper bound of the order $1-\Phi(x)$ should hold for this probability where $\Phi(x)$ denotes the standard normal distribution function. Since the standard normal distribution function satisfies the inequality $\left(\frac1x-\frac1{x^3}\right)\frac{e^{-x^2/2}}{\sqrt{2\pi}}<1-\Phi(x)<\frac1x\frac{e^{-x^2/2}}{\sqrt{2\pi}}$ for all $x>0$ it is natural to ask when the probability $P(S_n>xV_n)$ is comparable with the value $e^{-x^2/2}$. More generally, we say that we have a Gaussian type estimate for the probability $P(S_n>xV_n)$ if it can be bounded by e^{-Cx^2} with some constant C separated from zero.

First we discuss Bernstein's inequality which tells for which values x the probability $P(S_n>xV_n)$ satisfies a Gaussian type estimate. Such an estimate holds (for sums of random variables bounded by 1) if $x\le \text{const.}\,V_n$. For $x\ge\text{const.}\,V_n$ Bernstein's inequality yields almost no improvement if we have a better bound on the variance V_n of the sum S_n. Another estimate, Bennett's inequality yields a slight improvement, and as an example presented before this result shows it cannot be essentially improved without imposing some additional conditions. The main difficulties we meet in this paper are closely related to the weakness of the estimates we have for the probability of the event that a sum of independent random variables is larger than some value when this probability does not satisfy a Gaussian type estimate because of the small variance of the sum.

Let us formulate Bernstein's inequality. In its usual formulation a real number M is introduced and it is assumed that the terms in the sum we investigate are bounded by this number. But since the problem can be simply reduced to the special case $M=1$ we shall only deal with this special case.

Theorem 3.1 (Bernstein's inequality). *Let $X_1,\dots,X_n$ be independent random variables, $P(|X_j|\le 1)=1$, $EX_j=0$, $1\le j\le n$. Put $\sigma_j^2=EX_j^2$, $1\le j\le n$, $S_n=\sum\limits_{j=1}^n X_j$ and $V_n^2=\operatorname{Var} S_n=\sum\limits_{j=1}^n\sigma_j^2$. Then*

$$P\left(S_n>xV_n\right)\le\exp\left\{-\frac{x^2}{2\left(1+\frac13\frac{x}{V_n}\right)}\right\}\qquad \textit{for all } x>0. \tag{3.1}$$

Proof of Theorem 3.1. Let us give a good bound on the exponential moments Ee^{tS_n} for some appropriate parameters $t>0$. We can write $Ee^{tX_j}=\sum\limits_{k=0}^\infty\frac{t^k}{k!}EX_j^k\le 1+\frac{t^2\sigma_j^2}{2}\left(1+\sum\limits_{k=1}^\infty\frac{2t^k}{(k+2)!}\right)\le 1+\frac{t^2\sigma_j^2}{2}\left(1+\sum\limits_{k=1}^\infty 3^{-k}t^k\right)=1+\frac{t^2\sigma_j^2}{2}\frac{1}{1-\frac t3}\le\exp\left\{\frac{t^2\sigma_j^2}{2}\frac{1}{1-\frac t3}\right\}$ if $0\le t<3$. Hence $Ee^{tS_n}=\prod\limits_{j=1}^n Ee^{tX_j}\le\exp\left\{\frac{t^2V_n^2}{2}\frac{1}{1-\frac t3}\right\}$ for $0\le t<3$.

The above relation implies that

$$P\left(S_n>xV_n\right)=P(e^{tS_n}>e^{txV_n})\le Ee^{tS_n}e^{-txV_n}\le\exp\left\{\frac{t^2V_n^2}{2}\frac{1}{1-\frac t3}-txV_n\right\}$$

if $0\le t<3$. Choose the number t in this inequality as the solution of the equation $t^2V_n^2\frac{1}{1-\frac t3}=txV_n$, i.e. put $t=\frac{x}{V_n+\frac x3}$. Then $0\le t<3$, and we get that $P(S_n>xV_n)\le e^{-txV_n/2}=\exp\left\{-\frac{x^2}{2\left(1+\frac13\frac{x}{V_n}\right)}\right\}$.

If the random variables $X_1,\dots,X_n$ satisfy the conditions of the Bernstein inequality then also the random variables $-X_1,\dots,-X_n$ satisfy them. By applying the above result in both cases we get that $P(|S_n|>xV_n)\le 2\exp\left\{-\frac{x^2}{2\left(1+\frac13\frac{x}{V_n}\right)}\right\}$ under the conditions of the Bernstein inequality.

Bernstein's inequality states that for all $\varepsilon>0$ there is some sufficiently small number $\alpha(\varepsilon)>0$ such that in the case $\frac{x}{V_n}<\alpha(\varepsilon)$ $P(S_n>xV_n)\le e^{-(1-\varepsilon)x^2/2}$. Besides, for all fixed numbers $A>0$ there is some constant $C=C(A)>0$ such that in the case $\frac{x}{V_n}<A$ the inequality $P(S_n>xV_n)\le e^{-Cx^2}$ holds. This can be interpreted as a Gaussian type estimate for the probability $P(S_n>xV_n)$.

On the other hand, if $\frac{x}{V_n}$ is very large, then the Bernstein inequality yields a much worse estimate. The next example shows that this is not because of its weakness. There are sequences of independent, identically distributed random variables $X_1,\dots,X_n$ bounded by one and with expectation zero such that with the notations $S_n=\sum\limits_{j=1}^n X_j$, $\sigma^2=EX_j^2$, $V_n^2=\sum\limits_{j=1}EX_j^2=n\sigma^2$ the probability $P(S_n>xV_n)$ is relatively large if $\frac{x}{V_n}$ is large, it is much larger than the value suggested by the normal approximation. This example will be interesting for us mainly for the sake of some orientation. Hence I do not try to formulate it in such a general form as it could be done or to give the best possible constants in it. The method of proof shows that a wide class of examples could be constructed with similar properties. In the following discussion it will be convenient to replace the number x by $y = xV_n = \sqrt{n}\sigma x$.

Example 3.2. *Let us fix some positive integer n, real numbers $y \ge 200$ and $1 > \sigma^2 > 0$ such that $n > 16y > 64n\sigma^2$. Put $V_n^2 = n\sigma^2$ and take a sequence of independent, identically distributed random variables $X_1, \dots, X_n$ such that $P(X_j = 1) = P(X_j = -1) = \frac{\sigma^2}{2}$, and $P(X_j = 0) = 1 - \sigma^2$. Put $S_n = \sum\limits_{j=1}^n X_j$. Then $ES_n = 0$, $\operatorname{Var} S_n = V_n^2$, and*

$$P(S_n > y) > A \exp\left\{-By \log \frac{y}{V_n^2}\right\}$$

with some universal constants $A > 0$ and $B > 0$. We can choose for instance $A = \frac{1}{2}$, $B = \frac{22}{5}$ in this inequality.

Here I shall give a proof of the statement of Example 3.2. Let me remark that in the work [23] I gave a simpler and more elementary proof of this result under the name Example 2.4.

Proof of the statement of Example 3.2. In the proof some ideas of the large deviation theory will be applied. Let us introduce the measure μ, $\mu(\{1\}) = \mu(\{-1\}) = \frac{\sigma^2}{2}$, $\mu(\{0\}) = 1 - \sigma^2$ on the real line, which is actually the distribution of the random variables X_j, together with its conjugates μ_t, $\mu_t(\,dx) = \frac{e^{tx}}{\frac{\sigma^2}{2}(e^t + e^{-t}) + 1 - \sigma^2}\mu(\,dx)$, $x \in R^1$, for all real numbers t. Let $\mu^{(n)}$ denote the n-fold convolution of the measure μ and $\mu_t^{(n)}$ the n-fold convolution of the measure μ_t with itself. Then $P(S_n > y) = \mu^{(n)}((y, \infty))$, and it is not difficult to see (and it is a well-known fact in the theory of large deviations) that $\mu^{(n)}(A) = \left(\frac{\sigma^2}{2}(e^t + e^{-t}) + 1 - \sigma^2\right)^n \int_A e^{-tu} \mu_t^{(n)}(\,du)$ for all measurable sets $A \subset R^1$.

Let us consider the above defined measures μ_t and $\mu_t^{(n)}$ with $t = \log \frac{4y}{n\sigma^2}$. I claim that $\mu_t^{(n)}([y, \frac{11}{5}y]) \ge \frac{1}{2}$. To show this let us consider n independent μ_t distributed, independent random variables $\xi_1, \dots, \xi_n$, and estimate their expected value and variance. We have $E\xi_j = \frac{\frac{\sigma^2}{2}(e^t - e^{-t})}{\frac{\sigma^2}{2}(e^t + e^{-t}) + 1 - \sigma^2}$ for all $1 \le j \le n$, and since $1 \le \frac{\sigma^2}{2}(e^t + e^{-t}) + 1 - \sigma^2 \le 1 + \sigma^2 e^t = 1 + 4\frac{y}{n} \le \frac{5}{4}$, and besides, we get with the help of the estimate $e^{-t} = e^t e^{-2t} = e^t \left(\frac{n\sigma^2}{4y}\right)^2 \le \frac{1}{4}e^t$ the inequality $\frac{3}{2}\frac{y}{n} = \frac{3}{8}\sigma^2 e^t \le \frac{\sigma^2}{2}(e^t - e^{-t}) \le \frac{\sigma^2}{2}e^t = 2\frac{y}{n}$, hence $\frac{6}{5}\frac{y}{n} \le E\xi_j \le 2\frac{y}{n}$. Similarly, $\operatorname{Var} \xi_j \le E\xi_j^2 = \frac{\frac{\sigma^2}{2}(e^t + e^{-t})}{\frac{\sigma^2}{2}(e^t + e^{-t}) + 1 - \sigma^2} \le 4\frac{y}{n}$. The above estimates together with the Chebishev inequality imply that $\mu_t^{(n)}([y, \frac{11}{5}y]) = P\left(y \le \sum\limits_{j=1}^n \xi_j \le \frac{11}{5}y\right) \ge 1 - P\left(\left|\sum\limits_{j=1}^n (\xi_j - E\xi_j)\right| > \frac{y}{5}\right) \ge 1 - \frac{100y}{y^2} \ge \frac{1}{2}$. This inequality together with the relation between the measures $\mu^{(n)}$ and $\mu_t^{(n)}$ imply that

$$
\begin{aligned}
P(S_n > y) &= \mu^{(n)}([y, \infty]) \ge \mu^{(n)}\left(\left[y, \frac{11}{5}y\right]\right) \\
&= \left(\frac{\sigma^2}{2}(e^t + e^{-t}) + 1 - \sigma^2\right)^n \int_y^{11y/5} e^{-tu}\mu_t^{(n)}(\,du) \ge e^{-11ty/5}\mu_t^{(n)}\left(\left[y, \frac{11}{5}y\right]\right) \\
&\ge \frac{1}{2}e^{-11ty/5} = \frac{1}{2}\exp\left\{-\frac{11}{5}y\log\frac{4y}{V_n^2}\right\} \ge \frac{1}{2}\exp\left\{-\frac{22}{5}y\log\frac{y}{V_n^2}\right\}.
\end{aligned}
$$

In the case $y > V_n^2$ the Bernstein inequality yields the estimate $P(S_n > y) \le e^{-\alpha y}$ with some universal constant $\alpha > 0$, while the above example shows that we can expect at most an additional logarithmic factor in the exponent of the upper bound in an improvement of this estimate. The following result, called Bennett's inequality shows that such an improvement is really possible.

Theorem 3.3 (Bennett's inequality). *Let $X_1, \dots, X_n$ be independent random variables, $P(|X_j| \le 1) = 1$, $EX_j = 0$, $1 \le j \le n$. Put $\sigma_j^2 = EX_j^2$, $1 \le j \le n$, $S_n = \sum\limits_{j=1}^n X_j$ and $V_n^2 = \operatorname{Var} S_n = \sum\limits_{j=1}^n \sigma_j^2$. Then*

$$
P(S_n > y) \le \exp\left\{-V_n^2\left[\left(1 + \frac{y}{V_n^2}\right)\log\left(1 + \frac{y}{V_n^2}\right) - \frac{y}{V_n^2}\right]\right\} \quad \textit{for all } y > 0. \tag{3.2}
$$

As a consequence, for all $\varepsilon > 0$ there exists some $B = B(\varepsilon) > 0$ such that

$$
P\left(S_n > y\right) \le \exp\left\{-(1-\varepsilon)y\log\frac{y}{V_n^2}\right\} \quad \textit{if } y > BV_n^2, \tag{3.3}
$$

and there exists some positive constant $K > 0$ such that

$$
P\left(S_n > y\right) \le \exp\left\{-Ky\log\frac{y}{V_n^2}\right\} \quad \textit{if } y > 2V_n^2. \tag{3.4}
$$

Proof of Theorem 3.3. We have

$$
Ee^{tX_j} = \sum_{k=0}^{\infty}\frac{t^k}{k!}EX_j^k \le 1 + \sigma_j^2\sum_{k=2}^{\infty}\frac{t^k}{k!} = 1 + \sigma_j^2\left(e^t - 1 - t\right) \le e^{\sigma_j^2(e^t-1-t)}, \quad 1 \le j \le n,
$$

and $Ee^{tS_n} \le e^{V_n^2(e^t-1-t)}$ for all $t \ge 0$. Hence $P(S_n > y) \le e^{-ty}Ee^{tS_n} \le e^{-ty+V_n^2(e^t-1-t)}$ for all $t \ge 0$. We get relation (3.2) from this inequality with the choice $t = \log\left(1 + \frac{y}{V_n^2}\right)$.

(This is the place of minimum of the function $-ty+V_n^2(e^t-1-t)$ for fixed y in the parameter t.)

Relation (3.2) and the observation $\lim\limits_{u\to\infty}\frac{(u+1)\log(u+1)-u}{u\log u}=1$ with the choice $u=\frac{y}{V_n^2}$ imply formula (3.3). Because of relation (3.3) to prove formula (3.4) it is enough to check it for $2\le\frac{y}{V_n^2}\le B$ with some sufficiently large constant $B>0$. In this case relation (3.4) follows directly from formula (3.2). This can be seen for instance by observing that the expression $\frac{V_n^2\left[\left(1+\frac{y}{V_n^2}\right)\log\left(1+\frac{y}{V_n^2}\right)-\frac{y}{V_n^2}\right]}{y\log\frac{y}{V_n^2}}$ is a continuous and positive function of the variable $\frac{y}{V_n^2}$ in the interval $2\le\frac{y}{V_n^2}\le B$, hence its minimum in this interval is strictly positive.

Let us make a short comparison between Bernstein's and Bennett's inequality. Both results deal with the estimation of the probability $P(S_n>y)$, and their proofs are also very similar. In both cases first an estimate is given for the moment generating functions $R_j(t)=Ee^{tX_j}$ of the summands X_j. In Bennett's inequality a better estimate is given for them. (The worst case we have to handle is when $P(X_j=1)=\varepsilon_j$, $P\left(X_j=-\frac{e_j}{1-\varepsilon_j}\right)=1-\varepsilon_j$, and $\varepsilon_j+\frac{\varepsilon_j^2}{1-\varepsilon_j}=\sigma_j^2$. In this case the proof of Bennett's inequality contains an almost optimal estimate, while the estimate in Bernstein's inequality is weaker. In this estimate we are satisfied to give a good estimate for the first three coefficients in the Taylor expansion of the function $R_j(t)$.) With the help of this estimate a bound is given on the probability we are interested in which depends on the parameter t. In the proof of Bennett's inequality this parameter t is chosen optimally, while in Bernstein's inequality only an asymptotically optimal choice is taken. As a consequence, Bennett's inequality yields a sharper estimate. Actually Bernstein's inequality can be deduced from it. On the other hand, Bernstein's inequality gives a good, 'visible' bound for the probability $P(S_n>y)$ for not too large values of the number y which suffices for our purposes, while the magnitude of the estimate given by Bennett's inequality for small y cannot be directly seen. For large y Bennett's yields a better estimate, but this improvement seems to have a smaller importance.

I finish this section with another estimate due to Hoeffding which later will be useful for us when we want to carry out certain symmetrization arguments.

Theorem 3.4 (Hoeffding's inequality). *Let $\varepsilon_1,\dots,\varepsilon_n$ be independent random variables, $P(\varepsilon_j=1)=P(\varepsilon_j=-1)=\frac12$, $1\le j\le n$, and let $a_1,\dots,a_n$ be arbitrary real numbers. Put $V=\sum\limits_{j=1}^n a_j\varepsilon_j$. Then*

$$P(V>y)\le\exp\left\{-\frac{y^2}{2\sum_{j=1}^n a_j^2}\right\}\quad \textit{for all } y>0. \tag{3.5}$$

Remark: Clearly $EV=0$ and $\operatorname{Var}V=\sum\limits_{j=1}^n a_j^2$, hence Hoeffding's inequality yields such an estimate for $P(V > y)$ which the central limit theorem suggests. This estimate holds for all real numbers $a_1, \dots, a_n$.

Remark 2: If we consider the Rademacher functions $r_k(x)$, $r_k(x) = 1$ if $(2j-1)2^{-k} \le x < 2j2^{-k}$ and $r_k(x) = -1$ if $2(j-1)2^{-k} \le x < (2j-1)2^{-k}$, $1 \le j \le 2^k$, for all $k = 1, 2, \dots$, as random variables on the probability space $\Omega = [0,1]$ with the Borel σ-algebra and the Lebesgue measure as probability measure on the interval $[0,1]$, then they are independent random variables with the same distribution as the random variables $\varepsilon_1, \dots, \varepsilon_n$ considered in Theorem 3.4. Therefore such results which deal with random variables of this type are also called results about Rademacher functions in the literature. At some points we shall also use this terminology.

Proof of Theorem 3.4. Let us give a good bound on the exponential moment Ee^{tV} for all $t > 0$. We have $Ee^{tV} = \prod\limits_{j=1}^n Ee^{ta_j\varepsilon_j} = \prod\limits_{j=1}^n \frac{\left(e^{a_jt}+e^{-a_jt}\right)}{2}$, and $\frac{\left(e^{a_jt}+e^{-a_jt}\right)}{2} = \sum\limits_{k=0}^\infty \frac{a_j^{2k}}{(2k)!}t^{2k} \le \sum\limits_{k=0}^\infty \frac{(a_jt)^{2k}}{2^k k!} = e^{a_j^2t^2/2}$, since $(2k)! \ge 2^k k!$ for all $k \ge 0$. This implies that $Ee^{tV} \le \exp\left\{\frac{t^2}{2}\sum\limits_{j=1}^n a_j^2\right\}$. Hence $P(V > y) \le \exp\left\{-ty + \frac{t^2}{2}\sum\limits_{j=1}^n a_j^2\right\}$, and we get relation (3.5) with the choice $t = y\left(\sum\limits_{j=1}^n a_j^2\right)^{-1}$.

4. On the supremum of a nice class of partial sums

This section contains a result about the behaviour of the supremum of random integrals with respect to a normalized empirical measure in the special case when only one-fold integrals are considered. First we present an equivalent version of it about the supremum of a nice class of sums of independent, identically distributed random variables. We also discuss some natural problems related to them. In particular, we are interested in the question how restrictive the conditions of these results are. Also the natural Gaussian counterpart of these results will be given, but the proofs are postponed to a later section.

To formulate our results first we introduce the following notion.

Definition of L_p-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_p-dense class of functions, $1 \le p < \infty$, with parameter D and exponent L if for all numbers $1 \ge \varepsilon > 0$ and probability measures ν on the space $(Y, \mathcal{Y})$ there exists a finite ε-dense subset $\mathcal{G}_{\varepsilon,\nu} = \{g_1, \dots, g_m\} \subset \mathcal{G}$ in the space $L_p(Y, \mathcal{Y}, \nu)$ consisting of $m \le D\varepsilon^{-L}$ elements, i.e. there exists such a set $\mathcal{G}_{\varepsilon,\nu} \subset \mathcal{G}$ for which $\inf\limits_{g_j \in \mathcal{G}_{\varepsilon,\nu}} \int |g - g_j|^p \, d\nu < \varepsilon^p$ for all functions $g \in \mathcal{G}$. (Here the set $\mathcal{G}_{\varepsilon,\nu}$ may depend on the measure ν, but its cardinality is bounded by a number depending only on ε.)*

Now we formulate the following

Theorem 4.1. *Let us have a sequence of iid. random variables $\xi_1,\dots,\xi_n$, $n\ge 2$, taking values on a measurable space $(X,\mathcal{X})$ with some distribution μ together with an L_2-dense class $\mathcal{F}$ of functions of countable cardinality with some parameter D and exponent $L\ge 1$ on the space $(X,\mathcal{X})$ which satisfies the conditions*

$$\|f\|_\infty=\sup_{x\in X}|f(x)|\le 1, \qquad \text{for all } f\in\mathcal{F} \tag{4.1}$$

$$\|f\|_2^2=\int f^2(x)\mu(\,dx)\le\sigma^2 \qquad \text{for all } f\in\mathcal{F} \tag{4.2}$$

with some constant $\sigma>0$, and

$$\int f(x)\mu(\,dx)=0 \quad \text{for all } f\in\mathcal{F} \tag{4.3}$$

Define the normalized partial sums $S_n(f)=\frac{1}{\sqrt{n}}\sum\limits_{k=1}^{n} f(\xi_k)$ for all $f\in\mathcal{F}$ and introduce the number $\beta=\max\left(\frac{\log D}{\log n},0\right)$, where D is the parameter of the L_2-dense class $\mathcal{F}$.

There exist some 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

$$P\left(\sup_{f\in\mathcal{F}}|S_n(f)|\ge u\right)\le CD\exp\left\{-\alpha\left(\frac{u}{\sigma}\right)^2\right\} \qquad \text{if} \quad \sqrt{n}\sigma^2\ge u\ge\sqrt{M}(L+\beta)^{3/4}\sigma\log^{1/2}\frac{2}{\sigma}, \tag{4.4}$$

with the number β defined in this theorem, and the numbers D and L in formula (4.4) agree with the parameter and exponent of the L_2-dense class $\mathcal{F}$.

The condition about the countable cardinality of $\mathcal{F}$ can be weakened. For this goal we introduce the notion of countable approximability. For the sake of later applications it will be formulated more generally than needed in the present context.

Definition of countably approximable classes of random variables. *Let a class of random variables $U(f)$, $f\in\mathcal{F}$, indexed by a class of functions on a measure space $(Y,\mathcal{Y})$ be given. We say that this class of random variables $U(f)$, $f\in\mathcal{F}$, is countably approximable if there is a countable subset $\mathcal{F}'\subset\mathcal{F}$ such that for all numbers $u>0$ the sets $A(u)=\{\omega\colon \sup\limits_{f\in\mathcal{F}}|U(f)(\omega)|\ge u\}$ and $B(u)=\{\omega\colon \sup\limits_{f\in\mathcal{F}'}|U(f)(\omega)|\ge u\}$ satisfy the identity $P(A(u)\setminus B(u))=0$.*

Clearly, $B(u)\subset A(u)$. In the above definition we demanded that for all $u>0$ the set $B(u)$ should be almost as large as $A(u)$. The following corollary of Theorem 4.1 holds.

Corollary of Theorem 4.1. *Let a class of functions $\mathcal{F}$ satisfy the conditions of Theorem 4.1 with the only exception that instead of the condition about the countable* *cardinality of $\mathcal{F}$ it is assumed that the class of random variables $S_n(f)$, $f \in \mathcal{F}$, is countably approximable. Then the random variables $S_n(f)$, $f \in \mathcal{F}$, satisfy relation (4.4).*

This corollary can be simply proved, we only have to apply Theorem 4.1 for the class $\mathcal{F}'$. To do this we have to show that if $\mathcal{F}$ is an L_2-dense class with some parameter D and exponent L, and $\mathcal{F}' \subset \mathcal{F}$, then $\mathcal{F}'$ is also an L_2-dense class with the same exponent L, only with a possibly different parameter D'.

To prove this statement let us choose for all numbers $1 \ge \varepsilon > 0$ and probability measures ν on $(Y, \mathcal{Y})$ some functions $f_1, \dots, f_m \in \mathcal{F}$ with $m \le D\left(\frac{\varepsilon}{2}\right)^{-L}$ elements, such that the sets $\mathcal{D}_j = \left\{f\colon \int |f - f_j|^2\,d\nu \le \left(\frac{\varepsilon}{2}\right)^2\right\}$ satisfy the relation $\bigcup\limits_{j=1}^{m} \mathcal{D}_j = Y$. For all sets $\mathcal{D}_j$ for which $\mathcal{D}_j \cap \mathcal{F}'$ is non-empty choose a function $f_j' \in \mathcal{D}_j \cap \mathcal{F}'$. In such a way we get a collection of functions f_j' from the class $\mathcal{F}'$ containing at most $2^L D\varepsilon^{-L}$ elements which satisfies the condition imposed for L_2-dense classes with exponent L and parameter $2^L D$ for this number ε and measure ν.

Given a sequence of independent μ distributed random variables $\xi_1, \dots, \xi_n$ taking values on $(X, \mathcal{X})$ let us introduce their empirical distribution on $(X, \mathcal{X})$ as

$$\mu_n(A)(\omega) = \frac{1}{n}\#\left\{j\colon 1 \le j \le n,\ \xi_j(\omega) \in A\right\}, \quad A \in \mathcal{X}, \tag{4.5}$$

and define for all measurable (and integrable) functions f the (random) integral

$$J_n(f) = J_{n,1}(f) = \sqrt{n}\int f(x)(\mu_n(\,dx) - \mu(\,dx)). \tag{4.6}$$

Clearly $J_n(f) = \frac{1}{\sqrt{n}}\sum\limits_{j=1}^{n}(f(\xi_j) - Ef(\xi_j)) = S_n(\bar f)$ with $\bar f(x) = f(x) - \int f(x)\mu(\,dx)$. It is not difficult to see that $\sup\limits_{x\in X}|\bar f(x)| \le 2$ if $\sup\limits_{x\in X}|f(x)| \le 1$, $\int \bar f(x)\mu(\,dx) = 0$, $\int \bar f^2(x)\mu(\,dx) \le \int f^2(x)\mu(\,dx)$, if $\mathcal{F}$ is an L_2-dense class of functions with parameter D and exponent L, then the class of functions $\bar{\mathcal{F}}$ consisting of the functions $\bar f(x) = f(x) - \int f(x)\mu(\,dx)$, $f \in \mathcal{F}$, is an L_2-dense class of functions with parameter $2^L D$ and exponent L, since $\int(\bar f - \bar g)^2\,d\mu \le \varepsilon$ if $f, g \in \mathcal{F}$, and $\int (f-g)^2\,d\mu \le \left(\frac{\varepsilon}{2}\right)^2$. Hence Theorem 4.1 implies the following result which can be considered as its version reformulated for integrals with respect to normalized empirical measures.

Theorem 4.1′. *Let us have a sequence of iid. random variables $\xi_1, \dots, \xi_n$, $n \ge 2$, with distribution μ on a measurable space $(X, \mathcal{X})$ together with some class of functions $\mathcal{F}$ on this space which satisfy the conditions of Theorem 4.1 with the possible exception of condition (4.3). Then the estimate (4.4) remains valid if we replace the random sums $S_n(f)$ in it by the random integrals $J_n(f)$ defined in (4.6). Moreover, similarly to the corollary of Theorem 4.1, the countable cardinality of the set $\mathcal{F}$ can be replaced by the condition that the class of random variables $J_n(f)$, $f \in \mathcal{F}$, is countably approximable.*

All finite dimensional distributions of the set of random variables $S_n(f)$, $f \in \mathcal{F}$, converge to a Gaussian field $Z(f)$, $f \in \mathcal{F}$, as $n \to \infty$ with expectation $EZ(f) = 0$ and correlation $EZ(f)Z(g) = \int f(x)g(x)\mu(\,dx)$, $f, g \in \mathcal{F}$. (Here and in the subsequent part of the paper a collection of random variables indexed by some set of parameters will be called a Gaussian field if for all finite subsets of these parameters the random variables indexed by this finite set are jointly Gaussian.) Hence we can expect that the random variables of a Gaussian field with such properties satisfy a result similar to Proposition 4.1. The following result can be considered as the Gaussian counterpart of Theorem 4.1.

Theorem 4.2. *Let us fix some probability measure μ on a measurable space $(X, \mathcal{X})$ together with a countable set $\mathcal{F}$ of square integrable functions with respect to the measure μ such that there exists a parameter $D > 0$ and exponent $L \ge 1$ with the following property: For all $\varepsilon > 0$ there exist $m \le D\varepsilon^{-L}$ functions $f_j = f_j(\varepsilon) \in \mathcal{F}$, $1 \le j \le m$, such that for all $f \in \mathcal{F}$ $\inf\limits_{1 \le j \le m} \int (f_j(x) - f(x))^2 \mu(\,dx) < \varepsilon^2$. Let us also assume that the class of functions $\mathcal{F}$ satisfies condition (4.2) with some $1 \ge \sigma > 0$. Let us consider a Gaussian field $Z(f)$, $f \in \mathcal{F}$, such that $EZ(f) = 0$, $EZ(f)Z(g) = \int f(x)g(x)\mu(\,dx)$, $f, g \in \mathcal{F}$.*

Then there exist some constants $C > 0$ and $M > 0$ (for instance $C = 4$ and $M = 16$ can be chosen) such that the inequality

$$P\left(\sup_{f \in \mathcal{F}} |Z(f)| \ge u\right) \le C(D+1)\exp\left\{-\frac{1}{256}\left(\frac{u}{\sigma}\right)^2\right\} \quad \text{if } u \ge ML^{1/2}\sigma \log^{1/2}\frac{2}{\sigma} \tag{4.7}$$

holds with the parameter D and exponent L introduced in this theorem.

In the inequalities of the above results I did not try to find the best possible universal constants. One could choose for instance the coefficient $\frac{1-\varepsilon}{2}$ with arbitrary small $\varepsilon > 0$ instead of the coefficient $\frac{1}{256}$ in the exponent at the right-hand side of formula (4.7) if the other universal constants $C > 0$ and $M > 0$ are chosen sufficiently large in this inequality. This means that in the bound (4.7) we can get an estimate with an almost as good exponential term as in the estimate of the probability $P(Z(f) > u)$ for a single Gaussian random variable $Z(f)$ with $EZ(f) = 0$, $\operatorname{Var} Z(f) = \sigma^2$. Similarly, the constant $\alpha > 0$ can be chosen as $\alpha = \frac{1-\varepsilon}{2}$ with arbitrary small $\varepsilon > 0$ in formula (4.4).

The condition about the countable cardinality of the set $\mathcal{F}$ in Theorem 4.2 could be weakened similarly to Theorem 4.1. But here I omit the discussion of this question, since Theorem 4.2 was only introduced for the sake of a comparison between the Gaussian and non-Gaussian case. An essential difference between Theorems 4.1 and 4.2 is that in Theorem 4.1 the condition was imposed that the class of functions $\mathcal{F}$ has to be L_2-dense, while in Theorem 4.2 only a weaker version of this property was needed. In that result we only demanded that there exists a relatively small subset of $\mathcal{F}$ dense in the $L_2(\mu)$ norm. It may demand some explanation why the L_2-density property was imposed in Theorem 4.1, a property where also such probability measures ν are considered which seem to have no relation to the original problem. But as we shall see, the proof of Theorem 4.1 contains a conditioning argument where new conditional measures appear and the L_2-density property is needed to work with them. One would also like to know some results which enable us to check when this condition holds. In the next section we shall discuss a popular notion, the notion of Vapnik–Červonenkis classes and show that a Vapnik–Červonenkis class of functions bounded by 1 is L_2-dense.

Another difference between Theorems 4.1 and 4.2 is that the conditions of formula (4.4) contain the upper bound $n\sigma^2 > \sqrt{n}u$, and no such condition is imposed in formula (4.7). This difference can be simply explained, since as we have seen in Section 3 in the case $n\sigma^2 = \mathrm{Var}\,(\sqrt{n}S_n) \ll \sqrt{n}u$ we can guarantee only a weak non-Gaussian type estimate for the single probabilities $P(\sqrt{n}S_n(f) > \sqrt{n}u)$, $f \in \mathcal{F}$. It has a similar reason why condition (4.1) about the supremum of the functions $f \in \mathcal{F}$ appeared in Theorems 4.1 and 4.1′, and no such condition was needed in Theorem 4.2.

The lower bounds for the level u were imposed in formulas (4.4) and (4.7) because of a similar reason. To understand why such a condition is needed in formula (4.7) let us consider the following example. Take a Wiener process $W(t)$, $0 \le t \le 1$, define the functions $f_{s,t}(\cdot)$ on the interval $[0,1]$ by the formula $f_{s,t}(u) = 1$ if $s \le u \le t$, $f_{s,t}(u) = 0$ if $0 \le u < s$ or $t < u \le 1$, and put $Z(f_{s,t}) = \int f_{s,t}(u)W(\,du) = W(t) - W(s)$. Given some $\sigma > 0$ let us consider the class of functions $\mathcal{F}_\sigma = \{f_{s,t}\colon \int f^2_{s,t}(u)\,du = t - s \le \sigma^2, s \text{ and } t \text{ are rational numbers}\}$. It is not difficult to see that the above example satisfies the conditions of Theorem 4.2. It is natural to expect that $P\left(\sup\limits_{f\in\mathcal{F}_\sigma} Z(f) > u\right) \le e^{-\text{const.}\,(u/\sigma)^2}$. However, this relation does not hold if $u = u(\sigma) < (1-\varepsilon)\sqrt{2}\sigma\log^{1/2}\frac{1}{\sigma}$ with some $\varepsilon > 0$. In such cases $P\left(\sup\limits_{f\in\mathcal{F}_\sigma} Z(f) > u\right) \to 1$, as $\sigma \to 0$. This can be proved relatively simply with the help of the estimate $P(Z(f_{s,t}) > u(\sigma)) \ge \text{const.}\,\sigma^{1-\varepsilon}$ if $|t-s| = \sigma^2$ and the independence of the random integrals $Z(f_{s,t})$ if the functions $f_{s,t}$ are indexed by such pairs (s,t) for which the intervals (s,t) are disjoint. This means that in this example formula (4.7) holds only under the condition $u \ge M\sigma\log^{1/2}\frac{1}{\sigma}$ with $M = \sqrt{2}$.

Some additional work would show that a similar picture arises in the model where we consider the integrals $J_n(f_{s,t})$ of the functions from the same the class $\mathcal{F}_\sigma$ with respect to the normalized empirical measure of a sample of size n with uniform distribution on the interval $[0,1]$ instead of a Wiener process. In this example we have to impose the condition $\sqrt{n}u \ge M\sqrt{n}\sigma\log^{1/2}\frac{1}{\sigma}$ with $M = \sqrt{2}$ for the validity of relation (4.4). At a heuristic level it is clear that in the case of a class $\mathcal{F}$ with a large exponent L we have to put a larger coefficient of $\sqrt{n}\sigma\log^{1/2}\frac{2}{\sigma}$ in the condition of formula (4.4) for the validity of Theorem 4.1 or 4.1′, and a similar statement can be told about the condition (4.7) in Theorem 4.2. (I did not try to find the best possible coefficients in the conditions of relations (4.4) and (4.7), they could be improved considerably.)

In Theorem 4.1 (and in its version 4.1′) it was demanded that the class of functions $\mathcal{F}$ should be countable. Later this condition was replaced by a weaker condition about countable approximability. By restricting our attention to countable or countably approximable classes we could avoid some unpleasant measure theoretical problems which would have arisen if we had worked with the supremum of non-countable number of random variables which may be non-measurable. There are some papers where possibly non-measurable models are also considered with the help of some rather deep results of the analysis and measure theory. Actually, the problem we met here is the natural analog of an important problem in the theory of the stochastic processes about the smoothness property of the trajectories of an appropriate version of a stochastic process which we can get by exploiting our freedom to change all random variables on a set of probability zero.

The study of the problem in this work is simpler in one respect. Here the set of random variables $S_n(f)(\omega)$ or $J_n(f)(\omega)$, $f\in\mathcal{F}$, are constructed directly with the help of the underlying random variables $\xi_1(\omega),\dots,\xi_n(\omega)$ for all $\omega\in\Omega$ separately. We are interested in when the sets of random variables constructed in this way are countably approximable, i.e. we are not looking for a possibly different, better version of them with the same finite dimensional distributions. In the next simple Lemma 4.3 we give a sufficient condition for countable approximability. Its condition can be interpreted as a smoothness type condition for the trajectories of a stochastic process indexed by the functions $f\in\mathcal{F}$.

Lemma 4.3. *Let a class of random variables $U(f)$, $f\in\mathcal{F}$, indexed by some set $\mathcal{F}$ of functions on a space $(Y,\mathcal{Y})$ be given. If there exists a countable subset $\mathcal{F}'\subset\mathcal{F}$ of the set $\mathcal{F}$ such that the sets $A(u)=\{\omega\colon \sup\limits_{f\in\mathcal{F}}|U(f)(\omega)|\ge u\}$ and $B(u)=\{\omega\colon \sup\limits_{f\in\mathcal{F}'}|U(f)(\omega)|\ge u\}$ introduced for all $u>0$ in the definition of countable approximability satisfy the relation $A(u)\subset B(u-\varepsilon)$ for all $u>\varepsilon>0$, then the class of random variables $U(f)$, $f\in\mathcal{F}$, is countably approximable.*

The above property holds if for all $f\in\mathcal{F}$, $\varepsilon>0$ and $\omega\in\Omega$ there exists a function $\bar f=\bar f(f,\varepsilon,\omega)\in\mathcal{F}'$ such that $|U(\bar f)(\omega)|\ge|U(f)(\omega)|-\varepsilon$.

Proof of Lemma 4.3. If $A(u)\subset B(u-\varepsilon)$ for all $\varepsilon>0$, then $P^*(A(U)\setminus B(u))\le \lim\limits_{\varepsilon\to0}P(B(u-\varepsilon)\setminus B(u))=0$, where $P^*(X)$ denotes the outer measure of a not necessarily measurable set $X\subset\Omega$, since $\bigcap\limits_{\varepsilon\to0}B(u-\varepsilon)=B(u)$, and this is what we had to prove. If $\omega\in A(u)$, then for all $\varepsilon>0$ there exists some $f=f(\omega)\in\mathcal{F}$ such that $|U(f)(\omega)|>u-\frac{\varepsilon}{2}$. If there exists some $\bar f=\bar f(f,\frac{\varepsilon}{2},\omega)$, $f\in\mathcal{F}'$ such that $|U(\bar f)(\omega)|\ge|Uf(\omega)|-\frac{\varepsilon}{2}$, then $|U(\bar f)(\omega)|>u-\varepsilon$, and $\omega\in B(u-\varepsilon)$. This means that $A(u)\subset B(u-\varepsilon)$.

The question about countable approximability also appears in the case of multiple random integrals. To avoid some repetition we prove a result which also covers such cases. For this goal first we introduce the notion of multiple integrals with respect to a normalized empirical measure.

Given a measurable function $f(x_1,\dots,x_k)$ on the k-fold product space $(X^k,\mathcal{X}^k)$ and a sequence of independent random variables $\xi_1,\dots,\xi_n$ with some distribution μ on the space $(X,\mathcal{X})$ define the integral $J_{n,k}(f)$ of the function f with respect to the k-fold product of the normalized empirical measure μ_n introduced in (4.5) by the formula

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

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. (4.8)

Lemma 4.3 enables us to prove that certain classes of random variables $J_{n,k}(f)$, $f \in \mathcal{F}$, indexed by some set of functions $f \in \mathcal{F}$ of k variables are countably approximable. I present an example which is very important in certain applications.

Let us consider the case when $X = R^s$, the s-dimensional Euclidean space with some $s \ge 1$, and given some $u = (u^{(1)}, \dots, u^{(s)}) \in R^s$, $v = (v^{(1)}, \dots, v^{(s)}) \in R^s$ such that $u \le v$, i.e. $u^{(j)} \le v^{(j)}$ for all $1 \le j \le s$, let $B(u, v)$ denote the s-dimensional rectangle $B(u, v) = \{z \colon u \le z \le v\}$. Let us fix some function $f(x_1, \dots, x_k)$, $\sup |f(x_1, \dots, x_k)| \le 1$, on the space $(X^k, \mathcal{X}^k) = (R^{ks}, \mathcal{B}^{ks})$, where $\mathcal{B}^t$ denotes the Borel σ-algebra on the Euclidean space R^t together with some probability measure μ on $(R^s, \mathcal{B}^s)$. For all vectors $(u_1, \dots, u_k)$, $(v_1, \dots, v_k)$ such that $u_j, v_j \in R^s$ and $u_j \le v_j$, $1 \le j \le k$, let us define the function $f_{u_1,\dots,u_k,v_1,\dots,v_k}$ which equals the function f on the rectangle $[u_1, v_1] \times \cdots [u_k, v_k]$, and it is zero outside of this rectangle.

Let us consider a sequence of i.i.d. random variables $\xi_1, \dots, \xi_n$ taking value in the space $(R^s, \mathcal{B}^s)$ with distribution μ and define the empirical measure μ_n and random integrals $J_{n,k}(f_{u_1,\dots,u_k,v_1,\dots,v_k})$ by formulas (4.5) and (4.8), for all vectors $(u_1, \dots, u_k)$, $(v_1, \dots, v_k)$ such that $u_j, v_j \in R^s$ and $u_j \le v_j$, $1 \le j \le k$, with the above defined functions $f_{u_1,\dots,u_k,v_1,\dots,v_k}$. The following result will be proved.

Lemma 4.4. *Let us take n iid. random variables $\xi_1, \dots, \xi_n$ with values in the space $(R^s, \mathcal{B}^s)$. Let us define with the help of their distribution μ and the empirical distribution μ_n determined by them the class of random variables $J_{n,k}(f_{u_1,\dots,u_k,v_1,\dots,v_k})$ introduced in formula (4.8), where the class of kernel functions $\mathcal{F}$ in these integrals consists of all functions $f_{u_1,\dots,u_k,v_1,\dots,v_k} \in (R^{sk}, \mathcal{B}^{sk})$, $u_j, v_j \in R^s$, $u_j \le v_j$, $1 \le j \le k$, defined in the last but one paragraph. This class of random variables $J_{n,k}(f)$, $f \in \mathcal{F}$, is countably approximable.*

Proof of Lemma 4.4. We shall prove that the definition of countable approximability is satisfied in this model if the class of functions $\mathcal{F}'$ consists of those functions $f_{u_1,\dots,u_k,v_1,\dots,v_k}$, $u_j \le v_j$, $1 \le j \le k$, for which all coordinates of the vectors u_j and v_j are rational numbers.

Given some function $f_{u_1,\dots,u_k,v_1,\dots,v_k}$, a real number $1 > \varepsilon > 0$ and $\omega \in \Omega$ let us choose a function $f_{\bar u_1,\dots,\bar u_k,\bar v_1,\dots,\bar v_k} \in \mathcal{F}'$ determined with some vectors $\bar u_j = \bar u_j(\varepsilon, \omega)$, $\bar v_j = \bar v_j(\varepsilon, \omega)$ $1 \le j \le k$, with rational coordinates $\bar u_j \le u_j < v_j \le \bar v_j$ such that the sets $K_j = B(\bar u_j, \bar v_j) \setminus B(u_j, v_j)$ satisfy the relations $\mu(K_j) \le \varepsilon 2^{-2k+1} n^{-k/2}$, and $\xi_l(\omega) \notin K_j$ for all $j = 1, \dots, k$ and $l = 1, \dots, n$. Let us show that

$$|J_{n,k}(f_{\bar u_1,\dots,\bar u_k,\bar v_1,\dots,\bar v_k})(\omega) - J_{n,k}(f_{u_1,\dots,u_k,v_1,\dots,v_k})(\omega)| \le \varepsilon. \tag{4.9}$$

Lemma 4.3 (with the choice $U(f)=J_{n,k}(f)$) and relation (4.9) imply Lemma 4.4.

Relation (4.9) holds, since the expression in it can be written as the sum of the 2^k-1 integrals of the function f with respect to the k-fold product of the measure $\sqrt{n}(\mu_n-\mu)$ on the domains $D_1\times\cdots\times D_k$ with the omission of the diagonals $x_j=x_{\bar\jmath}$, $1\le j,\bar\jmath\le k$, $j\ne\bar\jmath$, where D_j is either the set K_j or $B(u_j,v_j)$ and $D_j=K_j$ for at least one index j. It is enough to show that the absolute value of all these integrals is less than $\varepsilon 2^{-k}$. This follows from the observations that $|f(x_1,\dots,x_k)|\le 1$, $\sqrt{n}(\mu_n-\mu)(K_j)=-\sqrt{n}\mu(K_j)$, $\mu(K_j)\le\varepsilon 2^{-2k+1}n^{-k/2}$, and the total variation of the signed measure $\sqrt{n}(\mu_n-\mu)$ (restricted to the set $B(u_j,v_j)$) is less than $2\sqrt{n}$.

Let us discuss the relation of the results in this section to an important result, the so-called fundamental theorem of the mathematical statistics. In that problem a sequence of independent random variables $\xi_1(\omega),\dots,\xi_n(\omega)$ is considered with distribution function $F(x)$, the empirical distribution function $F_n(x)=F_n(x,\omega)=\frac1n\#\{j\colon 1\le j\le n, \xi_j(\omega)<x\}$ is introduced, and the difference $F_n(x)-F(x)$ is considered. This result states that $\sup\limits_x|F_n(x)-F(x)|$ tends to zero with probability one.

Observe that $\sup\limits_x|F_n(x)-F(x)|=n^{-1/2}\sup\limits_{f\in\mathcal F}|J_n(f)|$, where $\mathcal F$ consists of the functions $f_x(\cdot)$, $x\in R^1$, defined by the relation $f_x(u)=1$ if $u<x$, and $f_x(u)=0$ if $u\ge x$. Theorem 4.1$'$ yields an estimate for the probabilities $P\left(\sup\limits_{f\in\mathcal F}|J_n(f)|>u\right)$. We have seen that the above class of functions $\mathcal F$ is countably approximable. The results of the next section imply that this class of functions is also L_2-dense. Otherwise it is not difficult to check this property directly. Hence we can apply Theorem 4.1 to the above defined class of functions with $\sigma=1$, and it yields that $P\left(n^{-1/2}\sup\limits_{f\in\mathcal F}|J_n(f)|>u\right)\le e^{-Cnu^2}$ if $1\ge u\ge\bar Cn^{-1/2}$ with some universal constants $C>0$ and $\bar C>0$. (The condition $1\ge u$ can actually be dropped.) The application of this estimate for the numbers $\varepsilon>0$ together with the Borel-Cantelli lemma imply the fundamental theorem of the mathematical statistics.

In short, the results of this section yield more information about the closeness the empirical distribution function F_n and distribution function F than the fundamental theorem of the mathematical statistics. Moreover, since these results can also be applied for other classes of functions they yield useful information about the closeness of the probability measure μ and empirical measure μ_n.

5. Vapnik–Červonenkis classes and L_2-dense classes of functions

In this section the most important notions and results will be presented about Vapnik–Červonenkis classes, and it will be explained how they help to show in some important cases that certain classes of functions are L_2-dense. Some proofs are put in the Appendix.

First I recall the following notions.

Definition of Vapnik-Červonenkis classes of sets and functions. *Let a set S be given, and let us select a class $\mathcal{D}$ consisting of certain subsets of this set S. We call $\mathcal{D}$ a Vapnik–Červonenkis class if there exist two real numbers B and K such that for all positive integers n and subsets $S_0(n) = \{x_1, \dots, x_n\} \subset S$ of cardinality n of the set S the collection of sets of the form $S_0(n) \cap D$, $D \in \mathcal{D}$, contains no more than Bn^K subsets of $S_0(n)$. We shall call B the parameter and K the exponent of this Vapnik–Červonenkis class.*

A class of real valued functions $\mathcal{F}$ on a space $(Y, \mathcal{Y})$ is called a Vapnik–Červonenkis class if the collection of graphs of these functions is a Vapnik–Červonenkis class, i.e. if the sets $A(f) = \{(y,t)\colon y \in Y,\ \min(0, f(y)) \le t \le \max(0, f(y))\}$, $f \in \mathcal{F}$, constitute a Vapnik–Červonenkis class of subsets of the product space $S = Y \times R^1$.

The following result which was first proved by Sauer is of fundamental importance in the theory of Vapnik–Červonenkis classes. Its proof is given in the Appendix.

Theorem 5.1 (Sauer's lemma). *Let a set S be given together with a class $\mathcal{D}$ of subsets of this set S. Fix some subset $S_0 = S_0(n)$ of the set S containing n point and consider the class of subsets $\mathcal{D}(S_0) = \{S_0 \cap D\colon D \in \mathcal{D}\}$ of S_0 consisting of the intersections of the set S_0 with the elements of the class $\mathcal{D}$. If there is some positive integer k such that all subsets $F \subset S_0$ of cardinality k have at least one "hidden" subset not contained in the collection of sets $\mathcal{D}(S_0, F) = \{F \cap B;\ B \in \mathcal{D}(S_0)\}$, then $\mathcal{D}(S_0)$ contains at most $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k-1}$ subsets of S.*

Theorem 5.1 has the remarkable consequence that if there exists some integer k such that for all subsets $S_0(k)$ of cardinality k of the set S the number of sets of the form $S_0(k) \cap D$, $D \in \mathcal{D}$, is less than 2^k, (i.e. not all subsets of $S_0(k)$ can be represented in this form,) then $S_0(n) \cap \mathcal{D}$ has at most $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k-1}$ elements for all subsets $S_0(n)$ of the set S with $n \ge k$ elements, since in this case the conditions of Theorem 5.1 hold for all $n \ge k$ and subset $S_0(n) \subset S$ of S of cardinality n and this number k. This means that in this case $\mathcal{D}$ is a Vapnik-Červonenkis class. It can be proved that $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k-1} \le 1.5 \frac{n^{k-1}}{(k-1)!}$ if $n \ge k+1$, and this relation enables us to give an explicit estimate on the exponent and parameter of this Vapnik–Červonenkis class. Hence we have to check a seemingly much weaker property to show that a class of subsets of a set S is a Vapnik–Červonenkis class. Moreover, Theorem 5.1 implies that there are two cases. Either there is some set $S_0(n)$ of cardinality n for all integers n such that $\mathcal{D}(S_0(n))$ contains all subsets of $S_0(n)$ or $\sup\limits_{S_0 \subset S, |S_0| = n} |\mathcal{D}(S_0)|$ tends to infinity in polynomial order as $n \to \infty$, where $|S_0|$ and $|\mathcal{D}(S_0)|$ denotes the cardinality of S_0 and $\mathcal{D}(S_0)$.

The upper bound given for $|\mathcal{D}(S_0)|$ in Theorem 5.1 appears in a natural way. If $\mathcal{D}(S_0)$ consists of the subsets of S_0 of cardinality less than or equal to $k-1$, then the above sum equals $|\mathcal{D}(S_0)|$. In such a case the conditions of Theorem 5.1 are satisfied, and the proof of Theorem 5.1 shows that this is the extreme case, this is the largest class of sets $S_0(n)\cap\mathcal{D}$ satisfying Theorem 5.1.

The following Theorem 5.2, an important result of Richard Dudley, states that a Vapnik–Červonenkis class of functions bounded by 1 is an L_1-dense class of functions.

Theorem 5.2. *Let $f(y)$, $f\in\mathcal{F}$, be a Vapnik–Červonenkis class of real valued functions on some measurable space $(Y,\mathcal{Y})$ such that $\sup\limits_{y\in Y}|f(y)|\le 1$ for all $f\in\mathcal{F}$. Then $\mathcal{F}$ is an L_1-dense class of functions on $(Y,\mathcal{Y})$. More explicitly, if $\mathcal{F}$ is a Vapnik–Červonenkis class with parameter $B\ge 1$ and exponent $K>0$, then it is an L_1-dense class with exponent $L=2K$ and parameter $D=CB^2(4K)^{2K}$ with some universal constant $C>0$.*

Proof of Theorem 5.2. Let us fix some probability measure ν on $(Y,\mathcal{Y})$ and a real number $1\ge\varepsilon>0$. We are going to show that the cardinality of any finite set $\mathcal{D}(\varepsilon,\nu)=\{f_1,\dots,f_M\}\subset\mathcal{F}$ such that $\int|f_j-f_k|\,d\nu\ge\varepsilon$ if $j\ne k$, $f_j,f_k\in\mathcal{D}(\varepsilon,\nu)$ has a cardinality $M\le D\varepsilon^{-L}$ with some $D>0$ and $L>0$. This implies that $\mathcal{F}$ is an L_1-dense class with parameter D and exponent L. Indeed, let us take a maximal subset $\bar{\mathcal{D}}(\varepsilon,\nu)=\{f_1,\dots,f_M\}$ such that the $L_1(\nu)$ distance of any two functions in this subset is at least ε. Maximality means in this context that no function f_{M+1} can be attached to $\bar{\mathcal{D}}(\varepsilon,\nu)$ without violating this condition. If we show that $M\le D\varepsilon^{-L}$, then this means that $\bar{\mathcal{D}}(\varepsilon,\nu)$ is an ε-dense subset of $\mathcal{F}$ in the space $L_p(Y,\mathcal{Y},\nu)$ with no more than $D\varepsilon^{-L}$ elements.

In the estimation of the cardinality M of $\mathcal{D}(\varepsilon,\nu)$ we exploit the Vapnik–Červonenkis class property of $\mathcal{F}$ in the following way. Let us choose relatively few p points (y_l,t_l), $y_l\in Y$, $-1\le t_l\le 1$, $1\le l\le p$, in the space $(Y\times[-1,1])$ in such a way that the set $S_0(p)=\{(y_l,t_l),\ 1\le l\le p\}$ and graphs $A(f_j)=\{(y,t)\colon y\in Y,\ \min(0,f_j(y))\le t\le\max(0,f_j(y))\}$, $f\in\mathcal{F}$, $f_j\in\mathcal{D}(\varepsilon,\nu)$ have the property that all sets $A(f_j)\cap S_0(p)$, $1\le j\le M$, are different. Then the Vapnik–Červonenkis class property of $\mathcal{F}$ implies that $M\le Bp^K$. Hence if we can construct a set $S_0(p)$ with the above property with a relatively small number p, then we get a useful estimate on M. Such a set $S_0(p)$ will be given by means of the following random construction.

Let us choose the p points (y_l,t_l), $1\le l\le p$, of the (random) set $S_0(p)$ independently of each other in such a way that the coordinate y_l is chosen with distribution ν on $(Y,\mathcal{Y})$ and the coordinate t_l with uniform distribution on the interval $[-1,1]$ independently of y_l. (The number p will be chosen later.) Let us fix some indices $1\le j,k\le M$ and estimate the probability that the sets $A(f_j)\cap S_0(p)$ and $A(f_k)\cap S_0(p)$ agree, where $A(f)$ denotes the graph of the function f. Consider the symmetric difference $A(f_j)\Delta A(f_k)$ of the sets $A(f_j)$ and $A(f_k)$. The sets $A(f_j)\cap S_0(p)$ and $A(f_k)\cap S_0(p)$ agree if and only if $(y_l,t_l)\notin A(f_j)\Delta A(f_k)$ for all $(y_l,t_l)\in S_0(p)$. Let us observe that for a fixed l $P((y_l,t_l)\in A(f_j)\Delta A(f_k))=\frac12(\nu\times\lambda)(A(f_j)\Delta A(f_k))=\frac12\int|f_j-f_k|\,d\nu\ge\frac{\varepsilon}{2}$, where λ denotes the Lebesgue measure. This implies that the probability that $A(f_j)\cap S_0(p)$ and $A(f_k)\cap S_0(p)$ agree can be bounded from above by $\left(1-\frac{\varepsilon}{2}\right)^p\le e^{-p\varepsilon/2}$. Hence the probability that all sets $A(f_j)\cap S_0(p)$ are different is greater than $1-\binom{M}{2}e^{-p\varepsilon/2}\ge 1-\frac{M^2}{2}e^{-p\varepsilon/2}$. Choose p such that $\frac{7}{4}e^{p\varepsilon/2}>e^{(p+1)\varepsilon/2}>M^2\ge e^{p\varepsilon/2}$. Then the above probability is greater than $\frac18$, and there exists some set $S_0(p)$ with the desired property.

The inequalities $M\le Bp^K$ and $M^2\ge e^{p\varepsilon/2}$ imply that $M\ge e^{\varepsilon M^{1/K}/4B^{1/K}}$, i.e. $\frac{\log M^{1/K}}{M^{1/K}}\ge\frac{\varepsilon}{4KB^{1/K}}$. As $\frac{\log M^{1/K}}{M^{1/K}}\le CM^{-1/2K}$ for $M\ge1$ with some universal constant $C>0$, this estimate implies that Theorem 5.2 holds with the exponent L and parameter D given in its formulation.

Let us observe that if $\mathcal{F}$ is an L_1-dense class of functions on a measure space $(Y,\mathcal{Y})$ with some exponent L and parameter D, and also the inequality $\sup\limits_{y\in Y}|f(y)|\le1$ holds for all $f\in\mathcal{F}$, then $\mathcal{F}$ is an L_2-dense class of functions with exponent $2L$ and parameter $D2^L$. Indeed, if we fix some measure ν on $(Y,\mathcal{Y})$ together with a number $1\ge\varepsilon>0$, and $\mathcal{D}(\varepsilon,\nu)=\{f_1,\dots,f_M\}$ is an $\frac{\varepsilon^2}{2}$-dense set of $\mathcal{F}$ in the space $L_1(Y,\mathcal{Y},\nu)$, $M\le 2^LD\varepsilon^{-2L}$ then for any $f\in\mathcal{F}$ we can choose some $f_j\in\mathcal{D}(\varepsilon,\nu)$ such that $\int(f-f_j)^2\,d\nu\le 2\int|f-f_j|\,d\nu\le\varepsilon^2$. This means that $\mathcal{F}$ is really an L_2-dense class with the given exponent and parameter.

It is not easy to check whether a collection of subsets $\mathcal{D}$ of a set S is a Vapnik–Červonenkis class even with the help of Theorem 5.1. Therefore the following Theorem 5.3 which enables us to construct many non-trivial Vapnik–Červonenkis classes is of special interest. Its proof is also put in the Appendix.

Theorem 5.3. *Let us consider a k-dimensional subspace $\mathcal{G}_k$ of the linear space of real valued functions defined on a set S, and define the level-set $A(g)=\{s\colon s\in S,\ g(s)\ge0\}$ for all functions $g\in\mathcal{G}_k$. Take the class of subsets $\mathcal{D}=\{A(g)\colon g\in\mathcal{G}_k\}$ of the set S consisting of the above introduced level sets. All subsets $S_0=S_0(k+1)\subset S$ of cardinality $k+1$ has a "hidden" subset which is not contained in the class of subsets $\mathcal{D}(S_0)=\{S_0\cap D\colon D\in\mathcal{D}\}$ of S_0 introduced in Theorem 5.1. By Theorem 5.1 this property implies that the class of sets $\mathcal{D}$ is a Vapnik–Červonenkis class.*

Theorem 5.3 enables us to construct many interesting Vapnik–Červonenkis classes. Thus for instance the class of all half-spaces in a Euclidean space, the class of all ellipses in the plane, or more generally the level sets of k-order algebraic functions with a fixed number k constitute a Vapnik–Červonenkis class. It can be proved that if $\mathcal{C}$ and $\mathcal{D}$ are Vapnik–Červonenkis classes of subsets of a set S, then also their intersection $\mathcal{C}\cap\mathcal{D}=\{C\cap D\colon C\in\mathcal{C},\ D\in\mathcal{D}\}$, their union $\mathcal{C}\cup\mathcal{D}=\{C\cup D\colon C\in\mathcal{C},\ D\in\mathcal{D}\}$ and complementers $\mathcal{C}^c=\{S\setminus C\colon C\in\mathcal{C}\}$ are Vapnik–Červonenkis classes. These results are less important for us and their proofs will be omitted. We are interested in Vapnik–Červonenkis classes not for their own sake. We are going to study L_2-dense classes of functions, and Vapnik–Červonenkis classes make possible to find some examples. Indeed, Theorem 5.2 implies that if $\mathcal{D}$ is a Vapnik–Červonenkis class of subsets of a set S, then their indicator functions constitute an L_1-dense, hence also an L_2-dense class of functions. Then the results of Lemma 5.4 formulated below enable us to construct new L_2-dense class of functions. The description of L_2-dense classes of functions are interesting for us, because they appear in the conditions of the results in Section 4.

Lemma 5.4. *Let $\mathcal{G}$ be an L_2-dense class of functions on some space $(Y,\mathcal{Y})$ whose absolute values are bounded by one, and let f be a function on $(Y,\mathcal{Y})$ also with absolute value bounded by one. Then $f\cdot\mathcal{G}=\{f\cdot g\colon g\in G\}$ is also an L_2-dense class of functions. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two L_2-dense classes of functions on some space $(Y,\mathcal{Y})$ whose absolute values are bounded by one. Then the classes of functions $\mathcal{G}_1+\mathcal{G}_2=\{g_1+g_2\colon g_1\in\mathcal{G}_1,\, g_2\in\mathcal{G}_2\}$, $\mathcal{G}_1\cdot\mathcal{G}_2=\{g_1g_2\colon g_1\in\mathcal{G}_1,\, g_2\in\mathcal{G}_2\}$, $\min(\mathcal{G}_1,\mathcal{G}_2)=\{\min(g_1,g_2)\colon g_1\in\mathcal{G}_1,\, g_2\in\mathcal{G}_2\}$, $\max(\mathcal{G}_1,\mathcal{G}_2)=\{\max(g_1,g_2)\colon g_1\in\mathcal{G}_1,\, g_2\in\mathcal{G}_2\}$ are also L_2-dense. If $\mathcal{G}$ is an L_2-dense class of functions, and $\mathcal{G}'\subset G$, then $\mathcal{G}'$ is also an L_2-dense class.*

The proof of Lemma 5.4 is rather straightforward. One has to observe for instance that if $g_1,\bar g_1\in\mathcal{G}_1$, $g_2,\bar g_2\in\mathcal{G}_2$ then $|\min(g_1,g_2)-\min(\bar g_1,\bar g_2)|\le|g_1-\bar g_1)|+|g_2-\bar g_2|$, hence if $g_{1,1},\dots,g_{1,M_1}$ is an $\frac{\varepsilon}{2}$-dense subset of $\mathcal{G}_1$ and $g_{2,1},\dots,g_{2,M_2}$ is an $\frac{\varepsilon}{2}$-dense subset of $\mathcal{G}_2$ in the space $L_2(Y,\mathcal{Y},\nu)$ with some probability measure ν, then the functions $\min(g_{1,j},g_{2,k})$, $1\le j\le M_1$, $1\le k\le M_2$ constitute an ε-dense subset of $\min(\mathcal{G}_1,\mathcal{G}_2)$ in $L_2(Y,\mathcal{Y},\nu)$. The last statement of Lemma 5.4 is proved after the Corollary of Theorem 4.1. The details are left to the reader.

The above results enable us to find some interesting classes of L_2-dense classes of functions. In particular, the indicator functions of Vapnik-Červonenkis class of sets is an L_2-dense class of functions, and then Lemma 5.4 enables us to construct new classes of L_2-dense classes of functions with their help. It is not difficult to see with the help of these results for instance that the random variables considered in Lemma 4.4 are not only countably approximable, but the class of functions $f_{u_1,\dots,u_k,v_1,\dots,v_k}$ taking part in their definition is L_2-dense.

6. The proof of Theorems 4.1 and 4.2 on the supremum of random sums

This section contains the proof of some results which can be proved by means of a simple but useful method, the so-called chaining argument. This method enables us to prove Theorem 4.2 completely, but it only helps to reduce Theorem 4.1 to a slightly simpler statement presented in Proposition 6.1. We also formulate another result in Proposition 6.2 and show that these two propositions together imply Theorem 4.1. The proof of Proposition 6.2 which is based on a symmetrization argument is left to the next section. The method of proof of Theorem 4.2 does not suffice in itself to prove Theorem 4.1, because we have relatively weak estimates about the distribution of sums of independent random variables with small variances. This does not allow to follow the chaining argument in the proof of Theorem 4.1 up to the end, we have to stop at a point. In such a way we only get a seemingly weak result, but as it turns out this is the result we need to cover that part of Theorem 4.1 which cannot be handled by means of the symmetrization method applied in the proof of Proposition 6.2. First we prove Theorem 4.2.

Proof of Theorem 4.2. Let us list the elements of $\mathcal{F}$ as $\{f_0,f_1,\dots\}=\mathcal{F}$, and choose for all $p=0,1,2,\dots$ a set of functions $\mathcal{F}_p=\{f_{a(p,1)},\dots,f_{a(p,m_p)}\}\subset\mathcal{F}$ with $m_p\le (D+1)\,2^{2pL}\sigma^{-L}$ elements in such a way that $\inf\limits_{1\le j\le m_p}\int(f-f_{a(p,j)})^2\,d\mu\le 2^{-4p}\sigma^2$ for all $f\in\mathcal{F}$, and $f_p\in\mathcal{F}_p$. For all indices $a(p,j)$ of the functions in $\mathcal{F}_p$, $p=1,2,\dots$, define a predecessor $a(p-1,j')$ from the indices of the set of functions $\mathcal{F}_{p-1}$ in such a way that the functions $f_{a(p,j)}$ and $f_{a(p-1),j')}$ satisfy the relation $\int(f_{(p,j)}-f_{(p-1,j')})^2\,d\mu\le 2^{-4(p-1)}\sigma^2$. With the help of the behaviour of the standard normal distribution function we can write the estimates

$$
\begin{aligned}
P(A(p,j)) &= P\left(|Z(f_{a(p,j)})-Z(f_{a(p-1,j')})|\ge 2^{-(1+p)}u\right)\le 2\exp\left\{-\frac{2^{-2(p+1)}u^2}{2\cdot 2^{-4(p-1)}\sigma^2}\right\}\\
&= 2\exp\left\{-\frac{2^{2p}u^2}{128\sigma^2}\right\}\qquad 1\le j\le m_p,\ p=1,2,\dots,
\end{aligned}
$$

and

$$
P(B(j))=P\left(|Z(f_{a(0,j)})|\ge\frac{u}{2}\right)\le\exp\left\{-\frac{u^2}{8\sigma^2}\right\},\qquad 1\le j\le m_0.
$$

The above estimates together with the relation $\bigcup\limits_{p=0}^{\infty}\mathcal{F}_p=\mathcal{F}$ which implies that $\{|Z(f)|\ge u\}\subset\bigcup\limits_{p=1}^{\infty}\bigcup\limits_{j=1}^{m_p}A(p,j)\cup\bigcup\limits_{s=1}^{m_0}B(s)$ for all $f\in\mathcal{F}$ yield that

$$
\begin{aligned}
P\left(\sup_{f\in\mathcal{F}}|Z(f)|\ge u\right)&\le P\left(\bigcup_{p=1}^{\infty}\bigcup_{j=1}^{m_p}A(p,j)\cup\bigcup_{s=1}^{m_0}B(s)\right)\\
&\le\sum_{p=1}^{\infty}\sum_{j=1}^{m_p}P(A(p,j))+\sum_{s=1}^{m_0}P(B(s))\\
&\le\sum_{p=1}^{\infty}2(D+1)2^{2pL}\sigma^{-L}\exp\left\{-\frac{2^{2p}u^2}{128\sigma^2}\right\}+2(D+1)\sigma^{-L}\exp\left\{-\frac{u^2}{8\sigma^2}\right\}.
\end{aligned}
$$

If $u\ge ML^{1/2}\sigma\log\frac{2}{\sigma}$ with $M\ge16$ (and $L\ge1$), then

$$
2^{2pL}\sigma^{-L}\exp\left\{-\frac{2^{2p}u^2}{256\sigma^2}\right\}\le\left(\frac12\right)^{-2pL}\sigma^{-L}\left(\frac{\sigma}{2}\right)^{2^{2p}M^2L/256}\le 2^{-pL}\le 2^{-p}
$$

for all $p=0,1\dots$, hence the previous inequality implies that

$$
P\left(\sup_{f\in\mathcal{F}}|Z(f)|\ge u\right)\le 2(D+1)\sum_{p=0}^{\infty}2^{-p}\exp\left\{-\frac{2^{2p}u^2}{256\sigma^2}\right\}=4(D+1)\exp\left\{-\frac{u^2}{256\sigma^2}\right\}.
$$

Theorem 4.2 is proved.

With the appropriate choice of the bound of the integrals in the definition of the sets $\mathcal{F}_p$ in the proof of Theorem 4.2 and some more calculation it can be proved that the coefficient $\frac{1}{256}$ in the exponent of the right-hand side (4.7) can be replaced by $\frac{1-\varepsilon}{2}$ with arbitrary small $\varepsilon > 0$ if the remaining constants in this estimate are chosen sufficiently large.

The proof of Theorem 4.2 was based on the fact that sufficiently good estimates can be given on the probabilities $P(|Z(f)-Z(g)|>u)$ for all $f,g\in\mathcal{F}$ and $u>0$. In the case of Theorem 4.1 we only have a weaker estimate for the corresponding probabilities, we cannot give a good estimate on the distribution of the difference $S_n(f)-S_n(g)$ if its variance is small. As a consequence the chaining argument supplies only a weaker result in this case. This result will be given in Proposition 6.1, where the supremum of the normalized random sums $S_n(f)$ is estimated on a relatively dense subset of the class of functions $f\in\mathcal{F}$ in the $L_2(\mu)$ norm. We present another result in Proposition 6.2 which will be proved in the next section and show that Theorem 4.1 follows from these two results.

Before the formulation of Proposition 6.1 I recall an estimate which is a simple consequence of Bernstein's inequality: If $S_n(f)=\frac{1}{\sqrt{n}}\sum\limits_{j=1}^n f(\xi_j)$ is the normalized sum of independent, identically random variables, $P(|f(\xi_1)|\le 1)=1$, $Ef(\xi_1)=0$, $Ef(\xi_1)^2\le\sigma^2$, then there exists some constant $\alpha>0$ such that

$$P(|S_n(f)|>u)\le 2e^{-\alpha u^2/\sigma^2} \quad \text{if} \quad 0<u<\sqrt{n}\sigma^2. \tag{6.1}$$

We can choose $\alpha=\frac{3}{8}$ in this estimate, and also could present a slightly more general version of it, but such additional information would not give a real help.

Proposition 6.1. *Let us have a countable L_2-dense class of functions $\mathcal{F}$ with parameter D and exponent L, $L\ge 1$, on a measurable space $(X,\mathcal{X})$ whose elements satisfy the conditions (4.1), (4.2) and (4.3) with some probability measure μ on $(X,\mathcal{X})$ and real number $0<\sigma\le 1$. Take a sequence of independent μ-distributed random variables $\xi_1,\dots,\xi_n$, $n\ge 2$, define the random sums $S_n(f)=\frac{1}{\sqrt{n}}\sum\limits_{l=1}^n f(\xi_l)$, for all $f\in\mathcal{F}$. Let us fix some number $\bar{A}\ge 2$. For all sufficiently large numbers $M\ge M_0=M_0(\bar{A})$ the following relation holds: For all numbers $u>0$ for which $n\sigma^2\ge\left(\frac{u}{\sigma}\right)^2\ge ML\log\frac{2}{\sigma}$ a number $\bar{\sigma}=\bar{\sigma}(u)$, $0\le\bar{\sigma}\le\sigma\le 1$, and a collection of functions $\mathcal{F}_{\bar{\sigma}}=\{f_1,\dots,f_m\}\subset\mathcal{F}$ with $m\le D\bar{\sigma}^{-L}$ elements can be chosen in such a way that the sets $\mathcal{D}_j=\{f\colon f\in\mathcal{F},\int|f-f_j|^2\,d\mu\le\bar{\sigma}^2\}$, $1\le j\le m$, satisfy the relation $\bigcup\limits_{j=1}^m\mathcal{D}_j=\mathcal{F}$, and the normalized partial sums $S_n(f)$, $f\in\mathcal{F}_{\bar{\sigma}}$, $n\ge 2$, satisfy the inequality*

$$P\left(\sup_{f\in\mathcal{F}_{\bar{\sigma}}}|S_n(f)|\ge\frac{u}{\bar{A}}\right)\le 4D\exp\left\{-\alpha\left(\frac{u}{10\bar{A}\sigma}\right)^2\right\} \quad \text{if} \quad n\sigma^2\ge\left(\frac{u}{\sigma}\right)^2\ge ML\log\frac{2}{\sigma} \tag{6.2}$$

with the constants α in formula (6.1) and the exponent L and parameter D of the L_2-dense class $\mathcal{F}$. Besides, also the inequalities $\frac{1}{4}\left(\frac{u}{\bar{A}\bar{\sigma}}\right)^2\ge n\bar{\sigma}^2\ge\frac{1}{64}\left(\frac{u}{\bar{A}\sigma}\right)^2$ and $n\sigma^2\ge\frac{M^{2/3}(L+\beta)\log n}{1000\bar{A}^{4/3}}$ hold with $\beta=\max\left(\frac{\log D}{\log n},0\right)$, provided that also the inequality $n\sigma^2\ge$

$\left(\frac{u}{\sigma}\right)^2 \ge M(L+\beta)^{3/2}\log\frac{2}{\sigma}$ holds. (We may assume that the sample size n is sufficiently large, so the set of numbers u for which $n\sigma^2 \ge \left(\frac{u}{\sigma}\right)^2 \ge M(L+\beta)^{3/2}\log\frac{2}{\sigma}$ is non-empty.)

Proposition 6.1 helps to reduce the proof of Theorem 4.1 to the case when the L_2 norm of the functions in the class $\mathcal{F}$ is bounded by a relatively small number $\bar\sigma$. In more detail, the proof of Theorem 4.1 can be reduced to a good estimate on the distribution of the supremum of random variables $\sup\limits_{f\in D_j}|S_n(f-f_j)|$ for all classes $\mathcal{D}_j$, $1\le j\le m$, by means of Proposition 6.1. We also have to know that the number m of the classes $\mathcal{D}_j$ is not too large, otherwise our estimates cannot be useful.

A result formulated in Proposition 6.2 helps us to complete the proof of Theorem 4.1. It contains some parameters, and we have to fit the constants in the estimates of Propositions 6.1 and 6.2. This was the reason to introduce the rather artificial parameter $\bar A\ge 2$ in Proposition 6.1 and to formulate the conditions of inequality (6.2) with a number $M\ge M_0(\bar A)$ instead of a number M_0. We want such a formulation of Proposition 6.1 in which it can achieved for any fixed number $A>0$ that the relation $n\bar\sigma^2\ge A\log n$ holds, where the number $\bar\sigma$ was defined in the proof of Proposition 6.1. The last two relations in Proposition 6.1 shows that this is possible if first the number $\bar A$ and then the number $M\ge M_0(\bar A)$ is chosen sufficiently large. Now we formulate Proposition 6.2 and prove Theorem 4.1 with its help.

Proposition 6.2. *Let us have a probability measure μ on a measurable space $(X,\mathcal{X})$ together with a sequence of independent and μ distributed random variables $\xi_1,\dots,\xi_n$, $n\ge 2$, and a countable, L_2-dense class of functions $f=f(x)$ on $(X,\mathcal{X})$ with some parameter D and exponent $L\ge 1$ which satisfies conditions (4.1), (4.2) and (4.3) with some $\sigma>0$ such that the inequality $n\sigma^2>K(L+\beta)\log n$ holds with an appropriate, sufficiently large universal number $K\ge 3$ and $\beta=\max\left(0,\frac{\log D}{\log n}\right)$. Then there exists some universal constant $\gamma>0$ and threshold index $A_0>0$ such that the random sums $S_n(f)$, $f\in\mathcal{F}$, introduced in Theorem 4.1 satisfy the inequality*

$$P\left(\sup_{f\in\mathcal{F}}|S_n(f)|\ge An^{1/2}\sigma^2\right)\le e^{-\gamma A^{1/2}n\sigma^2}\quad \text{if } A\ge A_0. \tag{6.3}$$

(A possible choice of the parameters is: $K=4$, $A_0=2^{10}\cdot 10^{16}$ and $\gamma=\frac12$.)

I did not try to find optimal parameters in formula (6.3). Even the exponent $\frac12$ of A in the exponent at its right-hand side could be improved. The result of Proposition 6.2 is similar to that of Theorem 4.1. Both of them give an estimate on a probability of the type $P\left(\sup\limits_{f\in\mathcal{F}}|S_n(f)|\ge u\right)$. The essential difference between them is that in Theorem 4.1 this probability is considered for $u\le \text{const.}\, n^{1/2}\sigma^2$ while in Proposition 6.2 the case $u>\text{const.}\, n^{1/2}\sigma^2$ is looked at. Let us observe that this is the case when no good Gaussian type estimate can be given for the probabilities $P(S_n(f)\ge u)$, $f\in\mathcal{F}$. In this case Bernstein's inequality yields the bound

$P(S_n(f) > An^{1/2}\sigma^2) = P\left(\sum\limits_{l=1}^n f(\xi_l) > xV_n\right) < e^{-\text{const.}\,An\sigma^2}$ with $x = A\sqrt{n}\sigma$ and $V_n = \sqrt{n}\sigma$ for each single function $f \in \mathcal{F}$ which takes part in the supremum of formula (6.3). The estimate (6.3) yields a slightly weaker estimate for the supremum of such random variables as it contains the coefficient $A^{1/2}$ instead of A in the exponent of the estimate at the right-hand side. But also such a bound will be sufficient for us.

In Proposition 6.2 that situation is considered when the irregularities of the summands provide a non-negligible contribution to the probabilities $P(|S_n(f)| \ge u)$, and the method of proof supplies a good estimate only in this case. This makes natural to separate the proof Theorem 4.1 to the proof of two different statements given in Proposition 6.1 and 6.2.

In the proof of Theorem 4.1 Propositions 6.1 will be applied with a sufficiently large number $\bar{A} \ge 2$ and Proposition 6.2 with $\sigma = \bar{\sigma}$ with the number $\bar{\sigma}$ defined in Proposition 6.1 and the classes $\mathcal{F} = \mathcal{D}_j$, more precisely the classes of functions $\mathcal{F} = \left\{\frac{g-f_j}{2} : g \in \mathcal{D}_j\right\}$ introduced in Proposition 6.1, where f_j is the function appearing in the definition of the class of functions $\mathcal{D}_j$. Clearly,

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}} |S_n(f)| \ge u\right) &\le P\left(\sup_{f\in\mathcal{F}_{\bar{\sigma}}} |S_n(f)| \ge \frac{u}{\bar{A}}\right) \\ &\quad + \sum_{j=1}^m P\left(\sup_{g\in\mathcal{D}_j} \left|S_n\left(\frac{f_j-g}{2}\right)\right| \ge \left(\frac{1}{2}-\frac{1}{2\bar{A}}\right)u\right), \end{aligned} \tag{6.4}$$

where m is the cardinality of the set of functions $\mathcal{F}_{\bar{\sigma}}$ appearing in Proposition 6.1. We want to show that if $\bar{A}$ and then $M \ge M_0(\bar{A})$ are chosen sufficiently large, then the second term at the right-hand side can be well bounded by means of Proposition 6.2, and Theorem 4.1 can be proved by means of this estimate.

Let us choose a number $\bar{A}_0$ in such a way that $\bar{A}_0 \ge A_0$ and $\gamma\bar{A}_0^{1/2} \ge \frac{1}{K}$ with the numbers A_0, K and γ in Proposition 6.2, put $\bar{A} = \max(2\bar{A}_0, 2)$, and apply Proposition 6.1 with this number $\bar{A}$. Then also the inequality $\left(\frac{u}{\bar{\sigma}}\right)^2 \ge 4\bar{A}^2 n\bar{\sigma}^2 \ge (4\bar{A}_0)^2 n\bar{\sigma}^2$, hence $u \ge 4\bar{A}_0\sqrt{n}\bar{\sigma}^2$ holds with the number $\bar{\sigma}$ in Proposition 6.1. (We assume that such numbers u are considered which satisfy the condition $n\sigma^2 \ge \left(\frac{u}{\sigma}\right)^2 \ge M(L+\beta)^{3/2}\log\frac{2}{\sigma}$ imposed in Proposition 6.1.) Choose such a number $M \ge M_0(\bar{A})$ in Proposition 6.1 (which also can be chosen as the number M in formula (4.4) of Theorem 4.1) which also satisfies the inequality $\frac{M^{2/3}(L+\beta)\log n}{1000\bar{A}^{4/3}} \ge K(L+\beta)\log n$ with the number K appearing in the conditions of Proposition 6.2. With such a choice we also have $n\bar{\sigma}^2 \ge K(L+\beta)\log n$.

Since $\left(\frac{1}{2}-\frac{1}{2\bar{A}}\right)u \ge \frac{u}{4} \ge \bar{A}_0\sqrt{n}\bar{\sigma}^2$ and $\bar{A}_0 \ge A_0$ Propositions 6.2 yields the estimation

$$\begin{aligned} P\left(\sup_{g\in\mathcal{D}_j}\left|S_n\left(\frac{f_j-g}{2}\right)\right| \ge \left(\frac{1}{2}-\frac{1}{2\bar{A}}\right)u\right) &\le P\left(\sup_{g\in\mathcal{D}_j}\left|S_n\left(\frac{f_j-g}{2}\right)\right| \ge \bar{A}_0\sqrt{n}\bar{\sigma}^2\right) \\ &\le e^{-\gamma\bar{A}_0^{1/2}n\bar{\sigma}^2} \quad \text{for all } 1 \le j \le m, \end{aligned}$$

(observe that the set of functions $\frac{f_j-g}{2}$, $g\in\mathcal{D}_j$ is an L_2-dense class with parameter D and exponent L), hence Proposition 6.1 and formula 6.4 imply that

$$P\left(\sup_{f\in\mathcal{F}}|S_n(f)|\ge u\right)\le 4D\exp\left\{-\alpha\left(\frac{u}{10\bar A\sigma}\right)^2\right\}+D\bar\sigma^{-L}e^{-\gamma\bar A_0^{1/2}n\bar\sigma^2}. \tag{6.5}$$

To get the result of Theorem 4.1 from inequality (6.5) we have to replace its second term at the right-hand side with a more appropriate expression where, in particular, we get rid of the coefficient $\bar\sigma^{-L}$. The condition $n\bar\sigma^2\ge K(L+\beta)\log n$ implies that $\bar\sigma\ge n^{-1/2}$, and by our choice of $\bar A_0$ we have $\gamma\bar A_0^{1/2}n\bar\sigma^2\ge\frac1K n\bar\sigma^2\ge L\log n\ge 2L\log\frac1{\bar\sigma}$, i.e. $\bar\sigma^{-L}\le e^{\gamma\bar A_0^{1/2}n\bar\sigma^2/2}$. By the estimates of Proposition 6.1 $n\bar\sigma^2\ge\frac1{64}\left(\frac{u}{\bar A\sigma}\right)^2$. The above relations imply that $\bar\sigma^{-L}e^{-\gamma\bar A_0^{1/2}n\bar\sigma^2}\le e^{-\gamma\bar A_0^{1/2}n\bar\sigma^2/2}\le\exp\left\{-\frac{\gamma}{128}\bar A_0^{1/2}\bar A^{-2}\left(\frac u\sigma\right)^2\right\}$. Then relation (6.5) gives that

$$P\left(\sup_{f\in\mathcal{F}}|S_n(f)|\ge u\right)\le 4D\exp\left\{-\frac{\alpha}{(10\bar A)^2}\left(\frac u\sigma\right)^2\right\}+D\exp\left\{-\frac{\gamma}{128}\bar A_0^{1/2}\bar A^{-2}\left(\frac u\sigma\right)^2\right\},$$

and this estimate implies Theorem 4.1.

Proof of Proposition 6.1. Let us list the members of $\mathcal{F}$, as $f_1,f_2,\dots$, and choose for all $p=0,1,2,\dots$ a set $\mathcal{F}_p=\{f_{a(p,1)},\dots,f_{a(p,m_p)}\}\subset\mathcal{F}$ with $m_p\le D\,2^{2pL}\sigma^{-L}$ elements in such a way that $\inf\limits_{1\le j\le m_p}\int(f-f_{a(p,j)})^2\,d\mu\le 2^{-4p}\sigma^2$ for all $f\in\mathcal{F}$. For all indices $a(p,j)$, $p=1,2,\dots$, $1\le j\le m_p$, choose a predecessor $a(p-1,j')$, $j'=j'(p,j)$, $1\le j'\le m_{p-1}$, in such a way that the functions $f_{a(p,j)}$ and $f_{a(p-1,j')}$ satisfy the relation $\int|f_{a(p,j)}-f_{a(p-1,j')}|^2\,d\mu\le\sigma^2 2^{-4(p-1)}$. Then we have $\int\left(\frac{f_{a(p,j)}-f_{a(p-1,j')}}{2}\right)^2d\mu\le 4\sigma^2 2^{-4p}$ and $\sup\limits_{x_j\in X,\,1\le j\le k}\left|\frac{f_{a(p,j)}(x_1,\dots,x_k)-f_{a(p-1,j')}(x_1,\dots,x_k)}{2}\right|\le 1$. Relation (6.1) yields that

$$\begin{aligned}P(A(p,j))=P\left(\frac12|S_n(f_{a(p,j)}-f_{a(p-1,j')})|\ge\frac{2^{-(1+p)}u}{2\bar A}\right)&\le 2\exp\left\{-\alpha\left(\frac{2^pu}{8\bar A\sigma}\right)^2\right\}\\ \text{if}\quad 4n\sigma^2 2^{-4p}\ge\left(\frac{2^pu}{8\bar A\sigma}\right)^2,\quad 1\le j\le m_p,\quad p=1,2,\dots,&\end{aligned}\tag{6.6}$$

and

$$\begin{aligned}P(B(s))=P\left(|S_n(f_{0,s})|\ge\frac{u}{2\bar A}\right)\le 2\exp\left\{-\alpha\left(\frac{u}{2\bar A\sigma}\right)^2\right\},\quad 1\le s\le m_0,\\ \text{if}\quad n\sigma^2\ge\left(\frac{u}{2\bar A\sigma}\right)^2.\end{aligned}\tag{6.7}$$

Choose the integer number R, $R\ge0$, in such a way that $\frac{2^{6(R+1)}}{256}\left(\frac{u}{\bar A\sigma}\right)^2\ge n\sigma^2\ge\frac{2^{6R}}{256}\left(\frac{u}{\bar A\sigma}\right)^2$, define $\bar\sigma^2=2^{-4R}\sigma^2$ and $\mathcal{F}_{\bar\sigma}=\mathcal{F}_R$. (As $n\sigma^2\ge\left(\frac u\sigma\right)^2$ and $\bar A\ge2$ by our conditions, there exists such a positive number R. The number R was chosen as the largest number p for which relation (6.6) holds.) Then the cardinality m of the set $\mathcal{F}_{\bar\sigma}$ equals $m_R \le D2^{2R}\sigma^{-L} = D\bar\sigma^{-L}$, and the sets $\mathcal{D}_j$ are $\mathcal{D}_j = \{f\colon f \in \mathcal{F}, \int (f_{a(R,j)} - f)^2\,d\mu \le 2^{-4R}\sigma^2\}$, $1 \le j \le m_R$, hence $\bigcup\limits_{j=1}^{m} \mathcal{D}_j = \mathcal{F}$. Besides, the number R was chosen in such a way that the inequalities (6.6) and (6.7) can be applied for $1 \le p \le R$. Hence the definition of the predecessor of an index (p,j) implies that

$$
\begin{aligned}
P\left(\sup_{f\in\mathcal{F}_{\bar\sigma}} |S_n(f)| \ge \frac{u}{\bar A}\right) &\le P\left(\bigcup_{p=1}^{R}\bigcup_{j=1}^{m_p} A(p,j) \cup \bigcup_{s=1}^{m_0} B(s)\right) \\
&\le \sum_{p=1}^{R}\sum_{j=1}^{m_p} P(A(p,j)) + \sum_{s=1}^{m_0} P(B(s)) \le \sum_{p=1}^{\infty} 2D\,2^{2pL}\sigma^{-L}\exp\left\{-\alpha\left(\frac{2^p u}{8\bar A\sigma}\right)^2\right\} \\
&\qquad + 2D\sigma^{-L}\exp\left\{-\alpha\left(\frac{u}{2\bar A\sigma}\right)^2\right\}.
\end{aligned}
$$

If the relation $\left(\frac{u}{\sigma}\right)^2 \ge ML\log\frac{2}{\sigma}$ holds with a sufficiently large constant M (depending on $\bar A$), then the inequalities

$$
2^{2pL}\sigma^{-L}\exp\left\{-\alpha\left(\frac{2^p u}{8\bar A\sigma}\right)^2\right\} \le 2^{-p}\exp\left\{-\alpha\left(\frac{2^p u}{10\bar A\sigma}\right)^2\right\}
$$

hold for all $p = 1, 2, \dots$, and

$$
\sigma^{-L}\exp\left\{-\alpha\left(\frac{u}{2\bar A\sigma}\right)^2\right\} \le \exp\left\{-\alpha\left(\frac{u}{10\bar A\sigma}\right)^2\right\}.
$$

Hence the previous estimate implies that

$$
\begin{aligned}
P\left(\sup_{f\in\mathcal{F}_{\bar\sigma}} |S_n(f)| \ge \frac{u}{\bar A}\right) &\le \sum_{p=1}^{\infty} 2D2^{-p}\exp\left\{-\alpha\left(\frac{2^p u}{10\bar A\sigma}\right)^2\right\} \\
&\quad + 2D\exp\left\{-\alpha\left(\frac{u}{10\bar A\sigma}\right)^2\right\} \le 4D\exp\left\{-\alpha\left(\frac{u}{10\bar A\sigma}\right)^2\right\},
\end{aligned}
$$

and relation (6.2) holds. We have

$$
2^{-4R}\cdot\frac{2^{6R}}{256}\left(\frac{u}{\bar A\sigma}\right)^2 \le n\bar\sigma^2 = 2^{-4R}n\sigma^2 \le 2^{-4R}\cdot\frac{2^{6(R+1)}}{256}\left(\frac{u}{\bar A\sigma}\right)^2 = \frac{1}{4}\cdot 2^{2R}\left(\frac{u}{\bar A\sigma}\right)^2,
$$

hence

$$
\frac{1}{64}\left(\frac{u}{\bar A\sigma}\right)^2 \le n\bar\sigma^2 \le \frac{1}{4}\cdot\left(\frac{\sigma}{\bar\sigma}\right)\left(\frac{u}{\bar A\sigma}\right)^2 = \frac{1}{4}\cdot\left(\frac{\bar\sigma}{\sigma}\right)\left(\frac{u}{\bar A\bar\sigma}\right)^2 \le \frac{1}{4}\left(\frac{u}{\bar A\bar\sigma}\right)^2,
$$

as we have claimed. It remained to show that $n\bar\sigma^2 \ge \frac{M^{2/3}(L+\beta)\log n}{1000\bar A^{4/3}}$.

This inequality clearly holds under the conditions of Proposition 6.1 if $\sigma \le n^{-1/3}$, since in this case $\log\frac{2}{\sigma} \ge \frac{\log n}{3}$, and $n\bar\sigma^2 \ge \frac{1}{64}\left(\frac{u}{\bar A\sigma}\right)^2 \ge \frac{1}{64}\bar A^{-2}M(L+\beta)^{3/2}\log\frac{2}{\sigma} \ge \frac{\bar A^{-2}}{192}M(L+\beta)\log n \ge \frac{M^{2/3}(L+\beta)\log n}{1000\bar A^{4/3}}$ if $M \ge M_0(\bar A)$ with a sufficiently large number $M_0(\bar A)$.

If $\sigma \ge n^{-1/3}$, then we apply that the inequality $2^{6R}\left(\frac{u}{\bar A\sigma}\right)^2 \le 256 n\sigma^2$ implies that $2^{-4R} \ge 2^{-16/3}\left[\frac{\left(\frac{u}{\bar A\sigma}\right)^2}{n\sigma^2}\right]^{2/3}$, and $n\bar\sigma^2 = 2^{-4R}n\sigma^2 \ge \frac{2^{-16/3}}{\bar A^{4/3}}(n\sigma^2)^{1/3}\left(\frac{u}{\sigma}\right)^{4/3}$. Since $n\sigma^2 \ge n^{1/3}$ and $(\frac{u}{\sigma})^2 \ge \frac{M}{3}(L+\beta)^{3/2}$, these estimates yield that

$$n\bar\sigma^2 \ge \frac{\bar A^{-4/3}}{50}(n\sigma^2)^{1/3}\left(\frac{u}{\sigma}\right)^{4/3} \ge \frac{\bar A^{-4/3}}{50}n^{1/9}\left(\frac{M}{3}\right)^{2/3}(L+\beta) \ge \frac{M^{2/3}(L+\beta)\log n}{1000\bar A^{4/3}}.$$

7. The completion of the proof of Theorem 4.1

In this section we prove Proposition 6.2 by which the proof of Theorem 4.1 is completed. First a symmetrization lemma is proved, and then with the help of this result and a conditioning argument the proof of Proposition 6.2 is reduced to the estimation of a probability which can be bounded by means of the Hoeffding inequality formulated in Theorem 3.4. Such an approach makes possible to prove Proposition 6.2.

First I formulate the symmetrization lemma we shall apply.

Lemma 7.1 (Symmetrization Lemma). *Let Z_n and $\bar Z_n$, $n=1,2,\dots$, be two sequences of random variables independent of each other, and let the random variables $\bar Z_n$, $n=1,2,\dots$, satisfy the inequality*

$$P(|\bar Z_n| \le \alpha) \ge \beta \quad \textit{for all } n=1,2,\dots \tag{7.1}$$

with some numbers $\alpha \ge 0$ and $\beta \ge 0$. Then

$$P\left(\sup_{1\le n<\infty}|Z_n| > \alpha+u\right) \le \frac{1}{\beta}P\left(\sup_{1\le n<\infty}|Z_n-\bar Z_n| > u\right) \quad \textit{for all } u>0.$$

Proof of Lemma 7.1. Put $\tau=\min\{n\colon |Z_n|>\alpha+u\}$ if there exists such an index n, and $\tau=0$ otherwise. Then the event $\{\tau=n\}$ is independent of the sequence of random variables $\bar Z_1,\bar Z_2,\dots$ for all $n=1,2,\dots$, and because of this independence

$$P(\{\tau=n\}) \le \frac{1}{\beta}P(\{\tau=n\}\cap\{|\bar Z_n|\le\alpha\}) \le \frac{1}{\beta}P(\{\tau=n\}\cap\{|Z_n-\bar Z_n|>u\})$$

for all $n=1,2,\dots$. Hence

$$P\left(\sup_{1\le n<\infty}|Z_n|>\alpha+u\right)=\sum_{n=1}^\infty P(\tau=n)\le\frac1\beta\sum_{n=1}^\infty P(\{\tau=n\}\cap\{|Z_n-\bar Z_n|>u\})$$
$$\le\frac1\beta P\left(\sup_{1\le n<\infty}|Z_n-\bar Z_n|>u\right).$$

Lemma 7.1 is proved.

We shall apply the following consequence Lemma 7.2 of the symmetrization lemma.

Lemma 7.2. *Let us fix a countable class of functions $\mathcal F$ on a measurable space $(X,\mathcal X)$ together with a real number $0<\sigma<1$. Consider a sequence of independent, identically distributed X-valued random variables $\xi_1,\dots,\xi_n$ such that $Ef(\xi_1)=0$, $Ef^2(\xi_1)\le\sigma^2$ for all $f\in\mathcal F$ together with another sequence $\varepsilon_1,\dots,\varepsilon_n$ of independent random variables with distribution $P(\varepsilon_j=1)=P(\varepsilon_j=-1)=\frac12$, $1\le j\le n$, independent also of the random sequence $\xi_1,\dots,\xi_n$. Then*

$$P\left(\frac1{\sqrt n}\sup_{f\in\mathcal F}\left|\sum_{j=1}^n f(\xi_j)\right|\ge An^{1/2}\sigma^2\right)$$
$$\le 4P\left(\frac1{\sqrt n}\sup_{f\in\mathcal F}\left|\sum_{j=1}^n\varepsilon_j f(\xi_j)\right|\ge\frac A3 n^{1/2}\sigma^2\right)\quad\textit{if } A\ge\frac{3\sqrt2}{\sqrt n\sigma}. \tag{7.2}$$

Proof of Lemma 7.2. Let us construct an independent copy $\bar\xi_1,\dots,\bar\xi_n$ of the sequence $\xi_1,\dots,\xi_n$ in such a way that all three sequences $\xi_1,\dots,\xi_n$, $\bar\xi_1,\dots,\bar\xi_n$ and $\varepsilon_1,\dots,\varepsilon_n$ are independent. Define the random variables $S_n(f)=\frac1{\sqrt n}\sum\limits_{j=1}^n f(\xi_j)$ and $\bar S_n(f)=\frac1{\sqrt n}\sum\limits_{j=1}^n f(\bar\xi_j)$ for all $f\in\mathcal F$. The inequality

$$P\left(\sup_{f\in\mathcal F}|S_n(f)|>A\sqrt n\sigma^2\right)\le 2P\left(\sup_{f\in\mathcal F}|S_n(f)-\bar S_n(f)|>\frac23 A\sqrt n\sigma^2\right). \tag{7.3}$$

follows from Lemma 7.1 if we apply it for the countable sets $Z_n(f)=S_n(f)$ and $\bar Z_n(f)=\bar S_n(f)$, $f\in\mathcal F$, of random variables and $x=\frac23A\sqrt n\sigma^2$, $\alpha=\frac13A\sqrt n\sigma^2$, since the fields $S_n(f)$ and $\bar S_n(f)$ are independent, and $P(|\bar S_n(f)|\le\alpha)>\frac12$ for all $f\in\mathcal F$. Indeed, $\alpha=\frac13A\sqrt n\sigma^2\ge\sqrt2\sigma$, $E\bar S_n(f)^2\le\sigma^2$, thus Chebishev's inequality implies that $P(|\bar S_n(f)|\le\alpha)\ge P(|\bar S_n(f)|\le\sqrt2\sigma)\ge\frac12$ for all $f\in\mathcal F$.

Let us observe that the random field

$$S_n(f)-\bar S_n(f)=\frac1{\sqrt n}\sum_{j=1}^n\left(f(\xi_j)-f(\bar\xi_j)\right),\quad f\in\mathcal F, \tag{7.4}$$

and its randomization

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j\left(f(\xi_j)-f(\bar\xi_j)\right),\quad f\in\mathcal{F}, \tag{7.4$'$}$$

have the same distribution. Indeed, even the conditional distribution of (7.4′) under the condition that the values of the ε_j-s are prescribed agrees with the distribution of (7.4) for all possible values of the ε_j-s. This follows from the observation that the distribution of the field (7.4) does not change if we exchange the random variables ξ_j and $\bar\xi_j$ for certain indices j, and this corresponds to considering the conditional distribution of the field in (7.4′) under the condition that $\varepsilon_j=-1$ for these indices j, and $\varepsilon_j=1$ for the remaining ones.

The above relation together with formula (7.3) imply that

$$\begin{aligned}
&P\left(\frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^{n}f(\xi_j)\right|\ge An^{1/2}\sigma^2\right)\\
&\qquad\le 2P\left(\frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^{n}\varepsilon_j\left[f(\xi_j)-\bar f(\xi_j)\right]\right|\ge \frac{2}{3}An^{1/2}\sigma^2\right)\\
&\qquad\le 2P\left(\frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^{n}\varepsilon_j f(\xi_j)\right|\ge \frac{A}{3}n^{1/2}\sigma^2\right)\\
&\qquad\qquad+2P\left(\frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^{n}\varepsilon_j f(\bar\xi_j)\right|\ge \frac{A}{3}n^{1/2}\sigma^2\right)\\
&\qquad=4P\left(\frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^{n}\varepsilon_j f(\xi_j)\right|\ge \frac{A}{3}n^{1/2}\sigma^2\right).
\end{aligned}$$

Lemma 7.2 is proved.

Let me briefly explain the approach to the proof of Proposition 6.2. We have to estimate a probability of the form $P\left(n^{-1/2}\sup\limits_{f\in\mathcal{F}}\left|\sum\limits_{j=1}^{n}f(\xi_j)\right|>u\right)$, and by Lemma 7.2 this can be replaced by the estimation of the probability $P\left(n^{-1/2}\sup\limits_{f\in\mathcal{F}}\left|\sum\limits_{j=1}^{n}\varepsilon_j f(\xi_j)\right|>\frac{u}{3}\right)$ with some independent random variables ε_j, $P(\varepsilon_j=1)=P(\varepsilon_j=-1)=\frac12$, $j=1,\dots,n$, which are also independent of the random variables ξ_j. We shall bound the conditional probability of the event appearing in this modified problem under the condition that the values of the random variables ξ_j are prescribed. This can be done with the help of Hoeffding's inequality formulated in Theorem 3.4 and the L_2-density property of the class of functions $\mathcal{F}$ we consider. By working out the details we are led to the estimation of the probability $P\left(n^{-1/2}\sup\limits_{f\in\mathcal{F}'}\left|\sum\limits_{j=1}^n f(\xi_j)\right|>u^{1+\alpha}\right)$ with some new nice L_2-dense class of bounded functions $\mathcal{F}'$ and some number $\alpha>0$. This problem is very similar to the original one, but it is simpler, since the number u is replaced by a larger number $u^{1+\alpha}$ in it. By repeating this argument successively, in finitely many steps we get the proof of Proposition 6.2.

The above sketched argument suggests a backward induction procedure to prove Proposition 6.2. To carry out such a program first we introduce a property we want to prove.

Definition of good tail behaviour for a class of normalized random sums. *Let us fix some measurable space $(X,\mathcal{X})$ and a probability measure μ on it together with some integer $n\ge 2$ and real number $\sigma>0$, and consider some class $\mathcal{F}$ of functions $f(x)$ on the space $(X,\mathcal{X})$. Take a sequence of independent μ distributed random variables $\xi_1,\dots,\xi_n$, and define with its help the normalized random sums $S_n(f)=\frac{1}{\sqrt{n}}\sum\limits_{j=1}^n f(\xi_j)$, $f\in\mathcal{F}$. Given some real number $T>0$ we say that the set of normalized random sums $S_n(f)$ determined by the class of functions $\mathcal{F}$ has a good tail behaviour at level T (with parameters n and σ^2 which we shall fix in the sequel) if the inequality*

$$P\left(\sup_{f\in\mathcal{F}}|S_n(f)|\ge A\sqrt{n}\sigma^2\right)\le\exp\left\{-A^{1/2}n\sigma^2\right\} \tag{7.5}$$

holds for all numbers $A>T$.

Now we formulate Proposition 7.3 and show that Proposition 6.2 follows from it.

Proposition 7.3. *Let us fix a positive integer $n\ge 2$, a real number $\sigma>0$ and a probability measure μ on a measurable space $(X,\mathcal{X})$ together with a countable L_2-dense class $\mathcal{F}$ of functions $f=f(x)$ on the space $(X,\mathcal{X})$ with some prescribed exponent $L\ge 1$ and parameter D. Let us also assume that all functions $f\in\mathcal{F}$ satisfy the conditions $\sup\limits_{x\in X}|f(x)|\le\frac14$, $\int f^2(x)\mu(\,dx)\le\sigma^2$, and $n\sigma^2>K(L+\beta)\log n$ with a sufficiently large fixed number K and $\beta=\max\left(\frac{\log D}{\log n},0\right)$.*

If there is a number $T>1$ such that for all classes of functions $\mathcal{F}$ which satisfy the above conditions the class of normalized random sums $S_n(f)=\frac{1}{\sqrt{n}}\sum\limits_{j=1}^n f(\xi_j)$, $f\in\mathcal{F}$, defined with the help of a sequence of independent μ distributed random variables $\xi_1,\dots,\xi_n$ have a good tail behaviour at level T, then there is a universal constant $\bar A_0$ such that the number $\bar T=T^{3/4}$ also have this property provided that $T\ge\bar A_0$. We can choose for instance $\bar A_0=64\cdot10^{12}$ and $K=1$.

Proposition 6.2 simply follows from Proposition 7.3. To show this let us first observe that the class of normalized random sums $S_n(f)$, $f\in\mathcal{F}$, has a good tail behaviour at level $T_0=\frac{1}{4\sigma^2}$ if the class of functions $\mathcal{F}$ satisfies the conditions of Proposition 7.3.

Indeed, in this case $P\left(\sup\limits_{f\in\mathcal{F}}|S_n(f)|\ge A\sqrt{n}\sigma^2\right)\le P\left(\sup\limits_{f\in\mathcal{F}}|S_n(f)|>\frac{\sqrt{n}}{4}\right)=0$ for all $A>T_0$. Then the repetitive application of Proposition 7.3 yields that the class of random sums $S_n(f)$ has a good tail behaviour at all levels $T\ge T_0^{(3/4)^j}$ if $T_0^{(3/4)^j}\ge\bar A_0$, hence for $T=\bar A_0^{4/3}$ if the class of functions $\mathcal{F}$ satisfies the conditions of Proposition 7.3. If the class of functions $f\in\mathcal{F}$ satisfies the conditions of Proposition 6.2, then the class of functions $\bar{\mathcal{F}}=\left\{\bar f=\frac{f}{4}\colon f\in\mathcal{F}\right\}$ satisfies the conditions of Proposition 7.2, (actually with $\bar\sigma=\frac{\sigma}{4}$, and a better parameter D for the class $\mathcal{F}$), hence the class of functions $S_n(\bar f)$, $\bar f\in\bar{\mathcal{F}}$, has a good tail behaviour at level $T=\bar A_0^{4/3}$. This implies that the original class of functions $\mathcal{F}$ satisfy formula (6.3) in Proposition 6.2 with $4K$, $A_0=4\bar A_0^{4/3}$ and $\gamma=\frac12$, and this is what we had to show.

The proof of Proposition 7.3. Fix a class of functions $\mathcal{F}$ which satisfies the conditions of Proposition 7.3 together with two independent sequences $\xi_1,\dots,\xi_n$ and $\varepsilon_1,\dots,\varepsilon_n$ of independent random variables, where ξ_j is μ-distributed, $P(\varepsilon_j=1)=P(\varepsilon_j=-1)=\frac12$, $1\le j\le n$, and investigate the conditional probability

$$P(f,A|\xi_1,\dots,\xi_n)=P\left(\frac{1}{\sqrt{n}}\left|\sum_{j=1}^n\varepsilon_jf(\xi_j)\right|\ge\frac{A}{6}\sqrt{n}\sigma^2\,\middle|\,\xi_1,\dots,\xi_n\right)$$

for all functions $f\in\mathcal{F}$, $A\ge T$ and values $(\xi_1,\dots,\xi_n)$ in the condition. By the Hoeffding inequality presented in Theorem 3.4

$$P(f,A|\xi_1,\dots,\xi_n)\le2\exp\left\{-\frac{\frac{1}{36}A^2n\sigma^4}{2\bar S^2(f,\xi_1,\dots,\xi_n)}\right\}\tag{7.6}$$

with

$$\bar S^2(f,x_1,\dots,x_n)=\frac1n\sum_{j=1}^nf^2(x_j),\qquad f\in\mathcal{F}.$$

Let us introduce the set

$$H=H(A)=\left\{(x_1,\dots,x_n)\colon\sup_{f\in\mathcal{F}}\bar S^2(f,x_1,\dots,x_n)\ge\left(1+A^{4/3}\right)\sigma^2\right\}.\tag{7.7}$$

I claim that

$$P((\xi_1,\dots,\xi_n)\in H)\le e^{-A^{2/3}n\sigma^2}\qquad\text{if } A\ge T.\tag{7.7$'$}$$

(The set H plays the role of the small exceptional set, where we cannot provide a good estimate for $P(f,A|\xi_1,\dots,\xi_n)$ for some $f\in\mathcal{F}$.)

To prove relation (7.7′) let us consider the functions $\bar f=\bar f(f)$, $\bar f(x)=f^2(x)-\int f^2(x)\mu(\,dx)$, and introduce the class of functions $\mathcal{F}'=\{\bar f(f)\colon f\in\mathcal{F}\}$. Let us show that the class of functions $\mathcal{F}'$ satisfies the conditions of Proposition 7.3, hence the estimate (7.5) holds for the class of functions $\mathcal{F}'$ if $A\ge T^{4/3}$.

The relation $\int \bar f(x)\mu(\,dx)=0$ clearly holds. The condition $\sup|\bar f(x)|\le\frac18<\frac14$ also holds if $\sup|f(x)|\le\frac14$, and $\int \bar f^2(x)\mu(\,dx)\le\int f^4(x)\mu(\,dx)\le\frac1{16}\int f^2(x)\,\mu(\,dx)\le\frac{\sigma^2}{16}<\sigma^2$ if $f\in\mathcal F$. It remained to show that $\mathcal F'$ is an L_2-dense class with exponent L and parameter D.

To show this observe that $\int(\bar f(x)-\bar g(x))^2\rho(\,dx)\le 2\int(f^2(x)-g^2(x))^2\rho(\,dx)+2\int(f^2(x)-g^2(x))^2\mu(\,dx)\le 2(\sup(|f(x)|+|g(x)|)^2\left(\int(f(x)-g(x))^2(\rho(\,dx)+\mu(\,dx)\right)\le\int(f(x)-g(x))^2\bar\rho(\,dx)$ for all $f,g\in\mathcal F$, $\bar f=\bar f(f)$, $\bar g=\bar g(g)$ and probability measure ρ, where $\bar\rho=\frac{\rho+\mu}{2}$. This means that if $\{f_1,\dots,f_m\}$ is an ε-dense subset of $\mathcal F$ in the space $L_2(X,\mathcal X,\bar\rho)$, then $\{\bar f_1,\dots,\bar f_m\}$ is an ε-dense subset of $\mathcal F'$ in the space $L_2(X,\mathcal X,\rho)$, and not only $\mathcal F$, but also $\mathcal F'$ is an L_2-dense class with exponent L and parameter D.

We get by applying the inductive hypothesis of Proposition 7.3 for the number $A^{4/3}\ge T^{4/3}$ and the class of functions $\mathcal F'$ that

$$\begin{aligned}P((\xi_1,\dots,\xi_n)\in H)&=P\left(\sup_{f\in\mathcal F}\left(\frac1n\sum_{j=1}^n\bar f(\xi_j)+\frac1n\sum_{j=1}^n Ef^2(\xi_j)\right)\ge\left(1+A^{4/3}\right)\sigma^2\right)\\&\le P\left(\sup_{\bar f\in\bar{\mathcal F}}\frac1{\sqrt n}\sum_{j=1}^n\bar f(\xi_j)\ge A^{4/3}n^{1/2}\sigma^2\right)\le e^{-A^{2/3}n\sigma^2},\end{aligned}$$

i.e. relation (7.7′) holds.

Formula (7.6) and the definition of the set H given in (7.7) yield the estimate

$$P(f,A|\xi_1,\dots,\xi_n)\le 2e^{-A^{2/3}n\sigma^2/144}\quad\text{if } (\xi_1,\dots,\xi_n)\notin H \tag{7.8}$$

for all $f\in\mathcal F$ and $A\ge T\ge1$. (Here we used the estimate $1+A^{4/3}\le 2A^{4/3}$.) Let us introduce the conditional probability

$$P(\mathcal F,A|\xi_1,\dots,\xi_n)=P\left(\sup_{f\in\mathcal F}\frac1{\sqrt n}\left|\sum_{j=1}^n\varepsilon_jf(\xi_j)\right|\ge\frac A3\sqrt n\sigma^2\,\middle|\,\xi_1,\dots,\xi_n\right)$$

for all $(\xi_1,\dots,\xi_n)$ and $A\ge T$. We shall estimate this conditional probability with the help of relation (7.8) if $(\xi_1,\dots,\xi_n)\notin H$. Given some set of n points $(x_1,\dots,x_n)$ in the space $(X,\mathcal X)$ let us introduce the measure $\nu=\nu(x_1,\dots,x_n)$ on $(X,\mathcal X)$ in such a way that ν is concentrated in the points $x_1,\dots,x_n$, and $\nu(\{x_j\})=\frac1n$. If $\int f^2(x)\nu(\,dx)\le\delta^2$ for a function f, then $\left|\frac1{\sqrt n}\sum\limits_{j=1}^n\varepsilon_jf(x_j)\right|\le n^{1/2}\int|f(x)|\nu(\,dx)\le n^{1/2}\delta$. Since the condition $n\sigma^2\ge K(L+\beta)\log n$ in Proposition 7.3 also implies that $n\sigma^2\ge1$ (if the constant K is chosen sufficiently large), the above estimate implies that if f and g are two functions such that $\int(f-g)^2\nu(\,dx)\le\delta^2$ with $\delta=\frac A{6n}$, then $\left|\frac1{\sqrt n}\sum\limits_{j=1}^n\varepsilon_jf(x_j)-\frac1{\sqrt n}\sum\limits_{j=1}^n\varepsilon_jg(x_j)\right|\le\frac A{6\sqrt n}\le\frac A6\sqrt n\sigma^2$.

Let us fix some (random) point $(\xi_1,\dots,\xi_n)\notin H$, consider the measure $\nu=\nu(\xi_1,\dots,\xi_n)$ corresponding to it and choose a $\bar\delta$-dense subset $\{f_1,\dots,f_m\}$ of $\mathcal{F}$ in the space $L_2(X,\mathcal{X},\nu)$ with $\bar\delta=\frac{1}{6n}\le\delta=\frac{A}{6n}$, whose cardinality m satisfies the inequality $m\le D\bar\delta^{-L}$. This is possible because of the L_2-dense property of the class $\mathcal{F}$. (This is the point where the L_2-dense property of the class of functions $\mathcal{F}$ is exploited in its full strength.) The above facts imply that if $\frac{1}{\sqrt{n}}\left|\sum\limits_{j=1}^n\varepsilon_jf(\xi_j)\right|\ge\frac{A}{3}\sqrt{n}\sigma^2$ for some function $f\in\mathcal{F}$, then $\frac{1}{\sqrt{n}}\left|\sum\limits_{j=1}^n\varepsilon_jf_l(\xi_j)\right|\ge\frac{A}{6}\sqrt{n}\sigma^2$ for some function f_l of the $\bar\delta$-dense subset $\{f_1,\dots,f_m\}$ of $\mathcal{F}$ with the fixed point $(\xi_1,\dots,\xi_n)\notin H$. Hence $P(\mathcal{F},A|\xi_1,\dots,\xi_n)\le\sum\limits_{l=1}^m P(f_l,A|\xi_1,\dots,\xi_n)$ with these functions $f_1,\dots,f_m$, and relation (7.8) yields that

$$P(\mathcal{F},A|\xi_1,\dots,\xi_n)\le 2D(6n)^Le^{-A^{2/3}n\sigma^2/144}\quad\text{if } (\xi_1,\dots,\xi_n)\notin H \text{ and } A\ge T.$$

This inequality together with Lemma 7.2 (under the restriction that $A\ge\bar A_0\ge\frac{3\sqrt2}{\sqrt{n}\sigma}\ge 3\sqrt2$) and estimate (7.7′) imply that

$$\begin{aligned}P\left(\frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^n f(\xi_j)\right|\ge An^{1/2}\sigma^2\right)&\le 4P\left(\frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^n\varepsilon_jf(\xi_j)\right|\ge\frac{A}{3}n^{1/2}\sigma^2\right)\\&\le 8D(6n)^Le^{-A^{2/3}n\sigma^2/144}+4e^{-A^{2/3}n\sigma^2}\quad\text{if } A\ge T.\end{aligned}\tag{7.9}$$

By the condition $n\sigma^2\ge K(L+\beta)\log n=KL\log n+K\log(\max(D,1))$, hence the first term at the right-hand side of (7.9) can be bounded as

$$\begin{aligned}&8D(6n)^Le^{-A^{2/3}n\sigma^2/144}\\&\qquad\le e^{-A^{1/2}n\sigma^2}\cdot 8D6^Ln^{L(1-A^{1/2}/3)}\max(D,1)^{-A^{1/2}/3}\le\tfrac12 e^{-A^{1/2}n\sigma^2}\end{aligned}$$

if $A\ge T\ge\bar A_0\ge 64\cdot10^{12}$ and $K\ge1$. (With such parameters $\frac{A^{2/3}}{144}-A^{1/2}\le\frac13A^{1/2}$.) With such a choice of the parameters the inequality $\frac{3\sqrt2}{\sqrt{n}\sigma}\le\frac{3\sqrt2}{\sqrt{K\log 2}}\le\bar A_0\le A$, needed for the validity of relation (7.2), also holds. The second term at the right-hand side of (7.9) be bounded as $4e^{-A^{2/3}n\sigma^2}\le\frac12e^{-A^{1/2}n\sigma^2}$. with the above choice of the numbers $\bar A_0$ and K.

By the above calculation formula (7.9) yields the inequality

$$P\left(\frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^n f(\xi_j)\right|\ge An^{1/2}\sigma^2\right)\le e^{-A^{1/2}n\sigma^2}$$

if $A\ge T$, and the constants $\bar A_0$ and K are chosen sufficiently large, for instance $\bar A_0=64\cdot10^{12}$ and $K=1$ is an appropriate choice.

8. Formulation of the main results of this work

This section contains the main results of this work about multiple stochastic integrals and their supremum. Section 4 contains these results in the special case of one-fold integrals together with their version about the supremum of appropriate classes of normalized sums of independent and identically distributed random variables with zero expectation. (See Theorem 4.1 and 4.1$'$ and their comparison.) The results about multiple stochastic integrals also have a similar version, and they will be also presented. Here the role of sums of independent, and identically distributed random variables are taken by degenerate U-statistics of independent and identically distributed random variables. The condition that the U-statistics have to be degenerate plays the a role similar to the condition about the zero expectation of the summands when the independent sum versions of the one-fold integral results are considered. The basic notions about U-statistics needed to understand the results will also be explained. The proof of the equivalence of the results about multiple integrals and U-statistics formulated in this section requires a detailed study of the property of U-statistics, a problem which has a special interest in itself. This will be the subject of the next section.

We also formulate some results about multiple Wiener–Itô integrals which are natural analogs of the results about multiple integrals with respect to normalized empirical measures. But these results are only briefly discussed, because they do not belong to the main subject of this work, and they demand a more detailed study of multiple Wiener–Itô integrals. Finally, this section is finished with a the two-dimensional version of Example 3.2 which shows that certain conditions of the results discussed here are really necessary.

Let us consider a sequence of iid. random variables $\xi_1,\dots,\xi_n$ taking values on a measurable space $(X,\mathcal{X})$. Let μ denote its distribution, and introduce the empirical distribution of this sequence defined in (4.5). Given a measurable function $f(x_1,\dots,x_k)$ on the k-fold product space $(X^k,\mathcal{X}^k)$ introduce its integral $J_{n,k}(f)$ with respect to the k-fold product of the normalized empirical measure $\sqrt{n}(\mu_n-\mu)$ defined in formula (4.8). Here we define the domain of integration by deleting the diagonals $x_j=x_l$, $1\le j<l\le k$, from the k-fold product space $(X^k,\mathcal{X}^k)$. The following Theorem 8.1 can be considered as the multiple integral version of Bernstein's inequality formulated in Theorem 3.1.

Theorem 8.1. *Let us take a measurable function $f(x_1,\dots,x_k)$ on the k-fold product $(X^k,\mathcal{X}^k)$ of a measure space $(X,\mathcal{X})$ with some $k\ge 1$ together with a non-atomic probability measure μ on $(X,\mathcal{X})$ and a sequence of iid. random variables $\xi_1,\dots,\xi_n$ with distribution μ on $(X,\mathcal{X})$. Let the function f satisfy the conditions*

$$\|f\|_\infty=\sup_{x_j\in X,\ 1\le j\le k}|f(x_1,\dots,x_k)|\le 1, \tag{8.1}$$

and

$$\|f\|_2^2=\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\le\sigma^2 \tag{8.2}$$

with some constant $0 < \sigma \le 1$. *There exist some constants* $C = C_k > 0$ *and* $\alpha = \alpha_k > 0$, *such that the random integral* $J_{n,k}(f)$ *defined by formulas (4.5) and (4.8) satisfies the inequality*

$$P\left(|J_{n,k}(f)| > u\right) \le C \max\left(\exp\left\{ -\alpha \left(\frac{u}{\sigma}\right)^{2/k} \right\}, \exp\left\{ -\alpha (nu^2)^{1/(k+1)} \right\} \right) \tag{8.3}$$

for all $u > 0$. *The constants* $C = C_k > 0$ *and* $\alpha = \alpha_k > 0$ *in formula (8.3) depend only on the parameter* k.

Theorem 8.1 can be reformulated in the following equivalent form.

Theorem 8.1′. *Under the conditions of Theorem 8.1*

$$P\left(|J_{n,k}(f)| > u\right) \le C \exp\left\{ -\alpha \left(\frac{u}{\sigma}\right)^{2/k} \right\} \quad \text{for all } 0 < u \le n^{k/2}\sigma^{k+1} \tag{8.3′}$$

with the number σ *appearing in (8.2) and some universal constants* $C = C_k > 0$, $\alpha = \alpha_k > 0$, *depending only on the multiplicity* k *of the integral* $J_{n,k}(f)$.

Theorem 8.1 clearly implies Theorem 8.1′, since in the case $u \le n^{k/2}\sigma^{k+1}$ the first term is larger than the second one in the maximum at the right-hand side of formula (8.3). On the other hand Theorem 8.1′ implies Theorem 8.1 also if $u > n^{k/2}\sigma^{k+1}$, since in this case Theorem 8.1′ can be applied with $\bar\sigma = \left(un^{-k/2}\right)^{1/(k+1)} \ge \sigma$. This yields that $P\left(|J_{n,k}(f)| > u\right) \le C \exp\left\{ -\alpha \left(\frac{u}{\bar\sigma}\right)^{2/k} \right\} = C \exp\left\{ -\alpha (nu^2)^{1/(k+1)} \right\}$ if $u > n^{k/2}\sigma^{k+1}$.

Theorem 8.1 or Theorem 8.1′ state that the tail probability $P(|J_{n,k}(f)| > u)$ of the k-fold random integral $J_{n,k}(f)$ can be bounded similarly to the probability $P(|\text{const.}\, \sigma\eta^k| > u)$, where η is a random variable with standard normal distribution and σ is the number appearing in relation (8.2), provided that the level u we consider is less than $n^{k/2}\sigma^{k+1}$. (The value of the number σ^2 in formula (8.2) is closely related to the variance of $J_{n,k}(f)$.) At the end of this section an example is given which shows that such a condition is really needed in the above results.

Now we formulate Theorem 8.2 which is the generalization of Theorem 4.1 for multiple integrals. Here we apply the notions of L_2-dense classes and countably approximability introduced in Section 4.

Theorem 8.2. *Let us have a non-atomic 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* $(X^k, \mathcal{X}^k)$ *which satisfies the conditions*

$$\|f\|_\infty = \sup_{x_j \in X,\ 1 \le j \le k} |f(x_1, \dots, x_k)| \le 1, \qquad \text{for all } f \in \mathcal{F} \tag{8.4}$$

and

$$\|f\|_2^2 = Ef^2(\xi_1,\dots,\xi_k) = \int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k) \le \sigma^2 \qquad \text{for all } f \in \mathcal{F} \tag{8.5}$$

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 (4.8) satisfies the inequality*

$$P\left(\sup_{f\in\mathcal{F}} |J_{n,k}(f)| \ge u\right) \le CD \exp\left\{-\alpha \left(\frac{u}{\sigma}\right)^{2/k}\right\} \qquad \text{if} \quad n\sigma^2 \ge \left(\frac{u}{\sigma}\right)^{2/k} \ge M(L+\beta)^{3/2}\log\frac{2}{\sigma}, \tag{8.6}$$

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* $\mathcal{F}$*.*

The condition about the countable cardinality of the class $\mathcal{F}$ *can be replaced by the weaker condition that the class of random variables* $J_{n,k}(f)$*,* $f \in \mathcal{F}$*, is countably approximable.*

To formulate that version of Theorems 8.1 and 8.2 which corresponds to the results about sums of independent random variables in the case $k = 1$ let us introduce the following notions:

Definition of U-statistics. *Let us consider a function* $f = f(x_1,\dots,x_k)$ *on the* k*-th power* $(X^k,\mathcal{X}^k)$ *of a space* $(X,\mathcal{X})$ *together with a sequence of independent and identically distributed random variables* $\xi_1,\dots,\xi_n$*,* $n \ge k$*, which take their values on this space* $(X,\mathcal{X})$*. The expression*

$$I_{n,k}(f) = \frac{1}{k!} \sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'} \text{ if } j\ne j'}} f\left(\xi_{l_1},\dots,\xi_{l_k}\right) \tag{8.7}$$

is called a U*-statistic of order* k *with the sequence* $\xi_1,\dots,\xi_n$*, and* f *is called its kernel function.*

To make our later notation non-ambiguous let us also consider functions of the form $f(x_{u_1},\dots,x_{u_k})$*, that is let us allow the possibility that the variables of the function* f *which take their values in the space* $(X,\mathcal{X})$ *are indexed in a general way. In the case of such an indexation we define*

$$I_{n,k}(f) = \frac{1}{k!} \sum_{\substack{1\le l_{u_j}\le n,\ j=1,\dots,k\\ l_{u_j}\ne l_{u'_j} \text{ if } j\ne j'}} f\left(\xi_{l_{u_1}},\dots,\xi_{l_{u_k}}\right). \tag{8.7'}$$

A similar convention will be applied in the definition of decoupled U-statistics introduced later, and the following definition of degenerate U-statistics and canonical functions can also be similarly reformulated in the case of general indexation.

The degenerate U-statistics which correspond to sums of identically distributed random variables with expectation zero constitute an important subclass of the U-statistics. We define it together with the notion of canonical kernel function which is closely related to it.

Definition of degenerate U-statistics. *A U-statistic $I_{n,k}(f)$ of order k with a sequence of iid. random variables $\xi_1,\dots,\xi_n$ is called degenerate if its kernel function $f(x_1,\dots,x_k)$ satisfies the relation*

$$
\begin{aligned}
&Ef(\xi_1,\dots,\xi_k|\xi_1=x_1,\dots,\xi_{j-1}=x_{j-1},\xi_{j+1}=x_{j+1},\dots,\xi_k=x_k)=0\\
&\qquad\qquad \text{for all } 1\le j\le k \text{ and } x_s\in X,\ s\ne j.
\end{aligned}
$$

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

$$
\int f(x_1,\dots,x_{j-1},u,x_{j+1},\dots,x_k)\mu(\,du)=0 \quad \text{for all } 1\le j\le k \quad \text{and} \quad x_s\in X,\ s\ne j. \tag{8.8}
$$

It is clear that a U-statistic $I_{n,k}(f)$ with kernel function f and independent μ-distributed random variables $\xi_1,\dots,\xi_n$ is degenerate if and only if its kernel function is canonical with respect to the probability measure μ. Now we can formulate two results about U-statistics which are, as we shall see in the next section, equivalent to Theorems 8.1 and 8.2.

Theorem 8.3. *Let us have a measurable function $f(x_1,\dots,x_k)$ on the k-fold product $(X^k,\mathcal{X}^k)$, $k\ge1$, of a measure space $(X,\mathcal{X})$ with some $k\ge1$ together with a probability measure μ on $(X,\mathcal{X})$ and a sequence of iid. random variables $\xi_1,\dots,\xi_n$ with distribution μ on $(X,\mathcal{X})$. Let us consider the k-fold U-statistic $I_{k,n}(f)$ with this sequence of random variables $\xi_1,\dots,\xi_n$. Assume that this U-statistic is degenerate, i.e. the kernel function $f(x_1,\dots,x_k)$ of this U-statistic is canonical with respect to the measure μ. Let us also assume that the function f satisfies conditions (8.1) and (8.2) with some number $0<\sigma\le1$. Then there exist some constants $C=C_k>0$ and $\alpha=\alpha_k>0$ such that the inequality*

$$
P\left(n^{-k/2}|I_{n,k}(f)|>u\right)\le C\exp\left\{-\alpha\left(\frac{u}{\sigma}\right)^{2/k}\right\} \tag{8.9}
$$

holds for all $0<u\le n^{k/2}\sigma^{k+1}$. The constants $C=C_k>0$ and $\alpha=\alpha_k>0$ depend only on the parameter k.

Theorem 8.4. *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 $(X^k, \mathcal{X}^k)$ which satisfies conditions (8.4) and (8.5) with some constant $0 < \sigma \le 1$. Besides these conditions let us assume that for a sequence of independent μ distributed random variables $\xi_1, \dots, \xi_n$ the U-statistics $I_{n,k}(f)$ with this sequence $\xi_1, \dots, \xi_n$ are degenerate for all $f \in \mathcal{F}$, or in an equivalent form all functions $f \in \mathcal{F}$ are canonical with respect to the measure μ. 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 inequality

$$
\begin{aligned}
P\left(\sup_{f\in\mathcal{F}} n^{-k/2}|I_{n,k}(f)| \ge u\right) &\le CD\exp\left\{-\alpha\left(\frac{u}{\sigma}\right)^{2/k}\right\} \\
&\qquad \text{if} \quad n\sigma^2 \ge \left(\frac{u}{\sigma}\right)^{2/k} \ge M(L+\beta)^{3/2}\log\frac{2}{\sigma},
\end{aligned}
\tag{8.10}
$$

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

The condition about the countable cardinality of the class $\mathcal{F}$ can be replaced by the weaker condition that the class of random variables $n^{-k/2}I_{n,k}(f)$, $f \in \mathcal{F}$, is countably approximable.

Let us briefly describe the Gaussian counterpart of the above results. Here some basic notions and results about multiple Wiener–Itô integrals are applied. But since the results about these Gaussian fields do not belong to the main subject of this work, they are mainly interesting for us for the sake of a comparison, most technical details will be omitted from our discussion.

Let us consider a measurable space $(X, \mathcal{X})$ together with a non-atomic σ-finite measure μ on it. Let Z_μ be an orthogonal Gaussian random measure with counting measure μ on $(X, \mathcal{X})$, i.e. assume that the random variables $Z_\mu(A)$, $A \in \mathcal{X}$, $\mu(A) < \infty$ are defined, they are jointly Gaussian, $EZ_\mu(A) = 0$ for all $A \in \mathcal{A}$, $\mu(A) < \infty$ and $EZ_\mu(A)Z_\mu(B) = \mu(A \cap B)$ for all $A \in \mathcal{A}$, $B \in \mathcal{A}$, $\mu(A) < \infty$, $\mu(B) < \infty$.

Let us observe that these relations imply that if $A \in \mathcal{X}$, $\mu(A) < \infty$ and $B \in \mathcal{X}$, $\mu(B) < \infty$ are disjoint sets, then $Z_\mu(A)$ and $Z_\mu(B)$ are independent, and $Z_\mu(A \cup B) = Z_\mu(A) + Z_\mu(B)$ with probability 1. The last relation follows from the fact that $E(Z_\mu(A \cup B) - Z_\mu(A) - Z_\mu(B))^2 = 0$ under these conditions.

If $f(x_1, \dots, x_k)$ is a measurable function on $(X^k, \mathcal{X}^k)$ such that

$$
\int f^2(x_1, \dots, x_k)\mu(\,dx_1)\dots\mu(\,dx_k) < \infty,
$$

then the multiple Wiener–Itô integral $Z_{\mu,k}(f) = \frac{1}{k!}\int f(x_1, \dots, x_k) Z_\mu(\,dx_1)\dots Z_\mu(\,dx_k)$ can be defined, and it satisfies similar estimates as the random integrals $J_{n,k}(f)$. This statement will be formulated more explicitly in the following Theorem 8.5.

Theorem 8.5 *Let us fix a measurable space $(X, \mathcal{X})$ together with a σ-finite non-atomic measure μ on it, and let Z_μ be an orthogonal Gaussian random measure with counting* *measure μ on $(X,\mathcal{X})$. If $f(x_1,\dots,x_k)$ is a measurable function on $(X^k,\mathcal{X}^k)$ such that $\frac{1}{k!}\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\le\sigma^2$ with some $0<\sigma<\infty$, then*

$$P(|Z_{\mu,k}(f)|>u)\le C\exp\left\{-\alpha\left(\frac{u}{\sigma}\right)^{2/k}\right\} \tag{8.11}$$

for all $u>0$ with some constants $C=C(k)$ and $\alpha=\alpha(k)$ depending only on k.

If $\mathcal{F}$ is a countable class of functions of k variables on $(X,\mathcal{X})$ such that

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

and there exist some constant $D>0$ and $L>0$ such that a subset $\{f_1,\dots,f_m\}\subset\mathcal{F}$ can be chosen with $m\le D\varepsilon^{-L}$ elements for which

$$\min_{1\le j\le m}\int\left(f(x_1,\dots,x_k)-f_j(x_1,\dots,x_k)\right)^2\mu(\,dx_1)\dots\mu(\,dx_k)\le\varepsilon \quad \text{for all } f\in\mathcal{F},$$

then the inequality

$$P\left(\sup_{f\in\mathcal{F}}|Z_{\mu,k}(f)|>u\right)\le C(D+1)\exp\left\{-\alpha\left(\frac{u}{\sigma}\right)^{2/k}\right\} \quad \textit{if } u\ge ML^{k/2}\sigma\log^{k/2}\frac{2}{\sigma} \tag{8.12}$$

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

Since the above result does not belong to the main part of this work, only a sketchy proof will be presented. Nelson's inequality, mentioned at the start of this section will be formulated and proved in the Appendix, and it will be explained how Theorem 8.5 can be proved with its help.

The above results show that multiple integrals with respect to a normalized empirical measure or degenerate U-statistics satisfy some estimates similar to multiple Wiener–Itô integrals, but they hold under more restrictive conditions. This difference between multiple integrals with respect to a normalized empirical measure and orthogonal Gaussian measures can be explained similarly to some arguments presented in Section 4 about the one-fold integral case. Here we do not repeat them, we only give an example similar to Example 3.2 which shows that the condition $u\le n^{k/2}\sigma^{k+1}$ cannot be dropped from the conditions of Theorem 8.2. For the sake of simplicity we restrict our attention to the case $k=2$.

Example 8.6. *Let $\xi_1,\dots,\xi_n$ be a sequence of independent, identically distributed random variables taking values on the plane $R^2=X$ such that $\xi_j=(\eta_{j,1},\eta_{j,2})$, $\eta_{j,1}$ and $\eta_{j,2}$ are independent, $P(\eta_{j,1}=1)=P(\eta_{j,1}=-1)=\frac{\sigma^2}{8}$, $P(\eta_{j,1}=0)=1-\frac{\sigma^2}{4}$, $P(\eta_{j,2}=1)=P(\eta_{j,2}=-1)=\frac{1}{2}$ for all $1\le j\le n$, 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*

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

of order 2 with the above kernel function f and the sequence of independent random variables $\xi_1,\dots,\xi_n$. Then $I_{n,2}(f)$ is a degenerate U-statistic. If $u\ge B_1n\sigma^3$ with some appropriate constant $B_1>0$, $\bar B_2^{-1}n\ge u\ge \bar B_2 n^{-2}$ with a sufficiently large fixed number $\bar B_2>0$ and $\sigma\ge\frac1n$, then the estimate

$$P(n^{-1}I_{n,2}(f)>u)\ge \exp\left\{-Bn^{1/3}u^{2/3}\log\left(\frac{u}{n\sigma^3}\right)\right\} \tag{8.13}$$

holds with some $B>0$.

Remark: The main content of the above example is that in the case $k=2$ the condition $\frac u\sigma\le n\sigma^2$ cannot be dropped from Theorem 8.3. Let us observe that in the case $u=n\sigma^3$ the right-hand side of (8.13) has the same order as Theorem 8.3 suggests. (In this model $\int f^2(x,y)\mu(\,dx)\mu(\,dy)=E(2\eta_{j,1}\eta_{j,2})^2=\sigma^2$.) If we consider the probability in (8.13) at the same level u, but with a much smaller parameter σ^2, then the probability at the right-hand side of (8.13) has a relatively small decrease, and the estimate of Theorem 8.3 does not hold any longer. Let me also remark that under some mild additional restrictions the estimate (8.13) can be slightly improved, the term log can be replaced by $\log^{2/3}$ in the exponent of the right-hand side of (8.13). To get this improvement some more calculation is needed, and the numbers u_1 and u_2 in the following calculations have to be replaced by $v_1=8n^{1/3}u^{2/3}\log^{-1/3}\left(\frac{u}{n\sigma^3}\right)$ and $v_2=\frac14n^{2/3}u^{1/3}\log^{1/3}\left(\frac{u}{n\sigma^3}\right)$.

It is simple to check that the U-statistic we considered in the above example is degenerate because of the independence properties of the model and the relation $E\eta_{j,1}=E\eta_{j,2}=0$. In the proof of the estimate (8.13) we shall apply for one hand the results of Section 3, in particular Example 3.2 for the sequence $\eta_{j,1}$, $j=1,2,\dots,n$, on the other hand the following result from the theory of large deviations: If $X_1,\dots,X_n$ are iid. random variables, $P(X_1=1)=P(X_1=-1)=\frac12$, then for any number $0\le\alpha<1$ there exists some numbers $C_1=C_1(\alpha)>0$ and $C_2=C_2(\alpha)>0$ such that $P\left(\sum\limits_{j=1}^n X_j>u\right)\ge C_1e^{-C_2u^2/n}$ for all $0\le u\le\alpha n$.

Proof of the statement of the example. We can write

$$P(n^{-1}I_{n,2}(f)>u)\ge P\left(2\left(\sum_{j=1}^n\eta_{j,1}\right)\left(\sum_{j=1}^n\eta_{j,2}\right)>2nu\right)-P\left(2\sum_{j=1}^n\eta_{j,1}\eta_{j,2}>nu\right). \tag{8.14}$$

Because of the independence of the random variables $\eta_{j,1}$ and $\eta_{j,2}$ the first probability at the right-hand side of (8.14) can be bounded from below with the choice $v_1=8n^{1/3}u^{2/3}$ and $v_2=\frac18n^{2/3}u^{1/3}$ by means of Example 3.2. (The estimate of Example 3.2 can be applied with the choice $y=v_1$, since by the inequality $\frac n8\ge v_1\ge n\sigma^2$ the conditions of Example 3.2 are satisfied), together with the large-deviation result mentioned after the remark. These estimates together yield that

$$P\left(\left(\sum_{j=1}^n \eta_{j,1}\right)\left(\sum_{j=1}^n \eta_{j,2}\right) > 2nu\right) \ge P\left(\sum_{j=1}^n \eta_{j,1} > v_1\right) P\left(\sum_{j=1}^n \eta_{j,2} > v_2\right)$$
$$\ge \exp\left\{-B_1 v_1 \log\left(\frac{v_1}{n\sigma^2}\right) - B_2\frac{v_2^2}{n}\right\} \ge \exp\left\{-B_3 n^{1/3}u^{2/3}\log\left(\frac{u}{n\sigma^3}\right)\right\}$$

with appropriate constants $B_1>0$, $B_2>0$ and $B_3>0$. On the other hand by applying Bennett's inequality, more precisely its consequence given in formula (3.4) for the sum of the random variables $X_j = 2\eta_{j,1}\eta_{j,2}$ and $y = nu$ we get the following upper bound for the second term at the right-hand side of (8.14):

$$P\left(2\sum_{j=1}^n \eta_{j,1}\eta_{j,2} > nu\right) \le \exp\left\{-B_4 nu \log\frac{u}{\sigma^2}\right\}$$
$$\le \exp\left\{-2B_5 n^{1/3}u^{2/3}\log\left(\frac{u}{n\sigma^3}\right)\right\},$$

since $nu \ge \bar B n^{1/3}u^{2/3} \ge \bar B n\sigma^2$, and the estimate (3.4) is applicable if $\bar B$ is sufficiently large. The above estimates imply the statement of the example.

9. Some results about U-statistics

This section contains the proof of an important result about U-statistics, the so-called Hoeffding decomposition theorem which states that all U-statistics can be represented as a sum of degenerate U-statistics. Let us consider the kernel function of a U-statistic together with the kernel functions of the U-statistics in its Hoeffding decomposition. It will also be shown that the L_2-norm of the kernel functions of the U-statistics in the Hoeffding decomposition are bounded by the L_2-norm of the kernel function of the original U-statistic. Besides, if a class of U-statistics is given with an L_2-dense class of kernel functions (with the same underlying sequence of independent and identically distributed random variables) and the Hoeffding decomposition of all of these U-statistics is taken, then the kernel functions of the degenerate U-statistics taking part in the Hoeffding decomposition also constitute an L_2-dense class. Another important result of this section is a decomposition of a k-fold random integral with respect to a normalized empirical measure to the linear combination of degenerate U-statistics presented in Theorem 9.4. These results enable us to prove the equivalence of Theorem 8.1 with Theorem 8.3 and of Theorem 8.2 with Theorem 8.4. They are also useful in the proof of Theorems 8.3 and 8.4.

In the special case $k=1$ Hoeffding's decomposition means that the sum $S_n = \sum\limits_{j=1}^n \xi_j$ of iid. random variables can be rewritten as $S_n = \sum\limits_{j=1}^n (\xi_j - E\xi_j) + \left(\sum\limits_{j=1}^n E\xi_j\right)$, i.e. the sum of independent random variables with zero expectation plus a constant. We may consider a constant as a U-statistic of order zero. (For the sake of a simpler terminology in the sequel let us consider a constant as a degenerate U-statistic of order zero, and define $I_{n,0}(c) = c$ for a constant c.) I wrote down this trivial calculation, because Hoeffding's decomposition is actually an adaptation of this procedure to the general case. To understand this let us see how to adapt this construction in the case $k = 2$. In this case we have to consider a sum of the form $I_{n,2}(f) = \sum\limits_{1\le j,k\le n, j\ne k} f(\xi_j, \xi_k)$. Write $f(\xi_j, \xi_k) = [f(\xi_j, \xi_k) - Ef(\xi_j, \xi_k|\xi_k)] + Ef(\xi_j, \xi_k|\xi_k) = f_1(\xi_j, \xi_k) + \bar f_1(\xi_k)$ with $f_1(\xi_j, \xi_k) = f(\xi_j, \xi_k) - Ef(\xi_j, \xi_k|\xi_k)$, and $\bar f_1(\xi_k) = Ef(\xi_j, \xi_k|\xi_k)$ to achieve that the conditional expectation of $f_1(\xi_j, \xi_k)$ for fixed ξ_k be zero. Repeating this procedure for the first coordinate we define $f_2(\xi_j, \xi_k) = f_1(\xi_j, \xi_k) - Ef_1(\xi_j, \xi_k|\xi_j)$ and $\bar f_2(\xi_j) = Ef_1(\xi_j, \xi_k|\xi_j)$. Simple calculation shows that $I_{n,2}(f_2)$ is a degenerate U-statistics of order 2, and the identity $I_{n,2}(f) = I_{n,2}(f_2) + I_{n,1}((n-1)(\bar f_1 - E\bar f_1)) + I_{n,1}((n-1)((\bar f_2 - E\bar f_2)) + n(n-1)E(\bar f_1 + \bar f_2)$ yields the decomposition of $I_{n,2}(f)$ for sums of degenerate U-statistics.

We get the Hoeffding decomposition by working out the details of the above argument in the general case. But it is simpler to calculate the appropriate conditional expectations with the help of the kernel functions of the U-statistics. To carry out such a program in the study of U-statistics of order k we introduce the following notations.

Let us consider the k-fold product $(X^k, \mathcal{X}^k, \mu^k)$ of a measure space $(X, \mathcal{X}, \mu)$ with some probability measure μ, and define for all integrable functions $f(x_1, \dots, x_k)$ and indices $1 \le j \le k$ the projection $P_j f$ of the function f to its j-th coordinate as

$$P_j f(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k) = \int f(x_1, \dots, x_k)\mu(\,dx_j), \quad 1 \le j \le k. \tag{9.1}$$

Let us also define the operators $Q_j = I - P_j$ as $Q_j f = f - P_j f$ on the space of integrable functions on $(X^k, \mathcal{X}^k, \mu^k)$, $1 \le j \le k$. In the definition (9.1) $P_j f$ is a function not depending on the coordinate x_j, but in the definition of Q_j we introduce the fictive coordinate x_j to make the expression $Q_j f = f - P_j f$ meaningful. Now we can formulate the following result.

Theorem 9.1 (Hoeffding decomposition). *Let $f(x_1, \dots, x_k)$ be an integrable function on the k-fold product space $(X^k, \mathcal{X}^k, \mu^k)$ of a space $(X, \mathcal{X}, \mu)$ with a probability measure μ. It has the decomposition*

$$f = \sum_{V \subset \{1,\dots,k\}} f_V, \quad \text{with} \quad f_V(x_j, \, j \in V) = \left(\prod_{j \in \{1,\dots,k\} \setminus V} P_j \prod_{j \in V} Q_j \right) f(x_1, \dots, x_k) \tag{9.2}$$

such that all functions f_V, $V \subset \{1, \dots, k\}$, in (9.2) are canonical with respect to the probability measure μ, and they depend on the $|V|$ arguments x_j, $j \in V$.

Let $\xi_1, \dots, \xi_n$ be a sequence of independent μ distributed random variables, and consider the U-statistics $I_{n,k}(f)$ and $I_{n,|V|}(f_V)$ corresponding to the kernel functions f,

f_V defined in (9.2) and random variables $\xi_1,\dots,\xi_n$. Then

$$I_{n,k}(f)=\sum_{V\subset\{1,\dots,k\}}(n-|V|)(n-|V|-1)\cdots(n-k+1)\frac{|V|!}{k!}I_{n,|V|}(f_V) \tag{9.3}$$

is a representation of $I_{n,k}(f)$ as a sum of degenerate U-statistics, where $|V|$ denotes the cardinality of the set V. (The product $(n-|V|)(n-|V|-1)\cdots(n-k+1)$ is defined as 1 for $V=\{1,\dots,k\}$, i.e. if $|V|=k$.) This representation is called the Hoeffding decomposition of $I_{n,k}(f)$.

The proof of Theorem 9.1. Write $f=\prod\limits_{j=1}^{k}(P_j+Q_j)f$. By carrying out the multiplications in this identity and applying the commutativity of the operators P_j and Q_j for different indices j we get formula (9.2). To show that the functions f_V in formula (9.2) are canonical let us observe that this property can be rewritten in the form $P_jf_V=0$ (in all coordinates x_s, $s\in V\setminus\{j\}$ if $j\in V$). Since $P_j=P_j^2$, and the identity $P_jQ_j=P_j-P_j^2=0$ holds for all $j\in\{1,\dots,k\}$ this relation follows from the above mentioned commutativity of the operators P_j and Q_j, as $P_jf_V=\left(\prod\limits_{s\in\{1,\dots,k\}\setminus V}P_s\prod\limits_{s\in V\setminus\{j\}}Q_s\right)P_jQ_jf=0$. By applying the identity (9.2) for all terms $f(\xi_{j_1},\dots,\xi_{j_k})$ in the sum defining the U-statistic $I_{n,k}(f)$ and then summing them up we get relation (9.3).

The next result enables us to estimate the kernel functions of the degenerate U-statistics in the Hoeffding-decomposition of a U-statistic by means of the properties kernel function of the original U-statistic.

Theorem 9.2. *Let $f(x_1,\dots,x_k)$ be a square integrable function on the k-fold product space $(X^k,\mathcal{X}^k,\mu^k)$, and take its decomposition defined in formula (9.2). The inequalities*

$$\int f_V^2(x_j,\,j\in V)\prod_{j\in V}\mu(\,dx_j)\le\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k) \tag{9.4}$$

and

$$\sup_{x_j,\,j\in V}|f_V(x_j,\,j\in V)|\le 2^{|V|}\sup_{x_j,\,1\le j\le k}|f(x_1,\dots,x_k)| \tag{9.4$'$}$$

hold for all $V\subset\{1,\dots,k\}$.

Let us consider an L_2-dense class $\mathcal{F}$ of functions with parameter D and exponent L on the space $(X^k,\mathcal{X}^k)$, take the decomposition (9.2) of all functions $f\in\mathcal{F}$ and define the classes of functions $\mathcal{F}_V=\{2^{-|V|}f_V\colon f\in\mathcal{F}\}$ for all $V\subset\{1,\dots,k\}$ with the help of the functions f_V taking part in this decomposition. The classes of functions $\mathcal{F}_V$ are also L_2-dense with the same parameter D and exponent L for all $V\subset\{1,\dots,k\}$.

Theorem 9.2 is a fairly simple consequence of Proposition 9.3 presented below. To formulate it first we introduce the following notations:

Let us consider the product $(Y \times Z, \mathcal{Y} \times \mathcal{Z})$ of two measurable spaces $(Y, \mathcal{Y})$ and $(Z, \mathcal{Z})$ together with a probability measure μ on $(Z, \mathcal{Z})$ and the operator

$$Pf(y) = P_\mu f(y) = \int f(y,z)\mu(\,dz), \quad y \in Y,\ z \in Z \tag{9.5}$$

for all measurable functions f on the space $Y \times Z$ for which this integral is finite. Let I denote the identity operator on the space of functions on $Y \times Z$, i.e. let $If(y,z) = f(y,z)$, and introduce the operator $Q = Q_\mu = I - P = I - P_\mu$ which maps the functions f on the space $Y \times Z$ to functions on the space $Y \times Z$ given by the formula

$$\begin{aligned} Q_\mu f(y,z) = (I - P_\mu)f(y,z) = f(y,z) - P_\mu f(y,z) = f(y,z) - \int f(y,z)\mu(\,dz), \\ y \in Y,\ z \in Z. \end{aligned} \tag{9.6}$$

(Here, and in the sequel we shall sometimes identify a function $g(y)$ defined on the space $(Y, \mathcal{Y})$ with the function $\bar g(y,z) = g(y)$ on the space $(Y \times Z, \mathcal{Y} \times \mathcal{Z})$ which actually does not depend on the coordinate z.) The following result will be proved:

Proposition 9.3. *Let us consider the direct product $(Y \times Z, \mathcal{Y} \times \mathcal{Z})$ of two measure spaces $(Y, \mathcal{Y})$ and $(Z, \mathcal{Z})$ together with a probability measure μ on the space $(Z, \mathcal{Z})$. Take the transformations P_μ and Q_μ defined in formulas (9.5) and (9.6). Given any probability measure ρ on the space $(Y, \mathcal{Y})$ consider the product measure $\rho \times \mu$ on $(Y \times Z, \mathcal{Y} \times \mathcal{Z})$. Then the transformations P_μ and Q_μ, as maps from the space $L_2(Y \times Z, \mathcal{Y} \times \mathcal{Z}, \mu \times \rho)$ to $L_2(Y, \mathcal{Y}, \rho)$ and $L_2(Y \times Z, \mathcal{Y} \times \mathcal{Z}, \rho \times \mu)$ respectively, have a norm less than or equal to 1, i.e.*

$$\int P_\mu f(y)^2 \rho(\,dy) \le \int f(y,z)^2 \rho(\,dy)\mu(\,dz), \tag{9.7}$$

and

$$\int Q_\mu f(y,z)^2 \rho(\,dy)\mu(\,dz) \le \int f(y,z)^2 \rho(\,dy)\mu(\,dz) \tag{9.8}$$

for all functions $f \in L_2(Y \times Z, \mathcal{Y} \times \mathcal{Z}, \rho \times \mu)$.

If $\mathcal{F}$ is an L_2-dense class of functions $f(y,z)$ in the product space $(Y \times Z, \mathcal{Y} \times \mathcal{Z})$, with parameter D and exponent L, then also the classes $\mathcal{F}_\mu = \{P_\mu f,\ f \in \mathcal{F}\}$ and $\mathcal{G}_\mu = \{\frac12 Q_\mu f = \frac12 (f - P_\mu f),\ f \in \mathcal{F}\}$ are L_2-dense classes with the same exponent L and parameter D in the spaces $(Y, \mathcal{Y})$ and $(Y \times Z, \mathcal{Y} \times \mathcal{Z})$ respectively.

The following corollary of Proposition 9.3 is formally more general, but it is a simple consequence of this result. Actually we shall need this corollary.

Corollary of Proposition 9.3. *Let us consider the product space $(Y_1 \times Z \times Y_2, \mathcal{Y}_1 \times \mathcal{Z} \times \mathcal{Y}_2)$, a probability measure μ on the space $(Z, \mathcal{Z})$ and define the transformations*

$$P_\mu f(y_1, y_2) = \int f(y_1, z, y_2)\mu(\,dz), \quad y_1 \in Y_1,\ z \in Z,\ y_2 \in Y_2 \tag{9.5'}$$

and

$$Q_\mu f(y_1,z,y_2)=(I-P_\mu)f(y_1,z,y_2)=f(y_1,z,y_2)-P_\mu f(y_1,z,y_2)$$
$$=f(y_1,z,y_2)-\int f(y_1,z,y_2)\mu(\,dz),\quad y_1\in Y_1,\ z\in Z,\ y_2\in Y_2 \tag{9.6$'$}$$

for the measurable functions f on the space $Y_1\times Z\times Y_2$. Then

$$\int P_\mu f(y_1,y_2)^2\rho(\,dy_1,\,dy_2)\le\int f(y,z)^2(\rho\times\mu)(\,dy_1,\,dz,\,dy_2), \tag{9.7$'$}$$

for all probability measures ρ on $(Y_1\times Y_2,\mathcal{Y}_1\times\mathcal{Y}_2)$, where $\rho\times\mu$ is the product of the probability measure ρ on $(Y_1\times Y_2,\mathcal{Y}_1\times\mathcal{Y}_2)$ and μ on $(Z,\mathcal{Z})$, i.e. $\rho\times\mu(\{y_1,z,y_2)\colon(y_1,y_2)\in A,z\in B\})=\rho(A)\mu(B)$ for all $A\in\mathcal{Y}_1\times\mathcal{Y}_2$, $B\in\mathcal{Z}$, and $\rho\times\mu$ is its unique extension as a probability measure on $(Y_1\times Z\times Y_2,\mathcal{Y}_1\times\mathcal{Z}\times\mathcal{Y}_2)$. Also the inequality

$$\int Q_\mu f(y_1,z,y_2)^2\rho(\,dy_1,\,dy_2)\mu(\,dz)\le\int f(y_1,z,y_2)^2\rho(\,dy_1,\,dy_2)\mu(\,dz) \tag{9.8$'$}$$

holds for all functions $f\in L_2(Y\times Z,\mathcal{Y}\times\mathcal{Z},\rho\times\mu)$.

If $\mathcal{F}$ is an L_2-dense class of functions $f(y_1,z,y_2)$ in the product space $(Y_1\times Z\times Y_2,\mathcal{Y}_1\times\mathcal{Z}\times Y_2)$, with parameter D and exponent L, then also the classes $\mathcal{F}_\mu=\{P_\mu f,\ f\in\mathcal{F}\}$ and $\mathcal{G}_\mu=\{\frac12 Q_\mu f=\frac12(f-P_\mu f),\ f\in\mathcal{F}\}$ are L_2-dense classes with exponent L and parameter D in the spaces $(Y_1\times Y_2,\mathcal{Y}_1\times\mathcal{Y}_2)$ and $(Y_1\times Z\times Y_2,\mathcal{Y}_1\times\mathcal{Z}\times\mathcal{Y}_2)$ respectively.

This corollary is a simple consequence of Proposition 9.3 if we apply it with $(Y,\mathcal{Y})=(Y_1\times Y_2,\mathcal{Y}_1\times\mathcal{Y}_2)$ and take the natural mapping $f((y_1,y_2),z)\to f(y_1,z,y_2)$ of a function from the space $(Y\times Z,\mathcal{Y}\times\mathcal{Z})$ to a function on $(Y_1\times Z\times Y_2,\mathcal{Y}_1\times\mathcal{Z}\times\mathcal{Y}_2)$, and use the correspondence between the product measure $\rho\times\mu$ in these spaces.

Proposition 9.3, more precisely its corollary implies Theorem 9.2, since it implies that the operators P_s, Q_s, $1\le s\le k$, applied in Theorem 9.2 do not increase the $L_2(\mu)$ norm of a function f, and it is also clear that the norm of P_s is bounded by 1 the norm of $Q_s=I-P_s$ is bounded by 2 as an operator from L_∞ spaces to L_∞ spaces. The corollary of Proposition 9.3 also implies that if $\mathcal{F}$ is an L_2-dense class of functions with parameter D and exponent L, then the same property holds for the classes of functions $\mathcal{F}_{P_s}=\{P_sf\colon f\in\mathcal{F}\}$ and $\mathcal{F}_{Q_s}=\{\frac12 O_sf\colon f\in\mathcal{F}\}$, $1\le s\le k$. These relations together with the identity $f_V=\left(\prod\limits_{s\in V}P_s\prod\limits_{s\in\{1,\dots,k\}\setminus V}Q_s\right)f$ imply Theorem 9.2.

Proof of Proposition 9.3. The Schwarz inequality yields that $P_\mu(f)^2\le\int f(y,z)^2\mu(\,dz)$, and integrating this inequality with respect to the probability measure $\rho(\,dy)$ we get inequality (9.7). Also the inequality

$$\int Q_\mu f(y,z)^2\rho(dy)\mu(dz)=\int[f(y,z)-P_\mu f(y,z)]^2\rho(du)\mu(dz)\le\int f(y,z)^2\rho(dy)\mu(dz)$$

holds, and this is relation (9.8). It follows for instance from the observation that the functions $f(y,z)-P_\mu f(y,z)$ and $P_\mu f(y,z)$ are orthogonal in the space $L_2(Y\times Z,\mathcal{Y}\times\mathcal{Z},\rho\times\mu)$.

Let us consider an arbitrary probability measure ρ on the space $(Y,\mathcal{Y})$. To prove that $\mathcal{F}_\mu$ is an L_2-dense class with parameter D and exponent L we have to find $m\le D\varepsilon^L$ functions $f_j\in\mathcal{F}_\mu$, $1\le j\le m$, such that $\inf\limits_{1\le j\le m}\int(f_j-f)^2\,d\rho\le\varepsilon^2$ for all $f\in\mathcal{F}_\mu$. But a similar property holds in the space $Y\times Z$ with the probability measure $\rho\times\mu$. This property together with the L_2 contraction property of P_μ formulated in (9.7) imply that $\mathcal{F}_\mu$ is an L_2-dense class.

To prove that $\mathcal{G}_\mu$ is also L_2-dense with parameter D and exponent L we have to find for all numbers $0<\varepsilon\le1$ and probability measures ρ on $Y\times Z$ a subset $\{g_1,\dots,g_m\}\subset\mathcal{G}_\mu$ with $m\le D\varepsilon^{-L}$ elements such that $\inf\limits_{1\le j\le m}\int(g_j-g)^2\,d\rho\le\varepsilon^2$ for all $g\in\mathcal{G}_\mu$.

Let us consider the probability measure $\tilde\rho=\frac12(\rho+\bar\rho\times\mu)$ on $(Y\times Z,\mathcal{Y}\times\mathcal{Z})$, where $\bar\rho$ is the projection of the measure ρ to $(Y,\mathcal{Y})$, i.e. $\bar\rho(A)=\rho(A\times Z)$ for all $A\in\mathcal{Y}$, take a class of function $\mathcal{F}_0(\varepsilon,\tilde\rho)=\{f_1,\dots,f_m\}\in\mathcal{F}$, $m\le D\varepsilon^{-L}$ such that $\inf\limits_{1\le j\le m}\int(f_j-f)^2\,d\tilde\rho\le\varepsilon^2$ for all $f\in\mathcal{F}$, and put $\{g_1,\dots,g_m\}=\{\frac12Q_\mu f_1,\dots,\frac12Q_\mu f_m\}$. All functions $g\in\mathcal{G}_\mu$ can be written in the form $g=\frac12Q_\mu f$ with some $f\in\mathcal{F}$, and there exists some function $f_j\in\mathcal{F}_0(\varepsilon,\tilde\rho)$ such that $\int(f-f_m)^2\,d\tilde\rho\le\varepsilon^2$. Hence to complete the proof of Proposition 9.3 it is enough to show that $\int\frac14(Q_\mu f-Q_\mu\bar f)^2\,d\rho\le\int(f-\bar f)^2\,d\tilde\rho$ for all pairs $f,\bar f\in\mathcal{F}$. This inequality holds, since $\int\frac14(Q_\mu f-Q_\mu\bar f)^2\,d\rho\le\int\frac12(f-\bar f)^2\,d\rho+\int\frac12(P_\mu f-P_\mu\bar f)^2\,d\rho$, and $\int(P_\mu f-P_\mu\bar f)^2\,d\rho=\int(P_\mu f-P_\mu\bar f)^2\,d\bar\rho\le\int(f-\bar f)^2\,d(\bar\rho\times\mu)$ by formula 9.7. The above relations imply that $\int\frac14(Q_\mu f-Q\mu\bar f)^2\,d\rho\le\int(f-\bar f)^2\frac12 d\,(\rho+\bar\rho\times\mu)=\int(f-\bar f)^2d\,\tilde\rho$ as we have claimed.

Let us turn to the proof of the equivalence of Theorem 8.1′ with 8.3 and of Theorem 8.2 with 8.4. In Theorems 8.2 and 8.4 we can restrict our attention to the case when the class of functions $\mathcal{F}$ is countable, since the case of countably approximable classes can be simply reduced to this situation. Let us remark that the integration with respect to the measure $\mu_n-\mu$ in the definition (4.8) of the integral $J_{n,k}(f)$ means some kind of normalization, and no such normalization appears in the definition of the U-statistics $I_{n,k}(f)$. This is the cause why degenerate U-statistics had to be considered in Theorems 8.3 and 8.4. The deduction of these results from Theorems 8.1′ and 8.2 is fairly simple if the underlying probability measure μ is non-atomic, since in this case the identity $I_{n,k}(f)=J_{n,k}(f)$ holds for a canonical function with respect to the measure μ. Let us remark that the non-atomic property of the measure μ is needed in this argument not only because of the conditions of Theorems 8.1′ and 8.2, but since in the proof of the above relation we need the identity $\int f(x_1,\dots,x_k)\mu(\,dx_j)=0$ in the case when the domain of integration is a set of the form $X\setminus\{x_1,\dots,x_{j-1},x_{j+1},\dots,x_k\}$.

The case of possibly atomic measures μ can be simply reduced to the case of non-atomic measures by means of the following enlargement of the space $(X,\mathcal{X},\mu)$. Let us introduce the product space $(\bar X,\bar{\mathcal{X}},\bar\mu)=(X,\mathcal{X},\mu)\times([0,1],\mathcal{B},\lambda)$, where $\mathcal{B}$ is the σ-algebra and λ is the Lebesgue measure on $[0,1]$. Define the function $\bar f((x_1,u_1),\dots,(x_k,u_k))=$

$f(x_1,\dots,x_k)$ on this enlarged space. Then $I_{n,k}(f)=I_{n,k}(\bar f)$, and the measure $\bar\mu=\mu\times\lambda$ is non-atomic. Hence we can deduce the estimates of Theorems 8.3 and 8.4 from Theorems 8.1′ and 8.2 by deducing them first for their counterpart in the above constructed enlarged space and the above defined functions.

The deduction of Theorems 8.1′ and 8.2 from Theorems 8.3 and 8.4 requires more work. Let us observe that an integral $J_{n,k}(f)$ can be written as a sum of U-statistics of different order, and by applying the Hoeffding decomposition for each term in this sum we can express the integral $J_{n,k}(f)$ as a sum of degenerate U-statistics. We show that the coefficients of the degenerate U-statistics in the above representation have relatively small coefficients. This is the content of the following Theorem 9.4. To make its content more understandable I formulated its main statement in the case of random integrals of multiplicity two in a more explicit form.

Theorem 9.4. *Let us have a non-atomic measure μ on a measurable space $(X,\mathcal X)$ together with a sequence of independent, μ-distributed random variables $\xi_1,\dots,\xi_n$, and take a function $f(x_1,\dots,x_k)$ of k variables on the space $(X^k,\mathcal X^k)$ such that*

$$\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)<\infty.$$

Let us consider the empirical distribution function μ_n of the sequence $\xi_1,\dots,\xi_n$ introduced in (4.5) together with the k-fold random integral $J_{n,k}(f)$ of the function f defined in (4.8). The identity

$$J_{n,k}(f)=\sum_{V\subset\{1,\dots,k\}}C(n,k,V)n^{-|V|/2}I_{n,|V|}(f_V), \tag{9.9}$$

holds with the set of (canonical) functions $f_V(x_j,\ j\in V)$ (with respect to the measure μ) defined in formula (9.2) together with some real numbers $C(n,k,V)$, $V\subset\{1,\dots,k\}$, where $I_{n,|V|}(f_V)$ denotes the (degenerate) U-statistic of order $|V|$ with the random variables $\xi_1,\dots,\xi_n$ and kernel function f_V. The constants $C(n,k,V)$ in formula (9.9) satisfy the inequality $|C(n,k,V)|\le C(k)$ with some constant $C(k)$ depending only on the order k of the integral $J_{n,k}(f)$. The relations $\lim\limits_{n\to\infty}C(n,k,V)=C(k,V)$ with some appropriate constant such that $0\le|C(k,V)|<\infty$ and $C(n,k,\{1,\dots,k\})=1$ for $V=\{1,\dots,k\}$ also hold.

Remark: Some considerations show that the coefficients $C(n,k,V)$ in formula (9.9) depend only on the cardinality $|V|$ of the set V, i.e. we can write $C(n,k,V)=C(n,k,|V|)$. We shall not need this observation.

Theorems 8.1′ and 8.2 can be simply deduced from Theorems 8.3 and 8.4 respectively with the help of Theorem 9.4. Indeed, to deduce Theorem 8.1′ we can write with the help of formula 9.9

$$P(|J_{n,k}(f)|>u)\le\sum_{V\subset\{1,\dots,k\}}P\left(n^{-|V|/2}|I_{n,|V|}(f_V)|>\frac{u}{2^kC(k)}\right) \tag{9.10}$$

with a constant $C(k)$ satisfying the inequality $C(n,k,|V|)\le C(k)$ for all coefficients $C(n,k,|V|)$ in (9.9). Then we get Theorem 8.1′ from Theorem 8.3 and relations (9.4) and (9.4′) in Theorem 9.2 by which the L_2-norm of the functions f_V are bounded by the L_2-norm of the function f and the L_∞-norm of f_V is bounded by the $2^{|V|}$-times the L_∞-norm or f if we estimate each term at the right-hand side of (9.10) by means of Theorem 8.3. Here we may assume that $2^kC(k)>1$ and let us first assume that also the inequality $\frac{u}{2^kC(k)\sigma}\ge 1$ holds. In this case we get formula (8.3′) in Theorem 8.1′ by estimating each term at the right-hand side of (9.10). Observe that $\exp\left\{-\alpha\left(\frac{u}{2^kC(k)\sigma}\right)^{2/s}\right\}\le\exp\left\{-\alpha\left(\frac{u}{2^kC(k)\sigma}\right)^{2/k}\right\}$ for all $s\le k$ if $\frac{u}{2^kC(k)\sigma}\ge 1$. If $\frac{u}{2^kC(k)\sigma}\le 1$, then formula (8.3′) holds with a sufficiently large $C>0$.

Theorem 8.2 can be similarly deduced from Theorem 8.4 if we observe that relation (9.10) remains valid if we replace $|J_{n,k}(f)|$ by $\sup\limits_{f\in\mathcal{F}}|J_{n,k}(f)|$ and $|I_{n,|V|}(f_V)|$ by $\sup\limits_{f_V\in\mathcal{F}_V}|I_{n,|V|}(f_V)|$ in it, and the constant M in formula (8.6) of Theorem 8.2 is chosen sufficiently large. The only difference is that now we have to exploit besides formulas (9.4) and (9.4′) of Theorem 9.2 the last statement of this result which tells that if $\mathcal{F}$ is an L_2-dense class of functions on a space $(X^k,\mathcal{X}^k)$, then the classes of functions $\mathcal{F}_V=\{2^{-|V|}f_V\colon f\in\mathcal{F}\}$ are also L_2-dense classes of functions for all $V\subset\{1,\dots,k\}$ with the same exponent and parameter.

In the definition of the random integrals $J_{n,k}(f)$ we have integrated in all coordinates with respect to the signed measure $\mu_n-\mu$, and this means some kind of normalization. Thus it is not surprising that the tail behaviour of the distribution of $J_{n,k}(f)$ is similar to that of certain degenerate U-statistics. Theorem 9.4 formulates such a relation. Formula (9.9) expresses the random integral $J_{n,k}(f)$ as a linear combination of degenerate U-statistics of different order. It is similar to the Hoeffding decomposition in that respect that the functions f_V in formula (9.9) agree with the functions f_V appearing in the Hoeffding decomposition of the U-statistic $I_{n,k}(f)$ with kernel function f. But the coefficients in the expansion (9.9) are small. On the other hand, these coefficients need not disappear. In particular, the expansion (9.9) may contain a non-zero constant term. In such a case the expected value $EJ_{n,k}(f)$ may not equal zero, but it can be bounded by a number not depending on the sample size n. In the next example I show that there are really random integrals $J_{n,k}(f)$ such that $EJ_{n,k}(f)\ne 0$.

Let us choose a sequence of independent random variables $\xi_1,\dots,\xi_n$ with uniform distribution on the unit interval, let μ_n denote its empirical distribution, let $f=f(x,y)$ denote the indicator function of the unit square, i.e. let $f(x,y)=1$ if $0\le x,y\le 1$, and $f(x,y)=0$ otherwise. Let us consider the random integral $J_{n,2}(f)=n\int_{x\ne y}f(x,y)(\mu_n(\,dx)-\,dx)(\mu_n(\,dy)-dy)$, and calculate its expected value $EJ_{n,2}(f)$. By adjusting the diagonal $x=y$ to the domain of integration and taking out the contribution obtained in this way we get that $EJ_{n,2}(f)=nE(\int_0^1(\mu_n(\,dx)-\mu(\,dx))^2-n^2\cdot\frac{1}{n^2}=-1$. (The last term is the integral of the function $f(x,y)$ on the diagonal $x=y$ with respect to the product measure $\mu_n\times\mu_n$ which equals $(\mu_n-\mu)\times(\mu_n-\mu)$ on the diagonal.)

The above considerations and the proof of Theorem 9.4 indicate that the equivalence between Theorems 8.1′ and 8.3 or between Theorems 8.2 and 8.4 is not self-evident. It is simpler to prove Theorems 8.3 and 8.4 of these theorem pairs about degenerate U-statistics, and this will be done in this work. On the other hand, Theorems 8.1′ and 8.2 seem to be more appropriate for applications, since here we do not have to restrict our attention to special, canonical kernel functions.

The proof of Theorem 9.4. Let us first introduce the (random) probability measures $\mu^{(l)}$, $1\le l\le n$, concentrated in the sample points ξ_l, i.e. let $\mu^{(l)}(A)=1$ if $\xi_l\in A$, and $\mu^{(l)}(A)=0$ if $\xi_l\notin A$, $A\in\mathcal{A}$. Then $\mu_n-\mu=\frac1n\left(\sum\limits_{l=1}^n\left(\mu^{(l)}-\mu\right)\right)$, and formula (4.8) can be rewritten as

$$J_{n,k}(f)=\frac{1}{n^{k/2}k!}\sum_{(l_1,\dots,l_k),\,1\le l_j\le n,\,1\le j\le k}\int' f(x_1,\dots,x_k) \left(\mu^{(l_1)}(\,dx_1)-\mu(\,dx_1)\right)\dots\left(\mu^{(l_k)}(\,dx_k)-\mu(\,dx_k)\right). \tag{9.11}$$

To rearrange the above sum in a way more appropriate for us let us introduce the class of all partitions $\mathcal{P}=\mathcal{P}_k$ of the set $\{1,2,\dots,k\}$. For a partition $P=\{R_1,\dots,R_u\}$ $\bigcup\limits_{j=1}^u R_j=\{1,\dots,k\}$, $R_j\cap R_l=\emptyset$, $1\le j<l\le u$, the sets R_j, $1\le j\le u$, will be called the components of the partition P. Given a sequence $(l_1,\dots,l_k)$, $1\le l_j\le n$, $1\le j\le k$, of length k let $P_H(l_1,\dots,l_k)$ denote that partition of $\mathcal{P}_k$ in which two points s and t, $1\le s,t\le k$, belong to the same component if and only if $l_s=l_t$. For a partition $P\in\mathcal{P}_k$ let us define the set of sequences $\mathcal{H}(P)=\mathcal{H}_n(P)$ as $\mathcal{H}(P)=\{(l_1,\dots,l_k)\colon 1\le l_j\le n,\,1\le j\le k, P_H(l_1,\dots,l_k)=P\}$.

Let us rewrite formula (9.11) in the form

$$J_{n,k}(f)=\frac{1}{n^{k/2}k!}\sum_{P\in\mathcal{P}}\sum_{(l_1,\dots,l_k)\colon\,(l_1,\dots,l_k)\in\mathcal{H}(P)}\int' f(x_1,\dots,x_k) \left(\mu^{(l_1)}(\,dx_1)-\mu(\,dx_1)\right)\dots\left(\mu^{(l_k)}(\,dx_k)-\mu(\,dx_k)\right). \tag{9.12}$$

Let us remember that the diagonals $x_s=x_t$, $s\ne t$, were omitted from the domain of integration in the formula defining $J_{n,k}(f)$. This implies that in the case $l_s=l_t$ the measure $\mu^{(l_s)}(\,dx_s)\mu^{(l_t)}(\,dx_t)$ has zero measure in the domain of integration. We have to understand the cancellation effects caused by this relation. It will be shown that because of these cancellations the expression in formula (9.12) can be rewritten as a linear combination of degenerate U-statistics with not too large coefficients. Besides, it will be seen from the calculations that the same degenerate U-statistics $I_{n,|V|}(f_V)$ appear in this representation of $J_{n,k}(f)$ which were defined in formula (9.2). This seems to be a natural approach, but the detailed proof demands some rather unpleasant calculations.

Let us fix some partition $P\in\mathcal{P}$ and investigate the integrals in the internal sum at the right-hand side of (9.12) corresponding to the sequences $(l_1,\dots,l_k)\in\mathcal{H}(P)$. For the sake of better understanding let us first consider such a partition $P \in \mathcal{P}$ which has a component of the form $\{1,\dots,s\}$ with some $s \ge 2$. The products of measures by which we have to integrate in this case contain a part of length s of the form $\big(\mu^{(l)}(dx_1)-\mu(dx_1)\big)\dots\big(\mu^{(l)}(dx_s)-\mu(dx_s)\big)$ This part of the product measure can be rewritten in the domain of integration as

$$
\begin{aligned}
&\sum_{j=1}^{s}(-1)^{s-1}\mu(\,dx_1)\dots\mu(\,dx_{j-1})\mu^{(l)}(\,dx_j)\mu(\,dx_{j+1})\dots\mu(\,dx_s)+(-1)^s\mu(dx_1)\dots\mu(dx_s)\\
&\qquad=\sum_{j=1}^{s}(-1)^{s-1}\mu(\,dx_1)\dots\mu(\,dx_{j-1})(\mu^{(l)}(\,dx_j)-\mu(\,dx_l))\mu(\,dx_{j+1})\dots\mu(\,dx_s)\\
&\qquad\qquad\qquad+(-1)^{s-1}(s-1)\mu(dx_1)\dots\mu(dx_s).
\end{aligned}
\tag{9.13}
$$

Here we exploit that all other terms of this product disappear in the domain of integration which does not contain the diagonals. Let us also observe that the term $(-1)^{s-1}(s-1)\mu(dx_1)\dots\mu(dx_j)$ appears n-times if we sum up for all $1\le l\le n$. We have assumed that $s\ge 2$, since the case $s=1$ is slightly different. In this case only the term $\mu^{(l)}(dx_1)-\mu(\,dx_1)$ appears, i.e. have to put no additional term consisting only of (deterministic) measures μ.

More generally, let us fix some partition $P=\{R_1,\dots,R_u\}$, consider the integral corresponding to a sequence $(l_1,\dots,l_k)\in\mathcal{H}(P)$ in the internal sum of (9.12), and let us rewrite it as the sum of integrals with respect to product measures with components of the form $\mu^{(l_s)}(\,dx_s)-\mu(\,dx_s)$ or $\mu(\,dx_s)$, where all measures $\mu^{(l_s)}$ appearing in a product measure are different. Such a representation can be given, similarly to the argument of relation (9.13), only the notations will be more complicated. To write down what we get first we define a class of subsets $\mathcal{T}(P)$ of the set $\{1,\dots,k\}$ depending on the partition $P=\{R_1,\dots,R_u\}$ together with a subclass $\bar{\mathcal{T}}(P)$ of it. Let $\mathcal{T}(P)$ consist of all such sets $\{j_1,\dots,j_{u'}\}\subset\{1,\dots,k\}$, $u'\le u$, for which all numbers $j_1,\dots,j_{u'}$ belong to a different component of the partition P. Let $\bar{\mathcal{T}}(P)\subset\mathcal{T}(P)$ consist of those sets $V=\{j_1,\dots,j_{u'}\}\in\mathcal{T}(P)$ which also satisfy the following additional condition: If some components $R_t=\{b_t\}$, $1\le t\le u$, of the partition P consists of only one point, then the sets V belonging to $\bar{\mathcal{T}}(P)\subset\mathcal{T}(P)$ contain this point b_t. With the help of the above quantities we can write in the case $(l_1,\dots,l_k)\in\mathcal{H}(P)$, similarly to the calculation in (9.13),

$$
\begin{aligned}
&\int' f(x_1,\dots,x_k)\left(\mu^{(l_1)}(\,dx_1)-\mu(\,dx_1)\right)\dots\left(\mu^{(l_k)}(\,dx_k)-\mu(\,dx_k)\right) \qquad (9.14)\\
&\qquad=\sum_{V\in\bar{\mathcal{T}}(P)}\alpha(V,P)\int f(x_1,\dots,x_k)\prod_{j\in V}\left(\mu^{(l_j)}(\,dx_j)-\mu(\,dx_j)\right)\prod_{j'\in\{1,\dots,k\}\setminus V}\mu(\,dx_{j'})
\end{aligned}
$$

with some appropriate finite constants $\alpha(V,P)$. These constants could be calculated explicitly, but it is enough for us to know that they depend only on the partition P and the set $V\in\bar{\mathcal{T}}(P)$. (Actually it was important for us to observe that we get a term with non-zero coefficient at the right-hand side of (9.14) only for $V \in \bar{\mathcal{T}}(P)$, and the class of functions $\bar{\mathcal{T}}(P)$ was introduced because of this reason. This property in the decomposition of the integral (9.14) holds, since in the case of a one-point component $R_t = \{b_t\}$ of the partition P only the term $\mu^{(l_{b_t})}(\,dx_{b_t}) - \mu(\,dx_{b_t})$ appears in the component of product of measures in (9.14), a component of the form $\mu(\,dx_{b_t})$ is missing.)

Let me remark that at the right-hand side of (9.14) I wrote $\int$ instead of integral $\int'$, i.e. I did not omit the diagonal from the domain of integration. This is allowed, since the measure μ is non-atomic, and this also has the consequence that the sample points $\xi_1, \dots, \xi_n$ are different with probability 1.

Formula (9.14) can be rewritten, by expressing its right-hand side with the help of the random variables ξ_l instead of the measures $\mu^{(l)}$ as

$$\int' f(x_1,\dots,x_k)\left(\mu^{(l_1)}(\,dx_1)-\mu(\,dx_1)\right)\dots\left(\mu^{(l_k)}(\,dx_k)-\mu(\,dx_k)\right) \tag{9.15}$$

$$= \sum_{V\in\bar{\mathcal{T}}(P)} \alpha(V,P)\left(\left(\prod_{j'\in\{1,\dots,k\}\setminus V} P_{\mu,j'} \prod_{j\in V} Q_{\mu,j}\right) f\right)(\xi_{l_j},\, j\in V).$$

Here $Q_{\mu,j} = I - P_{\mu,j}$ is the operator Q_μ defined in (9.6′), with the choice Y_1 which is the product of the first $j-1$ components of X^k, Z is the j-th component and Y_2 is the product of the last $k-j$ components of the product space X^k. The operator $P_{\mu,j'}$ is the operator P_μ defined in (9.5′) with the choice of Y_1 as the product of the first $j'-1$, Z the j-th component and Y_2 as the procuct of the last $k-j'$ components of the space X^k. To see why formula (9.15) holds we have to understand that integration with respect to $\left(\mu^{(l_j)}(\,dx_j)-\mu(\,dx_j)\right)$ means the application of the operator $Q_{\mu,j}$ and then putting the value ξ_{l_j} in the argument x_j, while integration with respect to $\mu(\,dx_{j'})$ means the application of the operator $P_{\mu,j'}$. Besides, the operators $Q_{\mu,j}$ and $P_{\mu,j'}$ are exchangeable.

Let us fix some partition $P \in \mathcal{P}_k$, a set $V \in \bar{\mathcal{T}}(P)$ and sum up the expressions at the right-hand side of (9.15) with this set V for all sequences $(l_1,\dots,l_k) \in \mathcal{H}(P)$. We get that

$$\alpha(V,P)\sum_{(l_1,\dots,l_k)\in\mathcal{H}(P)}\left(\prod_{j'\in\{1,\dots,k\}\setminus V} P_{\mu,j'} \prod_{j\in V} Q_{\mu,j}\right) f(\xi_{l_j},\, j\in V) = \bar{\alpha}(V,P,k,n) I_{n,|V|}(f_V) \tag{9.16}$$

where $I_{n,|V|}$ is a U-statistic of order $|V|$ with the kernel function $f_V(x_j, j\in V) = \left(\prod_{j'\in\{1,\dots,k\}\setminus V} P_{\mu,j'} \prod_{j\in V} Q_{\mu,j}\right) f$ with our function on $f \in (X^k, \mathcal{X}^k)$, and the coefficients $\bar{\alpha}(V,P,k,n)$ at the right-hand side of (9.16) (which could be calculated explicitly, but we do not need this formula) satisfy the inequality $|\bar{\alpha}(V,P,k,n)| \le D(k) n^{\beta(P,V)}$, where $\beta(P,V) = u - |V|$ is the number of those components R_j, $1 \le j \le u$, of the partition P for which $R_j \cap V = \emptyset$, and the constant $D(k) < \infty$ depends only on the multiplicity k of the integral $J_{n,k}(f)$.

To understand why $\bar\alpha(V,P,k,n)$ can be bounded by $D(k)n^{\beta(P,V)}$ let us observe that if we first fix the coordinates l_j, $j\in V$, and sum up for the remaining indices $l_{j'}$, $j'\notin V$, at the left-hand side of (9.16), then we get the term depending on the variables ξ_{l_j}, $j\in V$, in the sum defining the U-statistic $I_{n,|V|}(f_V)$ multiplied by $\bar\alpha(V,P,k)$. To get a good estimate on $\bar\alpha(V,P,k,n)$ we have to bound the number of choices for the non-fixed coordinates $l_{j'}$, $j'\notin V$. For this aim let us consider the class of vectors $(l_1,\dots,l_k)\in\mathcal{H}(P)$. Two coordinates $l_{j'}$ and $l_{j''}$ must agree if their indices j' and j'' belong to the same component of the partition P. Besides, if the number j is contained in such a component R_t of the partition P for which $R_t\cap V\neq\emptyset$, then the coordinate l_j of these vectors is fixed. Hence the value $l_{j'}$ of those non-fixed coordinates whose indices j' belong to the same component R_t of the partition P agree and only such components R_t have to be considered for which $R_t\cap V=\emptyset$. This yields the upper bound $n^{\beta(P,V)}$ for the number of possible choices of the indices $l_{j'}$, $j'\notin V$. A more careful consideration shows that the finite limit

$$C(k,V,P)=\lim_{n\to\infty} n^{-\beta(P,V)}\bar\alpha(V,P,k,n), \qquad |C(k,V,P)|<\infty \tag{9.17}$$

also exists.

We get by applying relation (9.12) and summing up relation (9.16) first for all $V\in\bar{\mathcal{T}}(P)$ for a partition $P\in\mathcal{P}_k$ and then for all $P\in\mathcal{P}$ that the identity

$$J_{n,k}(f)=\sum_{V\subset\{1,2,\dots,k\}} C(n,k,V)n^{-|V|/2}\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\\ l_j\neq l_{j'} \text{ if } j\neq j' \text{ for } j\in V}} f_V(\xi_{l_j},\, j\in V) \tag{9.18}$$

holds with the functions

$$f_V(x_j,\, j\in V)=\left(\prod_{j\in V} Q_{\mu,j}\prod_{j'\in\{1,\dots,k\}\setminus V} P_{\mu,j'}\right) f \quad \text{for all } V\subset\{1,\dots,k\} \tag{9.19}$$

and some coefficients $C(n,k,V)$. We shall show that these coefficients satisfy the inequality $|C(n,k,V)|\le C(k)$ with some constant $C(k)>0$. Besides, it is not difficult to see that the identity $C(n,k,\{1,\dots,k\})=1$ holds. To see that the estimate $|C(n,k,V)|\le C(k)$ really holds, observe that $n^{-|V|/2}C(n,k,|V|)$ can be written as a sum of finitely many terms, (the number of terms can be bounded by a number depending only on k) such that all of them can be bounded by a number of the form $D(k)n^{-k/2+\beta(P,V)}$ with some partition P and the number $\beta(P,V)$ introduced after formula (9.16) with some $P\in\mathcal{P}_k$ and $V\in\bar{\mathcal{T}}(P)$. Hence it is enough to show that $-\frac{k}{2}+\beta(P,V)\le-\frac{|V|}{2}$, i.e. $\beta(P,V)\le\frac{k-|V|}{2}$ if $V\in\bar{\mathcal{T}}(P)$. This relation clearly holds, since $\beta(P,V)$ is the number of components of a partition of a set with cardinality less than or equal to $k-|V|$, and all components of this partition have a cardinality at least 2.

Relation (9.18) can be rewritten as $J_{n,k}(f)=\sum\limits_{V\subset\{1,2,\dots,k\}} C(n,k,V)n^{-|V|/2}I_{n,|V|}(f_V)$, where $I_{n,|V|}(f_V)$ is the U-statistic with the random variables $\xi_1,\dots,\xi_n$ and the kernel function f_V defined in (9.19) agrees with the function f_V defined in (9.2). We have also seen that the coefficients $C(n,k,V)$ satisfy the inequality stated in Theorem 9.4. Relation (9.17) together with the bound on the terms $\beta(P,V)$ also imply that the finite limits $\lim\limits_{n\to\infty} C(n,k,V)=C(k,V)$ also exist. Theorem 9.4 is proved.

I formulate two corollaries of Theorem 9.4. The first one explains the content of conditions (8.2) and (8.5) in Theorems 8.1—8.4.

Corollary 1 of Theorem 9.4. *If $I_{n,k}(f)$ is a degenerate U-statistic of order k with some kernel function f, then $E\left(n^{-k/2}I_{n,k}(f)\right)^2\le \frac{1}{k!}\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)$, where μ is the distribution of the random variables taking part in the definition of the U-statistic $I_{n,k}(f)$. Analogously, the k-fold multiple random integral $J_{k,n}(f)$ satisfies the inequality $E\left(n^{-k/2}J_{n,k}(f)\right)^2\le \bar C(k)\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)$ with some constant $\bar C(k)$ depending only on the order k of the integral $J_{n,k}(f)$.*

Proof of Corollary 1 of Theorem 9.4. We have

$$E(n^{-k/2}I_{n,k}(f))^2=\frac{1}{(k!)^2n^k}\sum{}' Ef(\xi_{l_1},\dots,\xi_{l_s})f(\xi_{l'_1},\dots,\xi_{l'_s}),$$

where the prime in $\sum'$ means that summation is taken for such pairs of k-tuples $(l_1,\dots,l_k)$, $(l'_1,\dots,l'_k)$, $1\le l_j,l'_j\le n$, for which $l_j\ne l_{j'}$ and $l'_j\ne l'_{j'}$ if $j\ne j'$. The degeneracy of the U-statistic $I_{n,k}(f)$ implies that $Ef(\xi_{l_1},\dots,\xi_{l_s})f(\xi_{l'_1},\dots,\xi_{l'_s})=0$ if the two k-tuples $(l_1,\dots,l_s)$ and $(l'_1,\dots,l'_s)$ differ. This can be seen by taking such an index l_j from the first k-tuple which does not appear in the second one, and by observing that the conditional expectation of the product we consider equals zero by the degeneracy condition of the U-statistic under the condition that the value of all random variables except that of ξ_{l_j} is fixed in this product. There remains $k!n(n-1)\cdots(n-k+1)$ terms in the sum expressing $E\left(n^{-k/2}I_{n,k}(f)\right)^2$ which may be non-zero, and all of them can be bounded by $\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)$ because of our conditions and the Schwarz inequality. These estimates yield the bound given for $E\left(n^{-k/2}I_{n,k}(f)\right)^2$.

We can simply get the bound for $J_{n,k}(f)$ with the help of Theorem 9.4, formula (9.4) in Theorem 9.4 by which the L_2-norm of the functions f_V can be bounded by the L_2-norm of the function f and the bound given for the second moment of degenerate U-statistics $n^{-|V|/2}I_{n,|V|}(f_V)$ appearing in the expansion (9.9).

In Corollary 2 the decomposition (9.9) of a random integral $J_{n,2}(f)$ of order 2 is described in an explicit way.

Corollary 2 of Theorem 9.4. *Let the random integral $J_{n,2}(f)$ satisfy the conditions of Theorem 9.4. In this case formula (9.9) can be written in the following explicit form:*

$$J_{n,2}(f)=\frac{1}{n}I_{n,2}(f_{\{1,2\}})-\frac{1}{n}I_{n,1}(f_{\{1\}})-\frac{1}{n}I_{n,1}(f_{\{2\}})-f_\emptyset \tag{9.9$'$}$$

with the functions

$$f_{\{1,2\}}(x,y) = f(x,y) - \int f(x,y)\mu(\,dx) - \int f(x,y)\mu(\,dy) + \int f(x,y)\mu(\,dx)\mu(\,dy),$$
$$f_{\{1\}}(x) = \int f(x,y)\mu(\,dy) - \int f(x,y)\mu(\,dx)\mu(\,dy),$$
$$f_{\{2\}}(y) = \int f(x,y)\mu(\,dx) - \int f(x,y)\mu(\,dx)\mu(\,dy)$$

and $f_\emptyset = \int f(x,y)\mu(\,dx)\mu(\,dy)$.

10. The proof of Theorem 8.3 about the distribution of U-statistics

This section contains the proof of Theorem 8.3 about the distribution of degenerate U-statistics with the help of some results which are interesting in themselves. One of these results, called Borell's inequality, gives an estimate on the moments of homogeneous polynomials of Rademacher functions, another result we need is a symmetrization type estimate which can be considered as the multivariate version of the more interesting part of the Marcinkiewicz–Zygmund inequality. Finally there is a third result we apply which compares the distribution of a U-statistics with the distribution of an appropriate modification of it. The first two results will be proved in the next section, the third one in the Appendix.

Theorem 8.3 can be considered as the generalization of Bernstein's inequality (Theorem 3.1) for U-statistics. Bernstein's inequality was proved by means of an estimation of the moment-generating function of the partial sums of independent and bounded random variables. This approach has to be modified in the proof of Theorem 8.3. In such cases we cannot work well with the moment generating functions, since if the sample size tends to infinity, then the normalized version of degenerate U-statistics of order k have a limit distribution F with a tail-behaviour $1-F(x)\ge e^{-Cx^{2/k}}$ with some $C>0$ as $x\to\infty$. (This is a relatively well-known result, but we shall not need it in this work.) This means that a random variable with this limit distribution has no moment generating function for $k\ge3$. On the other hand, the proof of Theorem 8.3 is relatively simple, if we have a good estimate also for the high moments of degenerate U-statistics. Such a moment estimate is formulated in the following

Proposition 10.1. *Let us consider a canonical function $f=f(x_1,\dots,x_k)$ on the k-fold product $(X^k,\mathcal{X}^k,\mu^k)$ of a measure space $(X,\mathcal{X},\mu)$ together with a sequence of independent μ distributed random variables and the degenerate U-statistic $I_{n,k}(f)$ determined by this sequence of random variables $\xi_1,\dots,\xi_n$ and canonical function f. Let us also assume that the function f satisfies conditions (8.1) and (8.2) with some number $0<\sigma\le1$.*

Then there exists some constants $C=C_k>0$ such that the moments of the U-statistic $I_{n,k}(f)$ defined in formula (8.7) satisfy the inequality

$$E\left(\left|n^{-k/2}I_{n,k}(f)\right|^{2M}\right)\le C_k^M M^{kM}\sigma^{2M} \quad \text{if } 1\le M\le n\sigma^2. \tag{10.1}$$

Let us consider the k-th power of a standard normal random variable η and calculate the asymptotic magnitude of the $2M$-th moment $E\left(\sigma\eta^k\right)^{2M}$ of $\sigma\eta^k$ for large M. We have $E\left(\sigma\eta^k\right)^{2M}=1\cdot3\cdots(2kM-1)\sigma^{2M}=\frac{(2kM)!}{2^{kM}(kM)!}\sigma^{2M}\sim\left(\frac{2k}{e}\right)^{kM}M^{kM}\sigma^{2M}$ by the Stirling formula. This means that the estimate given for the $2M$-th moments of a normalized U-statistics $n^{-k/2}I_{n,k}(f)$ in formula (10.1) has the same order as the $2M$-th moment of the random variable const. $\sigma\eta^k$, at least if $1\le M\le n\sigma^2$. This estimate will imply Theorem 8.3 which also can be so interpreted that $P\left(n^{-k/2}I_{n,k}(f)>u\right)$ can be bounded by const. $P\left(\text{const.}\,\sigma\eta^k>u\right)$, at least if $0<u\le n^{k/2}\sigma^{k+1}$.

The hard part of the problem is to prove Proposition 10.1. There are methods to bound the moments of multiple Wiener–Itô integrals, and it is natural to try to adapt them to the proof of Proposition 10.1. I know of two different methods for estimating the moments of Wiener–Itô integrals. One of them is the so-called diagram formula which expresses the product of Wiener–Itô integrals as sums of appropriate new Wiener–Itô integrals, the other one is called Nelson's inequality which yields a direct comparison between the L_p-norms of Wiener–Itô integrals for different parameters p. Both of them can be adapted to our case, but they demand the solution of several non-trivial technical problems. The adaptation of Nelson's inequality seems to be the less complicated method, and this approach will be followed in this work. There is an important estimate, called Borell's inequality which will be applied. This inequality makes a comparison between the L_p norms of homogeneous polynomials of independent Rademacher functions for different parameters p. Borell's inequality in itself will be not sufficient for us, because we want to estimate more complicated objects. But we shall formulate and prove some additional results, and they will enable us together with Borell's inequality to prove a version of Proposition 10.1 which will be sufficient for our purposes.

Borell's inequality will be formulated below, but its proof is postponed to the next section.

Theorem 10.2 (Borell's inequality). *Let $\varepsilon_1,\dots,\varepsilon_n$ be independent, identically distributed random variables $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac12$, $1\le l\le n$, fix some real numbers $a(l_1,\dots,l_k)$ for all indices $(l_1,\dots,l_k)$ such that $1\le l_j\le n$, $1\le j\le k$, and $l_j\ne l_{j'}$ if $j\ne j'$, and define the random variable*

$$Z=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\,1\le j\le k\\ l_j\ne l_{j'}\text{ if }j\ne j'}}a(l_1,\dots,l_k)\varepsilon_{l_1}\cdots\varepsilon_{l_k}. \tag{10.2}$$

The inequality

$$E|Z|^p\le\left(\frac{p-1}{q-1}\right)^{kp/2}(E|Z|^q)^{p/q}\qquad\text{if}\quad 1<q\le p<\infty \tag{10.3}$$

holds.

Remark: The most interesting special case of Borell's inequality is when $q=2$, and we shall consider only this case. Since $EZ^2 \le \frac{1}{k!} \sum\limits_{\substack{1\le l_j\le n,\, 1\le j\le k\\ l_j\ne l_{j'} \text{ if } j\ne j'}} a^2(l_1,\dots,l_k)$, it yields that

$$E|Z|^p \le (p-1)^{kp/2} \left(\frac{1}{k!} \sum_{\substack{1\le l_j\le n,\, 1\le j\le k\\ l_j\ne l_{j'} \text{ if } j\ne j'}} a^2(l_1,\dots,l_k) \right)^{p/2} \qquad \text{if } 2\le p<\infty \tag{10.4}$$

We have the estimate written for EZ_n^2 because

$$E\varepsilon_{l_1}\cdots\varepsilon_{l_k} a(l_1,\dots,l_k)\varepsilon_{l'_1}\cdots\varepsilon_{l'_k} a(l'_1,\dots,l'_k)=0$$

if the sets of arguments $\{l_1,\dots,l_k\}$ and $\{l'_1,\dots,l'_k\}$ do not agree. In the inequality written for EZ_n^2 we have identity if all coefficients $a(l_1,\dots,l_k)$ are symmetric functions of their arguments, otherwise we can only write inequality.

Borell's inequality does not give a direct estimate for the moments $EI_{n,k}(f)^{2M}$ of the U-statistics we are interested in. But together with a symmetrization result formulated below it enables us to prove such a recursive estimate between the $2M$-th and $4M$-th moments of degenerate U-statistics which implies a version of Proposition 10.1 appropriate for our goals. This additional symmetrization result we need can be considered as a multivariate version of the Marcinkiewicz–Zygmund inequality about independent random variables with zero mean. First this symmetrization result will be given. Then for the sake of a better understanding the Marcinkiewicz–Zygmund inequality will be recalled, and its relation to the result considered as its multivariate version will be explained.

To formulate a good multivariate version of the Marcinkiewicz–Zygmund inequality first we introduce a notion which is called decoupled U-statistics in the literature.

The definition of decoupled and randomized decoupled U-statistics. *Let us have k independent copies $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, of a sequence $\xi_1,\dots,\xi_n$ of independent and identically distributed random variables taking their values on a measurable space $(X,\mathcal{X})$ together with a measurable function $f(x_1,\dots,x_k)$ on the product space $(X^k,\mathcal{X}^k)$ with values in a separable Banach space. Then the decoupled U-statistic determined by the random sequences $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, and kernel function f is defined by the formula*

$$\bar I_{n,k}(f)=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'} \text{ if } j\ne j'}} f\left(\xi_{l_1}^{(1)},\dots,\xi_{l_k}^{(k)}\right). \tag{10.5}$$

Let us have, besides the sequences $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, and function $f(x_1,\dots,x_k)$ a sequence of independent random variables $\varepsilon=(\varepsilon_1,\dots,\varepsilon_n)$, $P(\varepsilon_l=1)=P(\varepsilon_l=$

$-1)=\frac{1}{2}$, $1\le l\le n$, *which is independent also of the sequences of random variables* $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$. *We define the randomized decoupled U-statistic determined by the random sequences* $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, *the kernel function* f *and the randomizing sequence* $\varepsilon_1,\dots,\varepsilon_n$ *by the formula*

$$\bar I_{n,k}(f,\varepsilon)=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'} \text{ if } j\ne j'}}\varepsilon_{l_1}\cdots\varepsilon_{l_k}f\left(\xi_{l_1}^{(1)},\dots,\xi_{l_k}^{(k)}\right). \tag{10.6}$$

Now a symmetrization result will be formulated which will be applied in the proof of an appropriate version of Proposition 10.1. This result will be proved in the next section.

Proposition 10.3. *Let* $\xi_1,\dots,\xi_n$ *be a sequence of i.i.d. random variables which take their values on a measurable space* $(X,\mathcal{X})$ *with some distribution* μ, *and let* $f(x_1,\dots,x_k)$ *be a canonical function with respect to this measure* μ *such that* $E|f(\xi_1,\dots,\xi_k)|^p<\infty$ *with some* $p\ge1$. *Let us have* k *independent copies* $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, *of the sequence* $\xi_1,\dots,\xi_n$ *with the same distribution, and let* $\varepsilon=(\varepsilon_1,\dots,\varepsilon_n)$ *be a sequence of independent random variables,* $P(\varepsilon_l=1)=P(\varepsilon_l=-1)$, $1\le l\le n$ *which is also independent of the sequences* $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$. *The inequality*

$$E|\bar I_{n,k}(f)|^p\le 2^{kp}E|\bar I_{n,k}(f,\varepsilon)|^p \tag{10.7}$$

holds for the decoupled U-statistic $\bar I_{n,k}(f)$ *and its randomized version* $\bar I_{n,k}(f,\varepsilon)$ *defined in formulas (10.5) and (10.6) by means of the random sequences* $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, $\varepsilon_1,\dots,\varepsilon_n$ *and the kernel function* f.

In Proposition 10.1 we want to bound the moments of a U-statistic, while in Proposition 10.3 we have an estimate about decoupled U-statistics $\bar I_{n,k}(f)$. This results deals with decoupled statistics, because as we shall see, its proof does not work for the original U-statistics. This causes some difficulties, but they can be overcome with the help of a result of de la Peña and Montgomery–Smith. It will be formulated more generally than it is needed in the solution of the present problem to make it applicable also in the investigations of the subsequent part of the work. For its more general formulation let us slightly generalize the notion of U-statistics, let us allow also the case when the kernel function f in formula (8.7) takes its value in a separable Banach space. The result will be formulated in Theorem 10.4, and it will be proved in the Appendix.

Theorem 10.4. (de la Peña and Montgomery–Smith) *Let us consider a sequence of independent and identically distributed random variables* $\xi_1,\dots,\xi_n$ *on a measurable space* $(X,\mathcal{X})$ *together with* k *independent copies* $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$. *Let us also have a function* $f(x_1,\dots,x_k)$ *on the* k*-fold product space* $(X^k,\mathcal{X}^k)$ *which takes its values on a separable Banach space* B. *Define the U-statistic and decoupled U-statistic* $I_{n,k}(f)$ *and* $\bar I_{n,k}(f)$ *with the help of the above random sequences* $\xi_1,\dots,\xi_n$, $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le$

$j \le k$, and kernel function f. Then there exist some constants $\bar C = \bar C(k) > 0$ and $\gamma = \gamma(k) > 0$ depending only on the order k of the U-statistic such that

$$P\left(\|I_{n,k}(f)\| > u\right) \le \bar C P\left(\|\bar I_{n,k}(f)\| > \gamma u\right) \tag{10.8}$$

for all $u > 0$. Here $\|\cdot\|$ denotes the norm in the Banach space B where the function f takes its values.

More generally, if we have a countable sequence of functions f_s, $s = 1, 2, \dots$, taking their values in the same separable Banach-space, then

$$P\left(\sup_{1\le s<\infty} \|I_{n,k}(f_s)\| > u\right) \le \bar C P\left(\sup_{1\le s<\infty} \left\|\bar I_{n,k}(f_s)\right\| > \gamma u\right). \tag{10.8$'$}$$

We follow the following approach. We shall prove such a version of Proposition 10.1 and Theorem 8.3 where U-statistics are replaced by decoupled U-statistics. The proof of these results is simpler, because the arguments applied for U-statistics also work for decoupled U-statistics, and also Proposition 10.3 can be applied in this case. Theorem 8.3 can be obtained as a consequence of its version we shall prove and Theorem 10.4. More explicitly, we shall prove the following two results.

Proposition 10.1$'$**.** *Let the conditions of Proposition 10.1 be satisfied with some sequence of iid. μ-distributed random variables $\xi_1, \dots, \xi_n$ on a space $(X, \mathcal{X})$, a function f on the product space $(X^k, \mathcal{X}^k)$ and a number $0 < \sigma \le 1$. Take k independent copies $\xi_1^{(j)}, \dots, \xi_n^{(j)}$, $1 \le j \le k$, of the random sequence $\xi_1, \dots, \xi_n$, and define with their help the decoupled U-statistic $I_{n,k}(f)$ defined in (10.5). Then the inequality*

$$E\left(\left|n^{-k/2}\bar I_{n,k}(f)\right|^{2M}\right) \le C_k^M M^{kM}\sigma^{2M} \quad \text{if } 1 \le M \le n\sigma^2 \tag{10.1$'$}$$

holds with some constant C_k which depends only on the order k of the decoupled U-statistic.

Theorem 8.3$'$**.** *Let the conditions of Proposition 8.3 be satisfied with some sequence of iid. μ-distributed random variables $\xi_1, \dots, \xi_n$ on a space $(X, \mathcal{X})$, a function f on the product space $(X^k, \mathcal{X}^k)$ and a number $0 < \sigma \le 1$. Take k independent copies $\xi_1^{(j)}, \dots, \xi_n^{(j)}$, $1 \le j \le k$, of the random sequence $\xi_1, \dots, \xi_n$, and define with their help the decoupled U-statistic $\bar I_{n,k}(f)$ defined in (10.5). Then there exist some constants $C = C(k) > 0$ and $\alpha = \alpha(k) > 0$ such that the inequality*

$$P\left(n^{-k/2}|\bar I_{n,k}(f)| > u\right) \le C \exp\left\{-\alpha\left(\frac{u}{\sigma}\right)^{2/k}\right\} \tag{10.9}$$

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

It is clear that Theorem 8.3$'$ together with Theorem 10.4 imply Theorem 8.3. Let us continue our discussion with an explanation of the content of Proposition 10.3. As we have mentioned, it can be considered as a multivariate version of the Marcinkiewicz–Zygmund inequality which can be formulated in the following way:

Let $\xi_1,\dots,\xi_n$ be independent random variables such that $E\xi_j=0$, $1\le j\le n$. Then for all $p\ge2$ there exist some constants $0<B_p<C_p<\infty$ such that

$$B_pE\left(\sum_{l=1}^n \xi_l^2\right)^{p/2}\le E\left|\sum_{j=l}^n \xi_l\right|^p\le C_pE\left(\sum_{l=1}^n \xi_l^2\right)^{p/2}. \tag{10.12}$$

(This inequality also has a generalization for sums of martingale differences.) The really interesting part of formula (10.12) is his right-hand side part. It is useful, because the expression at the right-hand side of (10.12) can be well estimated even without exploiting the independence of the summands. The right-hand side of (10.12) can be deduced from Borell's inequality, more explicitly from its consequence (10.4) with $k=1$ and the inequality

$$E\left|\sum_{l=1}^n \xi_l\right|^p\le \bar C_pE\left|\sum_{l=1}^n \varepsilon_j\xi_l\right|^p. \tag{10.12$'$}$$

with some $\bar C_p>0$, where $\varepsilon_1,\dots,\varepsilon_n$, $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac12$ are independent random variables, independent also of the random sequence $\xi_1,\dots,\xi_n$. Indeed, formula (10.4) implies that $E\left|\sum\limits_{l=1}^n\varepsilon_l\xi_l\right|^p\le(p-1)^{p/2}E\left(\sum\limits_{l=1}^n\xi_l^2\right)^{p/2}$. Let us also observe that Proposition 10.3 is a multivariate generalization of formula (10.12$'$) with the additional (important) information that it gives a good explicit choice for the coefficient $\bar C_p$ in it.

We can prove with the help of Borell's inequality such an inequality which has similar relation to Proposition 10.3 as the right-hand side inequality in formula (10.12) to formula (10.12$'$). We shall give this result in the following corollary, and actually we shall apply this consequence of Proposition 10.3.

Corollary of Proposition 10.3. *Let the conditions of Proposition 10.3 hold with the additional restriction that the inequality $E|f(\xi_1,\dots,\xi_k)|^p<\infty$ holds with some $p\ge2$ (i.e. $p>1$ is not sufficient for us). Then also the inequality*

$$E|\bar I_{n,k}(f)|^p\le 2^{kp}p^{kp/2}E\bar I_{n,k}(f^2)^{p/2} \tag{10.13}$$

holds.

Proof of the Corollary of Proposition 10.3. Let $\mathcal F$ denote the σ-algebra generated by the random variables $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$. Then Proposition 10.3 implies that

$$E|\bar I_{n,k}(f)|^p\le 2^{kp}E|\bar I_{n,k}(f,\varepsilon)|^p=2^{kp}E(E(|\bar I_{n,k}(f,\varepsilon)|^p|\mathcal F)).$$

On the other hand, the consequence of Borell's inequality formulated in relation (10.4) yields that

$$E(|\bar I_{n,k}(f,\varepsilon)|^p|\mathcal F)=E_\varepsilon\left|\frac1{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\neq l_{j'} \text{ if } j\neq j'}}\varepsilon_{l_1}\cdots\varepsilon_{l_k}f\left(\xi^{(1)}_{l_1},\dots,\xi^{(k)}_{l_k}\right)\right|^p$$

$$\le p^{kp/2}\left(\frac1{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\neq l_{j'} \text{ if } j\neq j'}}f^2\left(\xi^{(1)}_{l_1},\dots,\xi^{(k)}_{l_k}\right)\right)^{p/2}=p^{kp/2}\bar I_{n,k}(f^2)^{p/2},$$

where E_ε means that we fix the values of the random variables $\xi^{(j)}_1,\dots,\xi^{(j)}_n$, $1\le j\le k$ and take expectation with respect to the random variables ε_j, $1\le j\le n$. We get, by taking expectation in the last inequality, that $E|\bar I_{n,k}(f,\varepsilon)|^p\le p^{kp/2}E\bar I_{n,k}(f^2)^{p/2}$. This inequality together with formula (10.7) imply relation (10.13).

Now we turn to the proof of Proposition 10.1′.

The proof of Proposition 10.1′. We have $En^{-k}\bar I_{n,k}(f)^2\le\frac1{k!^2}\sigma^2$ if f is a canonical function with respect to the probability measure μ, and $\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\le\sigma^2$, i.e. relation (10.1′) in Proposition 10.1′ holds for $M=1$ if $C_k\ge\frac1{k!^2}$, because

$$Ef(\xi^{(1)}_{l_1},\dots,\xi^{(k)}_{l_k})f(\xi^{(1)}_{l'_1},\dots,\xi^{(k)}_{l'_k})=0,\quad \text{if } l_j\neq l'_j \quad \text{for some index } 1\le j\le k,$$

and $Ef^2(\xi^{(1)}_{l_1},\dots,\xi^{(k)}_{l_k})\le\sigma^2$.

First we prove relation (10.1′) in the special case $M=2^m$ with $m=0,1,\dots$ if $1\le M\le 2n\sigma^2$ and the constants C_k are chosen appropriately in (10.1′). We have already proved this relation for $m=0$. We shall prove the inequality $E(n^{-k/2}I_{n,k}(f)^{2M})\le C_k^M M^{kM}\sigma^{2M}$ for all $k=1,2,\dots$ with some appropriate constant $C_k>0$ if $M=2^m$ and $M\le 2n\sigma^2$ by induction with respect to m. In the proof formula (10.13) of the Corollary of Proposition 10.3 will be applied with the choice $p=2M$. This yields the estimate

$$E\left(\left(n^{-k/2}\bar I_{n,k}(f)\right)^{2M}\right)\le 2^{2kM}(2M)^{Mk}E\left(n^{-k}\bar I_{n,k}(f^2)\right)^M. \tag{10.14}$$

The above inequality is not sufficient in its original form to carry out the inductive procedure we have in mind, since the function f^2 appearing at its right-hand side is not canonical. But this difficulty can be overcome if we apply the Hoeffding decomposition (9.2) for the function f^2.

This result yields a representation of the form

$$f^2(x_1,\dots,x_k)=\sum_{V\subset\{1,\dots,k\}}f_V(x_s,s\in V)$$

with some appropriate canonical functions $f_V(x_s, s\in V)$ with respect to the measure μ for all $V\subset\{1,\dots,k\}$. This relation implies that similarly to U-statistics decoupled U-statistics satisfy the relation

$$\bar I_{n,k}(f^2)=\sum_{V\subset\{1,\dots,k\}}(n-|V|)(n-|V|-1)\cdots(n-k+1)\frac{|V|!}{k!}\bar I_{n,|V|}(f_V). \tag{10.15}$$

In Theorem 9.1 the functions f_V appearing in formula (10.15) are described explicitly. (Here again we define the value of the product $(n-|V|)(n-|V|-1)\cdots(n-k+1)$ as 1 for $|V|=k$.) We do not need this formula, we only need that by formulas (9.4) and (9.4′) of Theorem 9.2 the integrals of the square of the functions f_V are bounded by σ^2, and these functions are bounded by $2^{|V|}$ in supremum norm, because the function f^2, similarly to the function f, is bounded by σ in $L_2(\mu)$-norm, and it is bounded by 1 in the supremum norm. The coefficient f_V with $V=\emptyset$ in the constant term of the sum at the right-hand side of (10.15) has to be considered separately. It equals $f_\emptyset=\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)$. This implies that $0\le f_\emptyset\le\sigma^2$, an estimate which does not follow directly from Theorem 9.2.

Formula (10.15) and the triangular inequality in L_M norm imply the inequality

$$\begin{aligned}E(n^{-k}\bar I_{n,k}(f^2))^M&\le\left(n^{-k}\sum_{V\subset\{1,\dots,k\}}\left(E\left(n^{k-|V|}\frac{(|V|)!}{k!}I_{n,|V|}(f_V)\right)^M\right)^{1/M}\right)^M\\&\le\left(\sum_{j=0}^k\binom{k}{j}\frac{j!}{k!}\sup_{V\colon V\subset\{1,\dots,k\},\,|V|=j}\left(n^{-jM/2}E(n^{-j/2}\bar I_{n,j}(f_V))^M\right)^{1/M}\right)^M.\end{aligned} \tag{10.16}$$

The function $2^{-j}|f_V|$ is bounded by 1 in the supremum norm and by $2^{-j/2}\sigma\le\sigma$ in the $L_2(\mu)$ norm if $|V|=j$, $1\le j\le k$. Our inductive hypothesis implies that the terms at the right-hand side of (10.16) can be estimated as

$$\begin{aligned}n^{-jM/2}E(n^{-j/2}\bar I_{n,j}(f_V))^M&\le 2^{jM}C_j^{M/2}\sigma^M\left(\frac M2\right)^{jM/2}n^{-jM/2}\\&=(2^jC_j)^{M/2}\sigma^{2M}\left(\frac{M}{n\sigma^{2/j}}\right)^{jM/2}\le 2^{jM}C_j^{M/2}\sigma^{2M},\\&\qquad\text{if } |V|=j,\ 1\le j\le k,\ \text{and } M\le 2n\sigma^2,\end{aligned}$$

since $\frac{M}{n\sigma^{2/j}}\le\frac{M}{n\sigma^2}\le 2$ in this case. (Observe that $\sigma^2\le1$, since $\sup|f(x_1,\dots,x_k)|\le1$.) Besides, $\binom kj\frac{j!}{k!}\sup\limits_{V\colon V\subset\{1,\dots,k\},\,|V|=j}\left(E(n^{-j}\bar I_{n,j}(f_V))^M\right)^{1/M}=\frac{f_\emptyset}{k!}\le\frac{\sigma^2}{k!}$ in the case $j=0$. These estimates yield that

$$E(n^{-k}I_{n,k}(f^2))^M\le\sigma^{2M}\left(\sum_{j=0}^k\frac{2^j}{(k-j)!}C_j^{1/2}\right)^M$$

if $M\le 2n\sigma^2$, and we choose $C_0\ge1$. By formula (10.14) and this estimate

$$E(n^{-k/2}\bar I_{n,k}(f))^{2M}\le 2^{3kM}M^{kM}E\left(n^{-k}\bar I_{n,k}(f^2)\right)^M \le \left(\sum_{j=0}^k \frac{2^{3k+j}}{(k-j)!}C_j^{1/2}\right)^M M^{kM}\sigma^{2M} \tag{10.17}$$

if $M\le 2n\sigma^2$. I show that with an appropriate choice of the coefficients C_k (which may depend only on k but not on M) the above estimate implies the inductive step. Indeed, we can choose such a sequence C_k, $=0,1,2,\dots$, with $C_0=1$ which satisfies the inequalities $C_k\ge\frac1{k!}$ and

$$\sum_{j=0}^k \frac{2^{3k+j}}{(k-j)!}C_j^{1/2}\le C_k \qquad \text{for all } k=1,2,\dots \tag{10.18}$$

Let us choose such a sequence C_k, $k=0,1,\dots$, which satisfies these relations. Then formula (10.1′) holds for $M=1$, and our inductive procedure together with relations (10.17) and (10.18) imply that it also holds for $M\le 2n\sigma^2$, i.e.

$$E(n^{-k/2}I_{n,k}(f))^{2M}\le C_k^M\sigma^{2M}M^{kM} \quad \text{if } M=2^m \text{ and } M\le 2n\sigma^2.$$

Thus we have proved Proposition 10.1′ in the special case when $M=2^m$, $m=0,1,\dots$, and $M\le 2n\sigma^2$. To estimate the moment $E|n^{-k/2}\bar I_{n,k}(f)|^{2M}$ for a general exponent $1\le M\le n\sigma^2$ (the number M may be non-integer) let us consider the number $\bar M=\bar M(M)$ of the form $\bar M=2^m$ with some integer m which satisfies the relation $\bar M\le M<2\bar M$. By applying the already proved part of Proposition 10.1′ for $\bar M$ we can write

$$E|n^{-k/2}\bar I_{n,k}(f)|^{2M}\le\left(E|n^{-k/2}\bar I_{n,k}(f)|^{2\bar M}\right)^{M/\bar M}\le\left(C_k^{\bar M}\sigma^{2\bar M}\bar M^{k\bar M}\right)^{M/\bar M} \le C_k^M\sigma^{2M}(2M)^{kM}=(2^kC_k)^M\sigma^{2M}M^{kM}.$$

Proposition 10.1′ is proved.

The proof of Theorem 8.3′. By the Markov inequality and Proposition 10.1′

$$P\left(|n^{-k/2}\bar I_{n,k}(f)|>u\right)\le\frac{E|n^{-k/2}\bar I_{n,k}(f)|^{2M}}{u^{2M}}\le\left(\frac{C_k\sigma^2M^k}{u^2}\right)^M$$

if $u>0$ and $1\le M\le n\sigma^2$. Let us choose $M=\frac1e\left(\frac{u^2}{C_k\sigma^2}\right)^{1/k}$. With this choice of the parameter M we get that

$$P\left(|n^{-k/2}I_{n,k}(f)|>u\right)\le e^{-kM}=\exp\left\{-\frac ke C_k^{-1/k}\left(\frac u\sigma\right)^{2/k}\right\} \tag{10.19}$$

if $\sqrt{C_k}e^{k/2}\sigma \le u \le e^{k/2}\sqrt{C_k}n^{k/2}\sigma^{k+1}$. Relation (10.19) implies formula (10.9). Indeed, formula (10.9) remains valid for $\sqrt{C_k}e^{k/2}n^{k/2}\sigma^{k+1} \le u \le n^{k/2}\sigma^{k+1}$ if the constant $kC_k^{-1/k}e^{-1}$ in the exponent at the right-hand side is replaced by $\alpha = \min(kC_k^{-1/k}e^{-1}, k)$, (here we exploit that $P\left(|n^{-k/2}I_{n,k}(f)| > u\right) \le P\left(|n^{-k/2}I_{n,k}(f)| > \sqrt{C_k}e^{k/2}n^{k/2}\sigma^{k+1}\right)$ if $u \ge \sqrt{C_k}e^{k/2}n^{k/2}\sigma^{k+1}$), and it holds also for $0 \le u \le \sqrt{C_k}e^{k/2}\sigma$ if the right-hand side is multiplied with a sufficiently large constant C.

As we have mentioned, Theorems 8.3′ and Theorem 10.4 together imply Theorem 8.3.

11. Some useful basic results

This section contains the proof of Borell's inequality and Proposition 10.3 which can be considered as the multivariate version of the Marcinkiewicz–Zygmund inequality, more precisely of its more important part.

11 A.) THE PROOF OF BORELL'S INEQUALITY FORMULATED IN THEOREM 10.2.

Borell's inequality will be proved as the consequence of the following hypercontractive inequality for Rademacher functions.

Theorem 11.1. The hypercontractive inequality for Rademacher functions. *Let us consider two copies $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu) = (X, \mathcal{X}, \mu)$ of the measure space $(X, \mathcal{X}, \mu)$, where $X = \{-1, 1\}$, $\mathcal{X}$ contains all subsets of X, and $\mu(\{1\}) = \mu(\{-1\}) = \frac{1}{2}$. Given a real number $\gamma > 0$ let us introduce the linear operator $\mathbf{T}_\gamma$ which maps the real (or complex) valued functions on the space X to the real (or complex) valued functions on the space Y which is defined by the relations $\mathbf{T}_\gamma r_0 = r_0$, and $\mathbf{T}_\gamma r_1 = \gamma r_1$, where $r_0(1) = r_0(-1) = 1$, and $r_1(1) = 1$, $r_1(-1) = -1$. For all $n = 1, 2, \dots$ let us consider the n-fold product $(X^n, \mathcal{X}^n, \mu^n)$ and $(Y^n, \mathcal{Y}^n, \nu^n)$ of the spaces $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ together with the n-fold product of the operator $\mathbf{T}_\gamma^n$ of the operator $\mathbf{T}_\gamma$ acting between these product spaces, (i.e. $\mathbf{T}_\gamma^n$ is the linear transformation for which $\mathbf{T}_\gamma^n(f_1(x_1)\cdots f_n(x_n)) = \mathbf{T}_\gamma f_1(x_1)\cdots \mathbf{T}_\gamma f_n(x_n)$ for all products of the functions f_s, $1 \le s \le n$, on the space $(X, \mathcal{X}, \mu)$). The transformation $\mathbf{T}_\gamma^n$ from the space $L_q(X^n, \mathcal{X}^n, \mu^n)$ to the space $L_p(Y^n, \mathcal{Y}^n, \nu^n)$ has norm 1 for all $n = 1, 2, \dots$ if $1 < p \le q < \infty$, and $0 \le \gamma \le \sqrt{\frac{q-1}{p-1}}$.*

The name hypercontractive inequality was given to this result because it states not only that $\|\mathbf{T}_\gamma^n f\|_q \le \|f\|_q$ for all functions f but also the inequality $\|\mathbf{T}_\gamma^n f\|_p \le \|f\|_q$ with some $1 \le q < p$, while $\|\mathbf{T}_\gamma^n f\|_q \le \|\mathbf{T}_\gamma^n f\|_p$ if $1 \le q < p$. It is not difficult to see that the hypercontractive inequality implies Borell's inequality.

The proof of Borell's inequality by means of the hypercontractive inequality. Let us define the function

$$f(x_1, \dots, x_n) = \sum_{\substack{1 \le l_j \le n,\ 1 \le j \le k \\ j_s \ne l_{j'} \text{ if } j \ne j'}} a(l_1, \dots, l_k) r_1(x_{l_1}) \cdots r_1(x_{l_k})$$

on the space $(X^n,\mathcal{X}^n,\mu^n)$. Observe that $\mathbf{T}^n_\gamma f=\gamma^k f$ for this function f and all $\gamma>0$, and $E|Z|^p=\|f\|_p^p$, $E|Z|^q=\|f\|_q^q$. Fix some numbers $1<q\le p\le\infty$ and put $\gamma=\sqrt{\frac{q-1}{p-1}}$. The norm of $\mathbf{T}^n_\gamma$ as a transformation from the space $L_q(X^n,\mathcal{X}^n,\mu^n)$ to the space $L_p(Y^n,\mathcal{Y}^n,\nu^n)$ is bounded by 1, i.e. $\|\mathbf{T}^n_\gamma f\|_p=\gamma^k\|f\|_p\le\|f\|_q$. The above relations imply that $(E|Z|^p)^{1/p}\le\left(\frac{q-1}{p-1}\right)^{k/2}E|Z|^q)^{1/q}$ in this case, and this is what we had to show.

The proof of the hypercontractive inequality can be reduced to a simpler statement by means of the following

Theorem 11.2. *Let us consider two pairs of measure spaces $(X_1,\mathcal{A}_1,\mu_1)$, $(Y_1,\mathcal{B}_1,\nu_1)$ and $(X_2,\mathcal{A}_2,\mu_2)$, $(Y_2,\mathcal{B}_2,\nu_2)$ together with two linear operators $\mathbf{T}_1$ and $\mathbf{T}_2$ which map the space $L_q(X_1,\mathcal{A}_1,\mu_1)$ to $L_p(Y_1,\mathcal{B}_1,\nu_1)$ and the space $L_q(X_2,\mathcal{A}_2,\mu_2)$ to $L_p(Y_2,\mathcal{B}_2,\nu_2)$ respectively. Assume that $1\le q\le p$, and the norm of both operators $\mathbf{T}_1$ and $\mathbf{T}_2$ is less than or equal to 1. Then also the norm of their direct product $\mathbf{T}_1\times\mathbf{T}_2$ which maps the space $L_q(X_1\times X_2,\mathcal{A}_1\times\mathcal{A}_2,\mu_1\times\mu_2)$ to the space $L_p(Y_1\times Y_2,\mathcal{B}_1\times\mathcal{B}_2,\nu_1\times\nu_2)$ is less than or equal to one.*

Proof of Theorem 11.2: We have to show that

$$\begin{aligned}
&\int_{Y_1\times Y_2}\left|\sum_{j=1}^n c_j\mathbf{T}_1f_j(y_1)\mathbf{T}_2g_j(y_2)\right|^p\nu_1(\,dy_1)\nu_2(\,dy_2)\\
&\qquad\le\left[\int_{X_1\times X_2}\left|\sum_{j=1}^n c_jf_j(x_1)g_j(x_2)\right|^q\mu_1(\,dx_1)\mu_2(\,dx_2)\right]^{p/q}
\end{aligned}\tag{11.1}$$

for arbitrary index n, real (or complex) numbers c_j and functions $f_j(\cdot)\in L_q(X_1,\mathcal{A}_1,\mu_1)$ and $g_j(\cdot)\in L_q(X_2,\mathcal{A}_2,\mu_2)$, $1\le j\le n$, since relation (11.1) is equivalent to the inequality $\|(\mathbf{T}_1\times\mathbf{T}_2)f(y_1,y_2\|_{L_p}\le\|f(x_1,x_2)\|_{L_q}$ for the function $f(x_1,x_2)=\sum\limits_{j=1}^n c_jf_j(x_1)g_j(x_2)$, and as functions of the above form are dense in the space $L_q(X_1\times X_2,\mathcal{A}_1\times\mathcal{A}_2,\mu_1\times\mu_2)$, this inequality implies that the norm of $\mathbf{T}_1\times\mathbf{T}_2$ is bounded by 1.

We get by integrating the left-hand side of (11.1) first by the variable y_1 and by exploiting the condition $|\mathbf{T}_1|\le1$ that

$$\begin{aligned}
&\int_{Y_1\times Y_2}\left|\sum_{j=1}^n c_j\mathbf{T}_1f_j(y_1)\mathbf{T}_2g_j(y_2)\right|^p\nu_1(\,dy_1)\nu_2(\,dy_2)\\
&\qquad\le\int_{Y_2}\left[\int_{X_1}\left|\sum_{j=1}^n c_jf_j(x_1)\mathbf{T}_2g_j(y_2)\right|^q\mu_1(\,dx_1)\right]^{p/q}\nu_2(\,dy_2).
\end{aligned}\tag{11.2}$$

We shall prove and apply the following result. Let a function $G(u,v)$ be given on a product space $(U\times V,\mathcal{U}\times\mathcal{V},\rho_1\times\rho_2)$, and let $1\le s<\infty$. Then

$$\left[\int_V\left[\int_U |G(u,v)|\rho_1(\,du)\right]^s\rho_2(\,dv)\right]^{1/s}\le\int_U\left[\int_V|G(u,v)|^s\rho_2(\,dv)\right]^{1/s}\rho_1(\,du). \qquad (11.3)$$

It is enough to prove this estimate for the following special type of functions $G(u,v)$. Let us consider a finite partition $A_1,\dots,A_m$ of the space U, choose for all $1\le j\le m$ a function $G_j(v)$ on the space $(V,\mathcal{V})$ and put $G(u,v)=G_j(v)$ if $u\in A_j$, $1\le j\le m$. Such kind of functions are dense in the $L_q(U\times V,\mathcal{U}\times\mathcal{V},\rho_1\times\rho_2)$ space, because such kind of functions are dense in the subspace consisting of functions of the form $G(u,v)=\sum\limits_{j=1}^n c_jf_j(u)g_j(v)$. If we prove inequality (11.3) for such special type of functions, then this inequality can be generalized for general functions $G(u,v)$ by an appropriate limiting procedure. Its details are left to the reader.

Inequality (11.3) in the special case we consider is equivalent to the triangular inequality in L_s spaces, $s\ge1$, (also called Minkowski inequality)

$$\left\|\sum_{j=1}^m\rho_1(A_j)|G_j(v)|\right\|_s\le\sum_{j=1}^m\|\rho_1(A_j)|G_j(v)|\|_s,$$

where $\|f\|_s$ denotes the L_s-norm of a function f in the space $(V,\mathcal{V},\rho_2)$.

Indeed,

$$\left\|\sum_{j=1}^m\rho_1(A_j)|G_j(v)|\right\|_s=\left\|\int_U|G(u,v)|\rho_1(\,du)\right\|_s=\left[\int_V\left[\int_U|G(u,v)|\rho_1(\,du)\right]^s\rho_2(\,dv)\right]^{1/s},$$

and this is the left-hand side of formula (11.3), while

$$\begin{aligned}\sum_{j=1}^m\|\rho_1(A_j)|G_j(v)|\|_s&=\sum_{j=1}^m\rho_1(A_j)\left[\int_V|G_j(v)|^s\rho_2(\,dv)\right]^{1/s}\\&=\int_U\left[\int_V|G(u,v)|^s\rho_2(\,dv)\right]^{1/s}\rho_1(\,du),\end{aligned}$$

and this is the right-hand side of (11.3).

Using inequality (11.3) in our case on the space $(X_1\times Y_2,\mathcal{A}_1\times\mathcal{B}_2,\mu_1\times\nu_2)$ with the choice $s=\frac pq$, $U=X_1$, $V=Y_2$, $G(u,v)=\left|\sum\limits_{j=1}^n c_jf_j(x_1)\mathbf{T}_2g_j(y_2)\right|^q$, $\rho_1=\mu_1$, $\rho_2=\nu_2$

we get with the help of formula (11.2) that

$$\int_{Y_1\times Y_2}\left|\sum_{j=1}^n c_j\mathbf{T}_1 f_j(y_1)\mathbf{T}_2 g_j(y_2)\right|^p \nu_1(\,dy_1)\nu_2(\,dy_2)$$
$$\le\left(\int_{X_1}\left[\int_{Y_2}\left|\sum_{j=1}^n c_j f_j(x_1)\mathbf{T}_2 g_j(y_2)\right|^p \nu_2(\,dy_2)\right]^{q/p}\mu_1(\,dx_1)\right)^{p/q}.$$

Then by exploiting that $|\mathbf{T}_2|\le 1$ we get that

$$\left[\int_{Y_2}\left|\sum_{j=1}^n c_j f_j(x_1)\mathbf{T}_2 g_j(y_2)\right|^p \nu_2(\,dy_2)\right]^{q/p}\le\int_{X_2}\left|\sum_{j=1}^n c_j f_j(x_1)g_j(x_2)\right|^q \mu_2(\,dx_2)$$

for all $x_1\in X_1$, and

$$\int_{Y_1\times Y_2}\left|\sum_{j=1}^n c_j\mathbf{T}_1 f_j(y_1)\mathbf{T}_2 g_j(y_2)\right|^p \nu_1(\,dy_1)\nu_2(\,dy_2)$$
$$\le\left(\int_{X_1}\left[\int_{X_2}\left|\sum_{j=1}^n c_j f_j(x_1)g_j(x_2)\right|^q \mu_2(\,dx_2)\right]\mu_1(\,dx_1)\right)^{p/q}.$$

By the Fubini theorem this inequality is equivalent to relation (11.1).

Theorem 11.2 enables us to reduce the proof of the hypercontractive inequality for Rademacher functions to the following simpler result.

Theorem 11.3. The reduced form of the hypercontractive inequality for Rademacher functions. *Let ε be a random variable such that $P(\varepsilon=1)=P(\varepsilon=-1)=\frac12$. Then the following inequality holds for all real (or complex) numbers a, b, and numbers $1\le q\le p<\infty$ together with some $0\le\gamma\le\sqrt{\frac{q-1}{p-1}}$:*

$$E\left(|a+\gamma b\varepsilon|^p\right)^{1/p}\le\left(E|a+b\varepsilon|^q\right)^{1/q} \tag{11.4}$$

Theorems 11.3 and 11.2 really imply Theorem 11.1, because Theorem 11.3 states the desired result in the special case $n=1$, and then by Theorem 11.2 it holds for arbitrary n.

Even the proof of Theorem 11.3 is far from trivial. On the other hand, Leonhard Gross has made a deep and interesting investigation in his paper *Logarithmic Sobolev* *inequalities* (Amer. J. Math. 97, 1061-1083, 1975) which supplies this result as a special case of a general theory. His approach is based on the following idea. Let us consider a continuous time Markov process $\xi(t)$, $t \ge 0$, with its stationary distribution and a function $f(x)$ on the state space of this Markov process. We can get good estimates on the moments $E|f(\xi(t))|^p$ if we have an appropriate estimate on the infinitesimal operator of the Markov process he calls logarithmic Sobolev inequality. In an informal way this approach can be interpreted as a good estimate of a function by means of its derivative.

Gross applies a rather hard analysis in his proof, but if we restrict our attention to that example which leads to the proof of Theorem 11.3, then the most difficult parts of his study do not appear. Here we shall follow this approach.

Let us define the Markov process ξ_t describing the movement of a particle on the state space $X = \{-1, 1\}$ consisting of two points, where the particle jumps from one state to the other one after exponential time with parameter $\lambda = \frac{1}{2}$. This means that the places of jumps constitute a Poisson process with parameter $\lambda = \frac{1}{2}$, and the transition probabilities of this Markov process are

$$p_t(1,1) = p_t(-1,-1) = e^{-t/2} \sum_{k=0}^{\infty} \frac{1}{(2k)!} \left(\frac{t}{2}\right)^{2k},$$
$$p_t(1,-1) = p_t(-1,1) = e^{-t/2} \sum_{k=0}^{\infty} \frac{1}{(2k+1)!} \left(\frac{t}{2}\right)^{2k+1}.$$

(The particle remains in the same place after time t if it made an even number of jumps in the time interval $[0,t]$, and changes his position if it made an odd number of jumps.) Let us calculate the semigroup U_t, $t \ge 0$, of this Markov process, defined as $U_t(f)(x) = E(f(\xi(t))|\xi(0) = x)$, for all $x \in X$, all functions f defined on X and parameters $t \ge 0$ together with the infinitesimal operator of this Markov process $Bf(x) = -\left.\frac{dU_t(f)(x)}{dt}\right|_{t=0}$. The above objects can be simply calculated in this model. Let us introduce the functions $r_0(x)$ and $r_1(x)$ on the state space X defined as $r_0(1) = r_0(-1) = 1$ and $r_1(1) = 1$, $r_1(-1) = -1$. Observe that $p_t(1,1) - p_t(1,-1) = e^{-t}$, $p_t(-1,-1) - p_t(-1,1) = e^{-t}$, hence $U_t r_1(1) = e^{-t} r_1(1)$, $U_t r_1(-1) = e^{-t} r_1(-1)$, i.e. $U_t r_1(x) = e^{-t} r_1(x)$ for all $t \ge 0$. On the other hand, clearly $U_t r_0(x) = r_0(x)$ for all $t \ge 0$. All functions f on the state space X can be written in the form $f(x) = a + b r_1(x)$ with some appropriate coefficients a and b, and $U_t(a + b r_1)(x) = a + e^{-t} b r_1(x)$. Clearly $B(a + b r_1)(x) = b r_1(x)$. Let μ, $\mu(1) = \mu(-1) = \frac{1}{2}$, denote the equilibrium state of the Markov process $\xi(t)$. Put $\|f\|_p = \left(\int |f(x)|^p \mu(\,dx)\right)^{1/p} = \left(\frac{1}{2}(|f(1)|^p + |f(-1)|^p\right)^{1/p}$. The following inequality will be proved, which is the logarithmic Sobolev inequality in the special model considered here. The notations introduced before will be preserved.

Proposition 11.4. *Let us consider a function $f(x) = a + b r_1(x)$ on the space $X = \{-1, 1\}$ with the probability measure μ, $\mu(1) = \mu(-1) = \frac{1}{2}$, on X such that both a and* *b are real numbers, and $a \ge |b|$. Then*

$$\int f^p(x)\ln f(x)\mu(\,dx) \le \frac{p}{2(p-1)}\int f^{p-1}(x)Bf(x)\mu(\,dx) + \|f\|_p^p \ln \|f\|_p, \qquad \text{for all } 1 < p < \infty. \tag{11.5}$$

(The letter B in formula (11.5) denotes the infinitesimal operator of the Markov process we consider.)

The corresponding result in Gross' paper is slightly more general. It contains such an estimate which holds for all functions f, i.e. the condition that a and b are real numbers, and $a \ge |b|$ in the expansion $f = a + br_1$ is not needed there. Our restriction makes the proof simpler, since this implies that the function $f(x)$ is real-valued and $f(x) \ge 0$ both for $x = 1$ and $x = -1$. Hence we do not have to work with absolute values. On the other hand, Proposition 11.4 is sufficient for us also in this restricted form. Before its proof we show that it implies Theorem 11.3.

The proof of Theorem 11.3 by means of Proposition 11.4. Let us introduce the function $p(t,q) = 1 + (q-1)e^{2t}$ for all $q > 1$, and $t \ge 0$. First we prove that

$$\left[\int |U_t f(x)|^{p(t,q)}\mu(\,dx)\right]^{1/p(t,q)} \le \left[\int |f(x)|^q \mu(\,dx)\right]^{1/q} \quad \text{for all } t \ge 0 \tag{11.6}$$

and functions f on X. (The general theory helps to find the 'right' definition of the function $p(t,q)$. It is defined as the solution of the differential equation $\frac{p}{2(p-1)}\frac{dp(t)}{dt} = p$, $p(0) = q$. The coefficient $\frac{p}{2(p-1)}$ in this equation agrees with the coefficient appearing in the logarithmic Sobolev inequality (11.5).) Let us prove inequality (11.6) first for such functions $f(x) = a + br_1(x)$ for which a and b are real numbers and $a \ge |b|$.

Given a function $f(x) = a + br_1(x)$ with $a \ge |b|$ define the function $F(t) = \left[\int (U_t f(x))^{p(t,q)}\mu(\,dx)\right]^{1/p(t,q)}$. Observe that $U_t f(x) = a + be^{-t}r_1(x)$, and $a \ge |b|e^{-t}$. Hence to prove (11.6) it is enough to show that

$$\frac{d\|U_t(f)\|_{p(t,q)}}{dt} = \frac{dF(t)}{dt} \le 0 \quad \text{for all } t > 0 \tag{11.7}$$

which means that the function $F(t)$ is monotone decreasing, and in the proof we can apply the logarithmic Sobolev inequality for the functions $f_t(x) = U_t f(x)$. We have

$$\frac{dF(t)}{dt} = F(t)\biggl[-\frac{p'(t,q)}{p(t,q)}\ln F(t) + \frac{p'(t,q)}{p(t,q)}\frac{\int U_t f(x)^{p(t,q)}\ln U_t f(x)\mu(\,dx)}{\int U_t f(x)^{p(t,q)}\mu(\,dx)} + \frac{\int U_t f(x)^{p(t,q)-1}(U_t f(x))'\mu(d\,x)}{\int U_t f(x)^{p(t,q)}\mu(\,dx)}\biggr],$$

where $G(t,\cdot)'$ means partial derivative with respect to the variable t. Since $F(t) = \|U_t(f)\|_{p(t,q)}$, $\int U_t f(x)^{p(t,q)}\mu(\,dx) = \|U_t(f)\|_{p(t,q)}^{p(t,q)}$, $(U_t f(x))' = -BU_t f(x)$ by the definition of the operator B,

$$\int U_t f(x)^{p(t,q)-1}(U_t f(x))'\mu(d\,x) = -\int U_t f(x)^{p(t,q)-1}B(U_t f)(x)\mu(d\,x),$$

and $\frac{p(t,q)}{p'(t,q)}=\frac{p(t,q)}{2(p(t,q)-1)}$ with our choice of functions, the last formula implies that the inequality $\frac{dF(t)}{dt}\le 0$ is equivalent to the relation

$$\begin{aligned}-\|U_t(f)\|_{p(t,q)}^{p(t,q)}\ln\|U_t(f)\|+\int U_tf^{p(t,q)}(x)\ln U_tf(x)\mu(\,dx)\\ -\frac{p}{2(p-1)}\int (U_tf)^{p(t,q)-1}(x)BU_tf(x)\mu((\,dx)\le 0.\end{aligned}$$

But this inequality follows from the logarithmic Sobolev inequality if it is applied for the function $U_t(f)$ with $\bar p=p(t,q)$.

To prove relation (11.6) for a general function f it is enough to check that $|U_t(f)|\le U_t(|f|)$, i.e. $|U_t(f)(1)|\le U_t(|f|)(1)$ and $|U_t(f)(-1)|\le U_t(|f|)(-1)$ for arbitrary function f and $t\ge 0$, since this relation has been already proved for the function $|f|$. But this relation simply follows from the following calculation. If $f(1)=A$, $f(-1)=B$, then $f(x)=\frac{A+B}{2}+\frac{A-B}{2}r_1(x)$, $U_tf(x)=\frac{A+B}{2}+e^{-t}\frac{A-B}{2}r_1(x)$, i.e. $U_tf(1)=\frac{1+e^{-t}}{2}A+\frac{1-e^{-t}}{2}B$, and $U_tf(-1)=\frac{1-e^{-t}}{2}A+\frac{1+e^{-t}}{2}B$, while $(U_t|f|)(\pm1)=\frac{1+e^{-t}}{2}|A|+\frac{1\mp e^{-t}}{2}|B|$.

Let us fix some numbers $1<p\le q<\infty$ and apply formula (11.6) for some function $f(x)=a+br_1(x)$ with the number t which is the solution of the equation $p(t,q)=p$. Then $e^{-t}=\gamma(p,q)=\sqrt{\frac{q-1}{p-1}}$, $U_t(a+br_1(x))=a+\gamma(p,q)r_1(x)$, hence $\|a+\gamma(p,q)br_1(x)\|_p\le\|a+br_1(x)\|_q$. Given some $\gamma\le\gamma_p$, let us define $\bar p$ as the solution of the equation $\gamma=\sqrt{\frac{q-1}{\bar p-1}}$. Then $\bar p\ge p$, hence $\|a+\gamma br_1(x)\|_p\le\|a+\gamma br_1(x)\|_{\bar p}\le\|a+br_1(x)\|_q$, and this relation is equivalent to formula (11.4). Thus Theorem 11.3 is proved with the help of Proposition 11.4.

The proof of Proposition 11.4. Let us prove relation (11.5) first in the special case $p=2$. We have to show that

$$\int Bf\cdot f\,d\mu+\frac12\int f^2\,d\mu\ln\left(\int f^2\,d\mu\right)-\int f^2\ln f\,d\mu\ge 0$$

for a function of the form $f=a+br_1$, $a\ge|b|$. Since the left-hand side of this inequality is homogeneous of order 2 it is enough to prove this inequality in the special case $f=1+sr_1$, $|s|\le1$. In this case the inequality we want to prove can be written as

$$h(s)=s^2+\frac12(1+s^2)\ln(1+s^2)-\frac12\left[(1+s)^2\ln(1+s)+(1-s)^2\ln(1-s)\right]\ge 0.$$

Simple calculation shows that $h'(s)=2s+s\ln(1+s^2)-(1+s)\ln(1+s)+(1-s)\ln(1-s)$, and $h''(s)=\frac{2s^2}{1+s^2}+\ln(1+s^2)-\ln(1-s^2)=\frac{2s^2}{1+s^2}-\ln\frac{1-s^2}{1+s^2}=\frac{2s^2}{1+s^2}-\ln\left(1-\frac{2s^2}{1+s^2}\right)\ge 0$ for all $0\le s\le 1$. This means that the function $h(s)$ convex. On the other hand $h(0)=h'(0)=0$. These relations imply that $h(s)\ge0$ for all $0\le s\le 1$ as we have claimed.

In the general case $p>1$ let us apply inequality (11.5) in the already proven case $p=2$ for the function $f^{p/2}$. We get that $\frac p2\int f^p(x)\ln f(x)\mu(\,dx)\le\int f^{p/2}(x)Bf^{p/2}(x)\mu(dx)+$

$\frac{p}{2}\|f\|_p^p \ln \|f\|_p$. Hence to prove Proposition 11.4 in the general case it is enough to show that

$$\int f^{p/2}(x)Bf^{p/2}(x)\mu(\,dx) \le \frac{p^2}{4(p-1)}\int f^{p-1}(x)Bf(x)\mu(\,dx)$$

for a function $f(x) = a + br_1(x)$ such that $a \ge |b|$.

The expressions in the last inequality can be simply calculated. As

$$\frac{1}{2^{p/2-1}} f^{p/2}(x) = \left[\left(\frac{a+b}{2}\right)^{p/2} + \left(\frac{a-b}{2}\right)^{p/2}\right] + \left[\left(\frac{a+b}{2}\right)^{p/2} - \left(\frac{a-b}{2}\right)^{p/2}\right] r_1(x),$$

$$\frac{1}{2^{p/2-1}} Bf^{p/2}(x) = \left[\left(\frac{a+b}{2}\right)^{p/2} - \left(\frac{a-b}{2}\right)^{p/2}\right] r_1(x),$$

and

$$\frac{1}{2^{p-2}} f^{p-1}(x) = \left[\left(\frac{a+b}{2}\right)^{p-1} + \left(\frac{a-b}{2}\right)^{p-1}\right] + \left[\left(\frac{a+b}{2}\right)^{p-1} - \left(\frac{a-b}{2}\right)^{p-1}\right] r_1(x)$$

this inequality, more precisely its version we get by multiplying it by $2^{-(p-2)}$ can be rewritten as

$$\begin{aligned}&\left[\left(\frac{a+b}{2}\right)^{p/2} - \left(\frac{a-b}{2}\right)^{p/2}\right]^2 \\ &\qquad \le \frac{p^2}{4(p-1)}\left[\left(\frac{a+b}{2}\right)^{p-1} - \left(\frac{a-b}{2}\right)^{p-1}\right]\left(\frac{a+b}{2} - \frac{a-b}{2}\right)\end{aligned}$$

or

$$\left(\int_u^v t^{(p-2)/2}\,dt\right)^2 \le \int_u^v t^{p-2}\,dt \cdot \int_u^v 1\,dt$$

with $u = \frac{a-|b|}{2}$ and $v = \frac{a+|b|}{2}$. But the last formula is a simple consequence of the Schwarz inequality. Proposition 11.4 is proved.

Remark: Theorem 11.3 is sharp in the following sense. The transformation T_γ, $T_\gamma(a + br_1(x)) = a + \gamma br_1(x)$ as a transformation from the $L_q(X, \mathcal{X}, \mu)$ space to the space $L_p(X, \mathcal{X}, \mu)$ with $1 < q < p$ has a norm greater then 1 if $\gamma > \sqrt{\frac{q-1}{p-1}}$. To see this let us compare the L_q norm of $1 + \delta r_1(x)$ with the L_p-norm of $T_\gamma r_1(x) = 1 + \gamma\delta r_1(x)$ for a small parameter $\delta > 0$. We have $\|1 + \delta r_1(x)\|_q = \left[\frac{1}{2}\left((1+\delta)^q + (1-\delta)^q\right)\right]^{1/q} = \left[1 + \frac{q(q-1)}{2}\delta^2 + O(\delta^3)\right]^{1/q} = 1 + \frac{q-1}{2}\delta^2 + O(\delta^3)$. Similarly, $\|1+\gamma\delta r_1(x)\|_p = 1 + \frac{p-1}{2}\gamma^2\delta^2 + O(\delta^3)$, and these relations imply the above remark.

11 B.) THE PROOF OF PROPOSITION 10.3.

Proof of Proposition 10.3. Let us use the notation introduced in the formulation of Proposition 10.3, and take another k independent copies $\bar\xi_1^{(j)}, \dots, \bar\xi_n^{(j)}$, $1 \le j \le k$, of the random sequences $\xi_1, \dots, \xi_n$ which are also independent of the sequence $\varepsilon_1, \dots, \varepsilon_n$ appearing in the formulation of Proposition 10.3. Let $\mathcal{F}$ denote the σ-algebra generated by the random variables $\xi_1^{(j)}, \dots, \xi_n^{(j)}$, $1 \le j \le k$, and let us introduce the notation $\xi_l^{(j,1)} = \xi_l^{(j)}$, $\xi_l^{(j,-1)} = \bar\xi_l^{(j)}$, $1 \le l \le n$ and $1 \le j \le k$. Let $\mathcal{V}_k$ denote the set of ± 1 sequences of length k, and for a $v \in \mathcal{V}_k$ let $m(v)$ denote the number of the digits -1 in the sequence $v = (v(1), \dots, v(k))$. Observe that $E\left(f\left(\xi_{l_1}^{(1,v(1))}, \dots, \xi_{l_k}^{(k,v(k))}\right) \middle| \mathcal{F}\right) = 0$ if the ± 1 sequence $(v(1), \dots, v(k))$ contains at least one coordinate -1, (this is the point of the proof where we exploit the canonical property of the function f), and

$$Ef\left(\xi_{l_1}^{(1,1)}, \dots, \xi_{l_k}^{(k,1)} \middle| \mathcal{F}\right) = f\left(\xi_{l_1}^{(1)}, \dots, \xi_{l_k}^{(k)}\right) \quad \text{for all indices } 1 \le l_j \le n,\ 1 \le j \le k.$$

These relations together with the Jensen-inequality for conditional expectations imply that

$$\begin{aligned}
|\bar I_{n,k}(f)|^p &= \left| E\left(\frac{1}{k!} \sum_{v \in \mathcal{V}_k} (-1)^{m(v)} \sum_{\substack{1 \le l_r \le n,\ r=1,\dots,k \\ l_r \ne l_{r'} \text{ if } r \ne r'}} f\left(\xi_{l_1}^{(1,v(1))}, \dots, \xi_{l_k}^{(k,v(k))}\right) \middle| \mathcal{F} \right) \right|^p \\
&\le E\left(\left| \frac{1}{k!} \sum_{v \in \mathcal{V}_k} (-1)^{m(v)} \sum_{\substack{1 \le l_r \le n,\ r=1,\dots,k \\ l_r \ne l_{r'} \text{ if } r \ne r'}} f\left(\xi_{l_1}^{(1,v(1))}, \dots, \xi_{l_k}^{(k,v(k))}\right) \right|^p \middle| \mathcal{F} \right).
\end{aligned}$$

Hence

$$E|\bar I_{n,k}(f)|^p \le E\left| \frac{1}{k!} \sum_{v \in \mathcal{V}_k} (-1)^{m(v)} \sum_{\substack{1 \le l_r \le n,\ r=1,\dots,k \\ l_r \ne l_{r'} \text{ if } r \ne r'}} f\left(\xi_{l_1}^{(1,v(1))}, \dots, \xi_{l_k}^{(k,v(k))}\right) \right|^p. \tag{11.8}$$

Let us introduce the random variables

$$\tilde I_{n,k}(f) = \frac{1}{k!} \sum_{v \in \mathcal{V}_k} (-1)^{m(v)} \sum_{\substack{1 \le l_r \le n,\ r=1,\dots,k \\ l_r \ne l_{r'} \text{ if } r \ne r'}} f\left(\xi_{l_1}^{(1,v(1))}, \dots, \xi_{l_k}^{(k,v(k))}\right) \tag{11.9}$$

and

$$\tilde I_{n,k}(f, \varepsilon) = \frac{1}{k!} \sum_{v \in \mathcal{V}_k} (-1)^{m(v)} \sum_{\substack{1 \le l_r \le n,\ r=1,\dots,k \\ l_r \ne l_{r'} \text{ if } r \ne r'}} \varepsilon_{l_1} \cdots \varepsilon_{l_k} f\left(\xi_{l_1}^{(1,v(1))}, \dots, \xi_{l_k}^{(k,v(k))}\right). \tag{11.9$'$}$$

Let us recall that the number $m(v)$ in these formula denotes the number of the digits -1 in the ± 1 sequence v of length k, i.e. it counts how many random variables $\xi_{l_j}^{(j,1)}$, $1\le j\le k$, were replaced by the 'secondary copy' $\xi_{l_j}^{(j,-1)}$ in the corresponding terms of the sum in (11.9) or (11.9′).

I claim that the above defined two random variables $\tilde I_{n,k}(f)$ and $\tilde I_{n,k}(f,\varepsilon)$ have the same distribution. This statement will be formulated in a slightly more general form which will be useful in the further part of this work.

Lemma 11.5. *Let us consider a (non-empty) class of functions $\mathcal F$ of k variables $f(x_1,\dots,x_k)$ on the space $(X^k,\mathcal X^k)$ together with the random variables $\tilde I_{n,k}(f)$ and $\tilde I_{n,k}(f,\varepsilon)$ defined in formulas (11.9) and (11.9′) for all $f\in\mathcal F$. The joint distributions of the set of random variables $\{\tilde I_{n,k}(f);\, f\in\mathcal F\}$ and $\{\tilde I_{n,k}(f,\varepsilon);\, f\in\mathcal F\}$ agree.*

The proof of Lemma 11.5. We even claim that fixing an arbitrary sequence $u=(u(1),\dots,u(n))$, $u(l)=\pm1$, $1\le l\le n$, of length n, the conditional distribution of the field $\{\tilde I_{n,k}(f,\varepsilon);\, f\in\mathcal F\}$ under the condition that $(\varepsilon_1,\dots,\varepsilon_n)=u=(u(1),\dots,u(n))$ agrees with the distribution of the field of $\{\tilde I_{n,k}(f);\, f\in\mathcal F\}$.

Indeed, the random variables $\tilde I_{n,k}(f)$, $f\in\mathcal F$, defined in (11.9) are functions of a random vector consisting of coordinates $(\xi_l^{(j)},\bar\xi_l^{(j)})=(\xi_l^{(j,1)},\xi_l^{(j,-1)})$, $1\le l\le n$, $1\le j\le k$, and the distribution of this random vector does not change if we replace the coordinates $(\xi_l^{(j)},\bar\xi_l^{(j)})=(\xi_l^{(j,1)},\xi_l^{(j,-1)})$, by $(\bar\xi_l^{(j)},\xi_l^{(j)})=(\xi_l^{(j,-1)},\xi_l^{(j,1)})$, for those indices (j,l) for which $u(l)=-1$ (independently of the value of the parameter j) and do not modify these random vectors for those coordinates (l,j) for which $u(l)=1$. Replacing the original vector in the definition of the expression $\tilde I_{n,k}(f)$ in (11.9) for all $f\in\mathcal F$ by this modified vector we carry out a measure preserving transformation. On the other hand, the random field we get in such a way has the same distribution as the conditional distribution of the random field $\tilde I_{n,k}(f,\varepsilon)$, $f\in\mathcal F$, with the elements defined in (11.9′) under the condition that $(\varepsilon_1,\dots,\varepsilon_n)=u$ with $u=(u(1),\dots,u(n))$.

To prove the last statement let us observe that the conditional distribution of the random field $\tilde I_{n,k}(f,\varepsilon)$, $f\in\mathcal F$, under the condition $(\varepsilon_1,\dots,\varepsilon_n)=u$ is the same as that of the random field we obtain by putting $u_l=\varepsilon_l$, $1\le l\le n$, in all coordinates ε_l of the random variables $\tilde I_{n,k}(f,\varepsilon)$. On the other hand, the random variables we get in such a way agree with the random variables we get by carrying out the above described transformation for the random variables $\tilde I_{n,k}(f)$, only the terms in the sums defining these random variables are listed in a different order.

Relation (11.8) and the agreement of the distribution of the random variables $\tilde I_{n,k}(f)$ in (11.9) and $\tilde I_{n,k}(f)$ (11.9′) imply that

$$E|\bar I_{n,k}(f)|^p\le E\left|\frac1{k!}\sum_{v\in\mathcal V_k}(-1)^{m(v)}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if }j\ne j'}}\varepsilon_{l_1}\cdots\varepsilon_{l_k}f\left(\xi_{l_1}^{(1,v(1))},\dots,\xi_{l_k}^{(k,v(k))}\right)\right|^p. \tag{11.10}$$

Let us define for all $v=(v(1),\dots,v(k))\in\mathcal{V}_k$ the random variable

$$\bar I_{n,k,v}(f,\varepsilon)=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'} \text{ if } j\ne j'}}\varepsilon_{l_1}\cdots\varepsilon_{l_k} f\left(\xi_{l_1}^{(1,v(1))},\dots,\xi_{l_k}^{(k,v(k))}\right),\quad v\in V_k.$$

The distribution of the random variables $\bar I_{n,k,v}(f,\varepsilon)$ agree with that of $\bar I_{n,k}(f,\varepsilon)$ introduced in (10.6) for all $v\in\mathcal{V}_k$. Hence relation (11.10) implies that

$$\begin{aligned}E|\bar I_{n,k}(f)|^p&\le E\left|\sum_{v\in\mathcal{V}_k}(-1)^{m(v)}\bar I_{n,k,v}(\varepsilon,f)\right|^p\\&\le 2^{(k-1)p}\sum_{v\in\mathcal{V}_k}E|\bar I_{n,k,v}(f,\varepsilon)|^p=2^{kp}E|\bar I_{n,k}(f,\varepsilon)|^p.\end{aligned}$$

Proposition 10.3 is proved.

12. Reduction of the main result in this work

The main result of this paper is Theorem 8.4 or its multiple integral version Theorem 8.2. It can be considered as the multivariate version of Theorem 4.1, and its proof is also based on a similar argument. Following the method of the proof of Theorem 4.1 first we prove a multivariate version of Proposition 6.1 in Proposition 12.1 and reduce Theorem 8.4 to a simpler statement formulated in Proposition 12.2.

The hard part of the problem is the proof of Proposition 12.2. In the first step of its proof we reduce it with the help of Theorem 10.4 (proved by de la Peña and Montgomery–Smith) to an analogous result formulated in Proposition 12.2′, where the U-statistics to be investigated are replaced by their decoupled U-statistics counterpart introduced in Section 10. The proof of this result is simpler, because here we have more independence. It is based on a symmetrization argument, similar to the proof of Proposition 6.2. The details of this symmetrization argument will be explained in the next section. This section contains only an important preliminary result needed in this argument, a multi-dimensional variant of Hoeffding's inequality (Theorem 3.4) formulated in Theorem 12.3. It yields an estimate about the distribution of homogeneous polynomials of Rademacher functions.

The first result of this Section, Proposition 12.1, can be proved in almost the same way as its simplified version Proposition 6.1. The only essential difference between their proof is that Bernstein's inequality applied in the proof of Proposition 6.1 is replaced now by its multivariate version Theorem 8.3. Theorem 12.1 can be considered as the result we can get by means of the Theorem 8.3 and the chaining argument. Its main content, formulated in relation (12.1) states that given a nice class of functions $\mathcal{F}$ it has a subclass $\mathcal{F}_{\bar\sigma}$ of relatively small cardinality which is also a relatively dense subclass of $\mathcal{F}$ in the L_2 norm, and the supremum of the U-statistics with kernel functions from $\mathcal{F}_{\bar\sigma}$ can be well bounded. To get an applicable result we also need some estimates on the number $\bar\sigma$ which measures how dense the subclass $\mathcal{F}_{\bar\sigma}$ in $\mathcal{F}$ is. Such estimates are contained at the end of this Proposition.

In the formulation of Proposition 12.1 we introduce, similarly to Proposition 6.1, two parameters $\bar A > 2^k$ and $M = M(\bar A, k)$, and this may seem at first sight unnatural. But the introduction of these parameters turned out to be useful, they help, similarly to the analogous problem in Section 6 to fit the parameters in Propositions 12.1 and 12.2 as we want to apply them simultaneously.

Proposition 12.1. *Let us have the k-fold power $(X^k,\mathcal{X}^k)$ of a measurable space $(X,\mathcal{X})$ with some probability measure μ on $(X,\mathcal{X})$ and a countable L_2-dense class $\mathcal{F}$ of functions $f(x_1,\dots,x_k)$ of k variables on $(X^k,\mathcal{X}^k)$ with parameter D and exponent L, $L\ge1$, such that all functions $f\in\mathcal{F}$ are canonical with respect to the measure μ, and they satisfy conditions (8.4) and (8.5) with some real number $0<\sigma\le1$. Take a sequence of independent μ-distributed random variables $\xi_1,\dots,\xi_n$, $n\ge\max(k,2)$, and consider the (degenerate) U-statistics $I_{n,k}(f)$, $f\in\mathcal{F}$, defined in formula (8.7). Let us fix some number $\bar A\ge 2^k$.*

For all numbers $M=M(k,\bar A)$ which are chosen sufficiently large in dependence of $\bar A$ and k the following relation depending on the numbers $\bar A$ and M holds: For all numbers $u>0$ for which $n\sigma^2\ge\left(\frac{u}{\sigma}\right)^{2/k}\ge ML\log\frac{2}{\sigma}$ a number $\bar\sigma=\bar\sigma(u)$, $0\le\bar\sigma\le\sigma\le1$, and a collection of functions $\mathcal{F}_{\bar\sigma}=\{f_1,\dots,f_m\}\subset\mathcal{F}$ with $m\le D\bar\sigma^{-L}$ elements can be chosen in such a way that the sets $\mathcal{D}_j=\{f\colon f\in\mathcal{F},\int|f-f_j|^2\,d\mu\le\bar\sigma^2\}$, $1\le j\le m$, satisfy the relation $\bigcup\limits_{j=1}^m\mathcal{D}_j=\mathcal{F}$, and the (degenerate) U-statistics $I_{n,k}(f)$, $f\in\mathcal{F}_{\bar\sigma(u)}$, satisfy the inequality

$$P\left(\sup_{f\in\mathcal{F}_{\bar\sigma(u)}} n^{-k/2}|I_{n,k}(f)|\ge\frac{u}{\bar A}\right)\le 2CD\exp\left\{-\alpha\left(\frac{u}{10\bar A\sigma}\right)^{2/k}\right\}$$
$$\text{if}\quad n\sigma^2\ge\left(\frac{u}{\sigma}\right)^{2/k}\ge ML\log\frac{2}{\sigma} \tag{12.1}$$

with the constants $\alpha=\alpha(k)$, $C=C(k)$ appearing in formula (8.9) of Theorem 8.3 and the exponent L and parameter D of the L_2-dense class $\mathcal{F}$.

The inequalities $4\left(\frac{u}{\bar A\bar\sigma}\right)^{2/k}\ge n\bar\sigma^2\ge\frac{1}{64}\left(\frac{u}{\bar A\sigma}\right)^{2/k}$ and $n\bar\sigma^2\ge\frac{M^{2/3}(L+\beta)\log n}{1000\bar A^{4/3}}$ also hold, provided that $n\sigma^2\ge\left(\frac{u}{\sigma}\right)^{2/k}\ge M(L+\beta)^{3/2}\log\frac{2}{\sigma}$ with $\beta=\max\left(\frac{\log D}{n},0\right)$.

Proof of Proposition 12.1. Let us list the elements of the countable set $\mathcal{F}$ as $f_1,f_2,\dots$. For all $p=0,1,2,\dots$ let us choose, by exploiting the L_2-density property of the class $\mathcal{F}$, a set $\mathcal{F}_p=\{f_{a(p,1)},\dots,f_{a(p,m_p)}\}\subset\mathcal{F}$ with $m_p\le D\,2^{2pL}\sigma^{-L}$ elements in such a way that $\inf\limits_{1\le j\le m_p}\int(f-f_{a(p,j)})^2\,d\mu\le2^{-4p}\sigma^2$ for all $f\in\mathcal{F}$. For all indices $a(j,p)$, $p=1,2,\dots$, $1\le j\le m_p$, choose a predecessor $a(j',p-1)$, $j'=j'(j,p)$, $1\le j'\le m_{p-1}$, in such a way that the functions $f_{a(j,p)}$ and $f_{a(j',p-1)}$ satisfy the relation $\int|f_{a(j,p)}-f_{a(j',p-1)}|^2\,d\mu\le\sigma^2 2^{-4(p-1)}$. Then we have $\int\left(\frac{f_{a(j,p)}-f_{a(j',p-1)}}{2}\right)^2 d\mu\le4\sigma^2 2^{-4p}$ and

$\sup_{x_j\in X,\,1\le j\le k}\left|\frac{f_{a(j,p)}(x_1,\dots,x_k)-f_{a(j',p-1)}(x_1,\dots,x_k)}{2}\right|\le 1$. Theorem 8.3 yields that

$$
\begin{aligned}
P(A(j,p)) &= P\left(n^{-k/2}|I_{n,k}(f_{a(j,p)}-f_{a(j',p-1)})|\ge \frac{2^{-(1+p)}u}{\bar A}\right)\\
&\le C\exp\left\{-\alpha\left(\frac{2^pu}{8\bar A\sigma}\right)^{2/k}\right\}\quad\text{if}\quad 4n\sigma^2 2^{-4p}\ge\left(\frac{2^pu}{8\bar A\sigma}\right)^{2/k},\\
&\qquad 1\le j\le m_p,\ p=1,2,\dots,
\end{aligned}
\tag{12.2}
$$

and

$$
P(B(s))=P\left(n^{-k/2}|I_{n,k}(f_{0,s})|\ge\frac{u}{2\bar A}\right)\le C\exp\left\{-\alpha\left(\frac{u}{2\bar A\sigma}\right)^{2/k}\right\},\quad 1\le s\le m,
$$
$$
\text{if}\quad n\sigma^2\ge\left(\frac{u}{2\bar A\sigma}\right)^{2/k}. \tag{12.3}
$$

Introduce an integer $R=R(u)$, $R>0$, which satisfies the relations

$$
2^{(4+2/k)(R+1)}\left(\frac{u}{\bar A\sigma}\right)^{2/k}\ge 2^{2+6/k}n\sigma^2\ge 2^{(4+2/k)R}\left(\frac{u}{\bar A\sigma}\right)^{2/k},
$$

and define $\bar\sigma^2=2^{-4R}\sigma^2$ and $\mathcal{F}_{\bar\sigma}=\mathcal{F}_R$ (i.e the class of functions $\mathcal{F}_p$ introduced before with $p=R$). (As $n\sigma^2\ge\left(\frac{u}{\sigma}\right)^{2/k}$ and $\bar A\ge 2^k$ by our conditions, there exists such a positive integer R.) Then the cardinality m of the set $\mathcal{F}_{\bar\sigma}$ is clearly not greater than $D\bar\sigma^{-L}$, and $\bigcup\limits_{j=1}^m\mathcal{D}_j=\mathcal{F}$. Besides, the number R was chosen in such a way that the inequalities (12.2) and (12.3) hold for $1\le p\le R$. Hence the definition of the predecessor of an index $a(j,p)$ implies that

$$
\begin{aligned}
P\left(\sup_{f\in\mathcal{F}_{\bar\sigma}}n^{-k/2}|I_{n,k}(f)|\ge\frac{u}{\bar A}\right)&\le P\left(\bigcup_{p=1}^R\bigcup_{j=1}^{m_p}A(j,p)\cup\bigcup_{s=1}^m B(s)\right)\\
&\le\sum_{p=1}^R\sum_{j=1}^{m_p}P(A(j,p))+\sum_{s=1}^m P(B(s))\le\sum_{p=1}^\infty CD\,2^{2pL}\sigma^{-L}\exp\left\{-\alpha\left(\frac{2^pu}{8\bar A\sigma}\right)^{2/k}\right\}\\
&\qquad+CD\sigma^{-L}\exp\left\{-\alpha\left(\frac{u}{2\bar A\sigma}\right)^{2/k}\right\}.
\end{aligned}
$$

If the condition $\left(\frac{u}{\sigma}\right)^{2/k}\ge ML^{3/2}\log\frac{2}{\sigma}$ holds with a sufficiently large constant M (depending on $\bar A$), then the inequalities

$$
2^{2pL}\sigma^{-L}\exp\left\{-\alpha\left(\frac{2^pu}{8\bar A\sigma}\right)^{2/k}\right\}\le 2^{-p}\exp\left\{-\alpha\left(\frac{2^pu}{10\bar A\sigma}\right)^{2/k}\right\}
$$

hold for all $p = 1, 2, \dots$, and

$$\sigma^{-L} \exp\left\{-\alpha\left(\frac{u}{2\bar A\sigma}\right)^{2/k}\right\} \le \exp\left\{-\alpha\left(\frac{u}{10\bar A\sigma}\right)^{2/k}\right\}.$$

Hence the previous estimate implies that

$$P\left(\sup_{f\in\mathcal{F}_{\bar\sigma}} n^{-k/2}|I_{n,k}(f)| \ge \frac{u}{\bar A}\right) \le \sum_{p=1}^{\infty} CD2^{-p}\exp\left\{-\alpha\left(\frac{2^p u}{10\bar A\sigma}\right)^{2/k}\right\}$$
$$+ CD\exp\left\{-\alpha\left(\frac{u}{10\bar A\sigma}\right)^{2/k}\right\} \le 2CD\exp\left\{-\alpha\left(\frac{u}{10\bar A\sigma}\right)^{2/k}\right\},$$

and relation (12.1) holds. We have

$$n\bar\sigma^2 = 2^{-4R}n\sigma^2 \le 2^{-4R}\cdot 2^{(4+2/k)(R+1)-2-6/k}\left(\frac{u}{\bar A\sigma}\right)^{2/k} = 2^{2-4/k}\cdot 2^{2R/k}\left(\frac{u}{\bar A\sigma}\right)^{2/k}$$
$$= 2^{2-4/k}\cdot\left(\frac{\sigma}{\bar\sigma}\right)^{1/k}\left(\frac{u}{\bar A\sigma}\right)^{2/k} = 2^{2-4/k}\cdot\left(\frac{\bar\sigma}{\sigma}\right)^{1/k}\left(\frac{u}{\bar A\bar\sigma}\right)^{2/k},$$

hence $n\bar\sigma^2 \le 4\left(\frac{u}{\bar A\bar\sigma}\right)^{2/k}$. Besides, as $n\sigma^2 \ge 2^{(4+2/k)R-2-6/k}\left(\frac{u}{\bar A\sigma}\right)^{2/k}$, $R \ge 1$,

$$n\bar\sigma^2 = 2^{-4R}n\sigma^2 \ge 2^{-2-6/k}\cdot 2^{2R/k}\left(\frac{u}{\bar A\sigma}\right)^{2/k} \ge \frac{1}{64}\left(\frac{u}{\bar A\sigma}\right)^{2/k}.$$

It remained to show that $n\bar\sigma^2 \ge \frac{M^{2/3}(L+\beta)\log n}{1000\bar A^{4/3}}$.

This inequality clearly holds under the conditions of Proposition 12.1 if $\sigma \le n^{-1/3}$, since in this case $\log\frac{2}{\sigma} \ge \frac{\log n}{3}$, and $n\bar\sigma^2 \ge \frac{1}{64}\left(\frac{u}{\bar A\sigma}\right)^{2/k} \ge \frac{1}{64}\bar A^{-2/k}M(L+\beta)^{3/2}\log\frac{2}{\sigma} \ge \frac{1}{192}\bar A^{-2/k}M(L+\beta)\log n \ge \frac{M^{2/3}(L+\beta)\log n}{1000\bar A^{4/3}}$ if $M = M(\bar A, k)$ is chosen sufficiently large.

If $\sigma \ge n^{-1/3}$, then the inequality $2^{(4+2/k)R}\left(\frac{u}{\bar A\sigma}\right)^{2/k} \le 2^{2+6/k}n\sigma^2$ holds. Hence $2^{-4R} \ge 2^{-4(2+6/k))/(4+2/k)}\left[\frac{\left(\frac{u}{\bar A\sigma}\right)^{2/k}}{n\sigma^2}\right]^{4/(4+2/k)}$, and

$$n\bar\sigma^2 = 2^{-4R}n\sigma^2 \ge \frac{2^{-16/3}}{\bar A^{4/3}}(n\sigma^2)^{1-\gamma}\left[\left(\frac{u}{\sigma}\right)^{2/k}\right]^{\gamma} \text{ with } \gamma = \frac{4}{4+\frac{2}{k}} \ge \frac{2}{3}.$$

Since $n\sigma^2 \ge (\frac{u}{\sigma})^{2/k} \ge \frac{M}{3}(L+\beta)^{3/2}$, and $n\sigma^2 \ge n^{1/3}$, the above estimates yield that $(n\sigma^2)^{1-\gamma}\left[\left(\frac{u}{\sigma}\right)^{2/k}\right]^{\gamma} \ge (n\sigma^2)^{1/3}\left[\left(\frac{u}{\sigma}\right)^{2/k}\right]^{2/3}$, and $n\bar\sigma^2 \ge \frac{\bar A^{-4/3}}{50}(n\sigma^2)^{1/3}\left[\left(\frac{u}{\sigma}\right)^{2/k}\right]^{2/3} \ge \frac{\bar A^{-4/3}}{50}n^{1/9}\left(\frac{M}{3}\right)^{2/3}(L+\beta) \ge \frac{M^{2/3}(L+\beta)\log n}{1000\bar A^{4/3}}$.

Now we formulate a multivariate analog of Proposition 6.2 in Proposition 12.2 and show that Propositions 12.1 and 12.2 imply Theorem 8.4.

Proposition 12.2. *Let us have a probability measure μ on a measurable space $(X,\mathcal{X})$ together with a sequence of independent and μ distributed random variables $\xi_1,\dots,\xi_n$ and a countable L_2-dense class $\mathcal{F}$ of canonical kernel functions $f=f(x_1,\dots,x_k)$ (with respect to the measure μ) with some parameter D and exponent L on the product space $(X^k,\mathcal{X}^k)$ such that all functions $f\in\mathcal{F}$ satisfy conditions (8.4) and (8.5) with some $0<\sigma\le 1$, and consider the (degenerate) U-statistics $I_{n,k}(f)$ with the random sequence $\xi_1,\dots,\xi_n$ and kernel functions $f\in\mathcal{F}$. There exists a sufficiently large constant $K=K(k)$ together with some numbers $\bar C=\bar C(k)>0$, $\gamma=\gamma(k)>0$ and threshold index $A_0=A_0(k)>0$ depending only on the order k of the U-statistics such that if $n\sigma^2>K(L+\beta)\log n$ with $\beta=\max\left(\frac{\log D}{\log n},0\right)$, then the degenerate U-statistics $I_{n,k}(f)$, $f\in\mathcal{F}$, satisfy the inequality*

$$P\left(\sup_{f\in\mathcal{F}}|n^{-k/2}I_{n,k}(f)|\ge An^{k/2}\sigma^{k+1}\right)\le \bar Ce^{-\gamma A^{1/2k}n\sigma^2}\quad \textit{if } A\ge A_0. \tag{12.4}$$

We shall prove formula (8.10) by applying Proposition 12.2 with the choice $\sigma=\bar\sigma=\bar\sigma(u)$ defined in Proposition 12.1 and the classes $\mathcal{F}=\mathcal{D}_j$, more precisely the classes $\mathcal{F}=\left\{\frac{g-f_j}{2}\colon g\in\mathcal{D}_j\right\}$ of functions introduced also in Proposition 12.1, where f_j is the function appearing in the definition of the class of functions $\mathcal{D}_j$. Clearly,

$$\begin{aligned}P\left(\sup_{f\in\mathcal{F}}n^{-k/2}|I_{n,k}(f)|\ge u\right)&\le P\left(\sup_{f\in\mathcal{F}_{\bar\sigma}}n^{-k/2}|I_{n,k}(f)|\ge\frac{u}{\bar A}\right)\\&\quad+\sum_{j=1}^m P\left(\sup_{g\in\mathcal{D}_j}n^{-k/2}\left|I_{n,k}\left(\frac{f_j-g}{2}\right)\right|\ge\left(\frac12-\frac{1}{2\bar A}\right)u\right),\end{aligned} \tag{12.5}$$

where m is the cardinality of the set of functions $\mathcal{F}_{\bar\sigma}$ appearing in Proposition 12.1. We want to show that if first $\bar A$ and then $M\ge M_0(\bar A)$ are chosen sufficiently large in Proposition 12.1, then the second term at the right-hand side of formula (12.5) can be well bounded by means of Proposition 12.2, and Theorem 8.4 can be proved by means of this estimate.

To carry out this program let us choose a number $\bar A_0$ in such a way that $\bar A_0\ge A_0$ and $\gamma\bar A_0^{1/2k}\ge\frac1K$ with the numbers A_0, K and γ in Proposition 12.2, put $\bar A=\max(2^{k+2}\bar A_0,2^k)$, and apply Proposition 12.1 with this number $\bar A$. Then by Proposition 12.1 and the choice of the numbers $\bar A$ and $\bar A_0$ also the inequality $\left(\frac{u}{\bar\sigma}\right)^{2/k}\ge\frac{\bar A^{2/k}}{4}n\bar\sigma^2\ge(4\bar A_0)^{2/k}n\bar\sigma^2$ holds, hence $u\ge 4\bar A_0n^{k/2}\bar\sigma^{k+1}$ with the number $\bar\sigma$ in Proposition 12.1. This implies that $\left(\frac12-\frac{1}{2\bar A}\right)u\ge\frac u4\ge\bar A_0n^{k/2}\bar\sigma^{k+1}$, $\bar A_0\ge A_0$, and by replacing the expression $\left(\frac12-\frac{1}{2\bar A}\right)u$ by $\bar A_0n^{k/2}\bar\sigma^{k+1}$ in the probabilities of the sum in the second term at the right-hand side of (12.5) we enlarge them.

The numbers u considered in these estimations satisfy the condition $n\sigma^{2/k}\ge\left(\frac u\sigma\right)^{2/k}\ge M(L+\beta)^{3/2}\log\frac2\sigma$ imposed in Proposition 12.1 with some appropriately chosen constant M. Choose the number $M\ge M(\bar A,k)$ in Proposition 12.1 (which also can be chosen as the number M in formula (8.10) of Theorem 8.4) in such a way that it also satisfies the inequality $\frac{M^{2/3}(L+\beta)\log n}{1000\bar A^{4/3}}\ge K(L+\beta)\log n$ with the number K appearing in the conditions of Proposition 12.2. With such a choice the inequality $n\bar\sigma^2\ge \frac{M^{2/3}(L+\beta)\log n}{1000\bar A^{4/3}}\ge K(L+\beta)\log n$ holds, and Proposition 12.2 can be applied to bound the terms in the sum at the right-hand side of (12.5). It yields the estimate

$$\begin{aligned}P&\left(\sup_{g\in\mathcal{D}_j} n^{-k/2}\left|I_{n,k}\left(\frac{f_j-g}{2}\right)\right|\ge\left(\frac12-\frac{1}{2\bar A}\right)u\right)\\&\le P\left(\sup_{g\in\mathcal{D}_j} n^{-k/2}\left|I_{n,k}\left(\frac{f_j-g}{2}\right)\right|\ge \bar A_0 n^{k/2}\bar\sigma^{k+1}\right)\le \bar C e^{-\gamma\bar A_0^{1/2k}n\bar\sigma^2}\end{aligned}$$

for all $1\le j\le m$. (Observe that the set of functions $\frac{f_j-g}{2}$, $g\in\mathcal{D}_j$ is an L_2-dense class with parameter D and exponent L.) Hence Proposition 12.1 (relation (12.1) together with the inequality $m\le D\bar\sigma^{-L}$) and formula 12.4 with $A=\bar A_0$ imply that

$$P\left(\sup_{f\in\mathcal{F}} n^{-k/2}|I_{n,k}(f)|\ge u\right)\le 2CD\exp\left\{-\alpha\left(\frac{u}{10\bar A\sigma}\right)^{2/k}\right\}+\bar C D\bar\sigma^{-L}e^{-\gamma\bar A_0^{1/2k}n\bar\sigma^2}. \tag{12.6}$$

To get the result of Theorem 8.4 from inequality (12.6) we have to replace its second term at the right-hand side with a more appropriate expression where, in particular, we get rid of the coefficient $\bar\sigma^{-L}$. The condition $n\bar\sigma^2\ge K(L+\beta)\log n$ implies that $\bar\sigma\ge n^{-1/2}$, and by our choice of $\bar A_0$ we have $\gamma\bar A_0^{1/2k}n\bar\sigma^2\ge\frac1K n\bar\sigma^2\ge L\log n\ge 2L\log\frac{1}{\bar\sigma}$, i.e. $\bar\sigma^{-L}\le e^{\gamma\bar A_0^{1/2k}n\bar\sigma^2/2}$. By the estimates of Proposition 12.1 $n\bar\sigma^2\ge\frac{1}{64}\left(\frac{u}{\bar A\sigma}\right)^{2/k}$. The above relations imply that $\bar\sigma^{-L}e^{-\gamma\bar A_0^{1/2k}n\bar\sigma^2}\le e^{-\gamma\bar A_0^{1/2k}n\bar\sigma^2/2}\le\exp\left\{-\frac{\gamma}{128}\bar A_0^{1/2k}\bar A^{-2/k}\left(\frac{u}{\sigma}\right)^{2/k}\right\}$. Hence relation (12.6) yields that

$$\begin{aligned}P&\left(\sup_{f\in\mathcal{F}} n^{-k/2}|I_{n,k}(f)|\ge u\right)\\&\le 2CD\exp\left\{-\frac{\alpha}{(10\bar A)^2}\left(\frac{u}{\sigma}\right)^{2/k}\right\}+\bar C D\exp\left\{-\frac{\gamma}{128}\bar A_0^{1/2k}\bar A^{-2/k}\left(\frac{u}{\sigma}\right)^{2/k}\right\},\end{aligned}$$

and this estimate implies Theorem 8.4.

Thus to complete the proof of Theorem 8.4 it is enough to prove Proposition 12.2. It turned out to be useful to apply an approach similar to the proof of Theorem 8.3. In the proof of Theorem 8.3 first an appropriate counterpart of this result was proved, where the U-statistics were replaced by their decoupled U-statistics analogs defined in formula (10.5), and then the desired result was deduced from this estimate and Theorem 10.4. Similarly, Proposition 12.2 will be deduced from the following result.

Proposition 12.2′. *Let a class of functions $f\in\mathcal{F}$ on the k-fold product $(X^k,\mathcal{X}^k)$ of a measurable space $(X,\mathcal{X})$, a probability measure μ on $(X,\mathcal{X})$ together with a sequence* *of independent and μ distributed random variables $\xi_1,\dots,\xi_n$ satisfy the conditions of Proposition 12.2. Let us take k independent copies $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, of the random sequence $\xi_1,\dots,\xi_n$, and consider the decoupled U-statistics $\bar I_{n,k}(f)$, $f\in\mathcal{F}$, defined with their help by formula (10.5). Then there exists a sufficiently large constant $K=K(k)$ together with some number $\gamma=\gamma(k)>0$ and threshold index $A_0=A_0(k)>0$ depending only on the order k of the decoupled U-statistics we consider such that if $n\sigma^2>K(L+\beta)\log n$ with $\beta=\max\left(\frac{\log D}{\log n},0\right)$, then the (degenerate) decoupled U-statistics $\bar I_{n,k}(f)$, $f\in\mathcal{F}$, satisfy the following version of inequality (12.4):*

$$P\left(\sup_{f\in\mathcal{F}}|n^{-k/2}\bar I_{n,k}(f)|\ge An^{k/2}\sigma^{k+1}\right)\le e^{-\gamma A^{1/2k}n\sigma^2}\quad \text{if } A\ge A_0. \tag{12.7}$$

It is clear that Proposition 12.2′ and Theorem 10.4, more explicitly formula (10.8′) in it imply Proposition 12.2. The proof of Proposition 12.2 is based on a symmetrization argument which will be explained in the next section. Here an important ingredient of it will be proved, the multivariate version of Hoeffding's inequality formulated in Theorem 3.4.

Theorem 12.3. The multivariate version of Hoeffding's inequality. *Let $\varepsilon_1,\dots,\varepsilon_n$ be independent random variables, $P(\varepsilon_j=1)=P(\varepsilon_j=-1)=\frac12$, $1\le j\le n$. Fix a positive integer k, and define the random variable*

$$Z=\sum_{\substack{(j_1,\dots,j_k)\colon\ 1\le j_l\le n \text{ for all } 1\le l\le k\\ j_l\ne j_{l'} \text{ if } l\ne l'}} a(j_1,\dots,j_k)\varepsilon_{j_1}\cdots\varepsilon_{j_k} \tag{12.8}$$

with the help of some real numbers $a(j_1,\dots,j_k)$ which are given for all sets of indices such that $1\le j_l\le n$, $1\le l\le k$, and $j_l\ne j_{l'}$ if $l\ne l'$. Put

$$S^2=\sum_{\substack{(j_1,\dots,j_k)\colon\ 1\le j_l\le n \text{ for all } 1\le l\le k\\ j_l\ne j_{l'} \text{ if } l\ne l'}} a^2(j_1,\dots,j_k) \tag{12.9}$$

Then

$$P(|Z|>u)\le C\exp\left\{-B\left(\frac{u}{S}\right)^{2/k}\right\}\quad \text{for all } u\ge 0 \tag{12.10}$$

with some constants $B>0$ and $C>0$ depending only on the parameter k. Relation (12.10) holds for instance with the choice $B=\frac{k}{2e(k!)^{1/k}}$ and $C=e^k$.

Proof of Theorem 12.3. We get with the help of formula (10.4) (which is a consequence of Borell's inequality) that

$$E|Z|^q\le(q-1)^{kq/2}\left(EZ^2\right)^{q/2}\le q^{kq/2}\left(EZ^2\right)^{q/2}=q^{kq/2}\bar S^q$$

with

$$\bar S^2 = \sum_{1\le j_1<j_2\cdots<j_k\le n} \left(\sum_{\pi\in\Pi_k} a((j_{\pi(1)},\dots,j_{\pi(k)})\right)^2,$$

where Π_k denotes the set of all permutations of the set $\{1,\dots,k\}$. Observe that

$$\left(\sum_{\pi\in\Pi_k} a(j_{\pi(1)},\dots,j_{\pi(k)})\right)^2 \le k!\sum_{\pi\in\Pi_k} a^2(j_{\pi(1)},\dots,j_{\pi(k)}) \quad \text{for all } 1\le j_1<\cdots j_k\le n,$$

hence $\bar S^2 \le k!S^2$, and $E|Z|^q \le q^{kq/2}(k!)^{q/2}S^q$ with the number S^2 defined in (12.9). Thus the Markov inequality implies that

$$P(|Z|>u) \le \left(q^{k/2}\frac{\sqrt{k!}S}{u}\right)^q \quad \text{for all } u>0 \quad \text{and } q\ge 2.$$

Choose the number q as the solution of the equation $q\left(\frac{\sqrt{k!}S}{u}\right)^{2/k}=\frac1e$. Then we get that $P(|Z|>u)\le \exp\left\{-B\left(\frac uS\right)^{2/k}\right\}$ with $B=\frac{k}{2e(k!)^{1/k}}$, provided that $q=\frac{1}{ek!^{1/k}}\left(\frac uS\right)^{2/k}\ge 2$, i.e. $B\left(\frac uS\right)^{2/k}\ge k$. By multiplying the above upper bound with $C=e^k$ we get such an estimate for $P(|Z|>u)$ which holds for all $u>0$.

Remark: The result of Theorem 12.3 will be good enough for our purposes, although the constants B and C we have given in formula (12.10) are not optimal. Thus Theorem 3.4 yields that in the special case $k=1$ the estimate (12.10) holds with $B=\frac12$ and $C=1$ (and not only with $B=\frac{1}{2e}$ and $C=e$). The reason for this relative weakness of Theorem 12.3 is that the moment estimate given for a homogeneous polynomial of Rademacher functions in formula (10.4) is not always sharp. In Theorem 16.6 I present (without proof) an improved version of Theorem 12.3 which yields the estimate (12.10) with the right constant C in the exponent. The proof can be found in paper [22]. It is based on a sharp estimate on the moments EZ^{2M} for large positive integers M formulated in Theorem 16.7. This estimate can be considered as the improvement of Bernstein's inequality in a most important special case.

13. The strategy of the proof for the main result of this paper

We have reduced the proof of the main result of this paper, the proof of Theorem 8.4 to that of Proposition 12.2$'$. It is a multivariate version of Proposition 6.2, and also its proof is based on similar ideas. In particular, a multivariate version of Lemma 7.2 will be proved, which means some kind of randomization. In this result we consider a class of decoupled, degenerate U-statistics $\bar I_{n,k}(f)$ together with a class of randomized, decoupled U-statistics $\bar I_{n,k}(f,\varepsilon)$ defined in formulas (10.5) and (10.6) respectively with the same countable class of functions $f\in\mathcal F$ and want to bound the probability $P\left(n^{-k/2}\sup\limits_{f\in\mathcal F}\bar I_{n,k}(f)|>u\right)$ with the help of a probability of the form $P\left(n^{-k/2}\sup\limits_{f\in\mathcal F}\bar I_{n,k}(f,\varepsilon)|>Bu\right)$ with some appropriate universal constant $B>0$.

To carry out such a program we introduce $2k$ independent copies $\xi_1^{(j)},\dots,\xi_n^{(j)}$ and $\bar\xi_1^{(j)},\dots,\bar\xi_n^{(j)}$, $1\le j\le n$, of the sequence of random variables $\xi_1,\dots,\xi_n$ we have at the start. We shall work with these $2k$ copies and a sequence of independent random variables $\varepsilon=(\varepsilon_1,\dots,\varepsilon_n)$, $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac12$, $1\le l\le n$, independent also of the sequences $\xi_1^{(j)},\dots,\xi_n^{(j)}$ and $\bar\xi_1^{(j)},\dots,\bar\xi_n^{(j)}$, $1\le j\le k$. Given some function $f(x_1,\dots,x_k)$ of k variables in the k-fold power $(X^k,\mathcal X^k)$ of some measurable space $(X,\mathcal X)$ let us consider the random sums $\tilde I_{n,k}(f)$ and $\tilde I_{n,}(f,\varepsilon)$ defined in formulas (11.9) and (11.9$'$) with the help of the above random sequences. We shall use Lemma 11.5 which states that given a class of functions of k-variables $f\in\mathcal F$, the joint distribution of the random variables $\tilde I_{n,k}(f)$ and $\tilde I_{n,}(f,\varepsilon)$ agree.

As we shall see later, Lemma 11.5 enables us to reduce the multivariate version of Lemma 7.2 we would like to prove to an appropriate bounding of the distribution of $\sup\limits_{f\in\mathcal F}\bar I_{n,k}(f)$ by that of $\sup\limits_{f\in\mathcal F}\tilde I_{n,k}(f)$. In the proof of Lemma 7.2 we met a simple special case of this problem, and it could be solved by means of the Symmetrization Lemma (Lemma 7.1). In the general case Lemma 7.1 is not sufficient for our purposes, since we have to work with not necessarily independent random variables. Hence we prove a generalized version of it.

Lemma 13.1 (Generalized version of the Symmetrization Lemma.) *Let Z_n and $\bar Z_n$, $n=1,2,\dots$, be two sequences of random variables on a probability space $(\Omega,\mathcal A,P)$. Let a σ-algebra $\mathcal B\subset\mathcal A$ be given on the probability space $(\Omega,\mathcal A,P)$ together with a $\mathcal B$-measurable set B and two numbers $\alpha>0$ and $\beta>0$ such that the random variables Z_n, $n=1,2,\dots$, are $\mathcal B$ measurable, and the inequality*

$$P(|\bar Z_n|\le\alpha|\mathcal B)(\omega)\ge\beta\quad\textit{for all } n=1,2,\dots\ \textit{if } \omega\in B \tag{13.1}$$

holds. Then

$$P\left(\sup_{1\le n<\infty}|Z_n|>\alpha+u\right)\le\frac1\beta P\left(\sup_{1\le n<\infty}|Z_n-\bar Z_n|>u\right)+(1-P(B))\quad\textit{for all } u>0. \tag{13.2}$$

Proof of Lemma 13.1. Put $\tau=\min\{n\colon |Z_n|>\alpha+u)$ if there exists such an n, and $\tau=0$ otherwise. Then

$$
\begin{aligned}
P(\{\tau=n\}\cap B)&\le\int_{\{\tau=n\}\cap B}\frac1\beta P(|\bar Z_n|\le\alpha|\mathcal B)\,dP=\frac1\beta P(\{\tau=n\}\cap\{|\bar Z_n|\le\alpha\}\cap B)\\
&\le\frac1\beta P(\{\tau=n\}\cap\{|Z_n-\bar Z_n|>u\})\quad\text{for all } n=1,2,\dots.
\end{aligned}
$$

Hence

$$
\begin{aligned}
P\left(\sup_{1\le n<\infty}|Z_n|>\alpha+u\right)-(1-P(B))&\le P\left(\left\{\sup_{1\le n<\infty}|Z_n|>\alpha+u\right\}\cap B\right)\\
&=\sum_{n=1}^\infty P(\{\tau=n\}\cap B)\le\frac1\beta\sum_{n=1}^\infty P(\{\tau=n\}\cap\{|Z_n-\bar Z_n|>u\})\\
&\le\frac1\beta P\left(\sup_{1\le n<\infty}|Z_n-\bar Z_n|>u\right).
\end{aligned}
$$

Thus Lemma 13.1 is proved.

The main difficulty we meet when we try to prove Proposition 12.2′ instead of its simpler version, Proposition 6.2 is that now we have to check an estimate of the form (13.1) with some appropriately chosen random variables Z_n, σ-algebra $\mathcal B$ and set B instead of the estimate (7.1) applied in the proof of Proposition 6.2. The cause of this difference is that now we have to work with not completely independent random variables. In the symmetrization argument needed in the proof of Proposition 6.2 we could simply check inequality (7.1) by calculating the variance of the random variables we were working with. On the other hand, to check inequality (13.1) in the symmetrization argument we want to apply in the present case we shall bound the conditional variance of certain random variables, and we can only state that this conditional variance is relatively small with great probability.

In the proof of Proposition 12.2′ we formulate and prove a multivariate version of the definition of *good tail behaviour for a class of normalized random sums*, where the normalized random sums are replaced by degenerate decoupled U-statistics. It is enough to prove the good tail-behaviour of decoupled U-statistics introduced below by means of an appropriate induction, and Proposition 12.2′ follows from it. But to carry out such a program we have to formulate and check another property which will be called the *good tail behaviour for a class of integrals of decoupled U-statistics.* This property helps us to carry out the induction procedure needed in the proof of Proposition 12.2′. Its introduction and proof corresponds to the symmetrization argument formulated in Lemma 7.2 in the proof of Proposition 6.2. The above mentioned two properties will be proved simultaneously. Before their formulation I make some comments about the idea behind the introduction of the property ‘good tail behaviour for a class of integrals of decoupled U-statistics’.

In the introduction of this property we consider a class of functions $f(x_1,\dots,x_k,y)$ depending on a parameter $y\in Y$, but in all further applications we shall apply this property with the choice $Y = X^l$ and $\rho = \mu^l$ (i.e. with the l-th power of the space X and the probability measure μ on it) with some integer l. The property 'good tail behaviour for a class of integrals of decoupled U-statistics' with the above choice will be useful for us for the following reason.

We shall consider the expression introduced in formula (11.9), which is some sort of linear combination of decoupled U-statistics, and want to bound the inner sums at the right-hand side of this expression. More explicitly, we consider those inner sum terms for which $m(v) = l \geq 1$, i.e. for which the original sample elements are replaced by their independent copies in $l \geq 1$ coordinates. We want to calculate the conditional variance of such sums under the condition that the values of the elements of the original sample are prescribed. The property of good tail behaviour for a class of integrals of decoupled U-statistics helps us in getting a good estimate for these expressions. If we want to bound the conditional variance of such an inner sum where the original sample elements are replaced in l coordinates, then the application of the property of good tail behaviour for a class of integrals of U-statistics with $k-l$ instead of k parameters and with the choice $(Y, \mathcal{Y}, \rho) = (X^l, \mathcal{X}^l, \mu^l)$ will be useful. By applying this property with such a choice together with the canonical property of the function $f(x_1, \dots, x_k)$ we shall work with we can prove the estimate we need.

Let me also remark that the estimate (13.5) we have imposed in the definition of the property of 'good tail behaviour for a class of integrals of U-statistics' is fairly natural. We have applied the natural normalization, and with such a normalization it is natural to expect that the distribution of $\sup\limits_{f \in \mathcal{F}} n^{-k} H_{n,k}(f)$ behaves similarly to that of const. $\left(\sigma\eta^2\right)^k$, where η is a standard normal random variables. Formula (13.5) expresses such a behaviour, only the power of the number A in the exponent at the right-hand side was chosen in a non-optimal way.

Naturally, we want to prove the property of good tail behaviour for a class of integrals of decoupled U-statistics under appropriate, not too restrictive conditions. Let me remark that in the conditions of Proposition 13.3 we want to prove we have imposed besides formula (13.6) a fairly weak condition (13.7). Most difficulties arise in the proof because we want to work with this condition. Here we did not demand that the L_2-norm of the functions $f(x_1, \dots, x_k, y)$ should be small for all parameters y. We only assumed that some average of these L_2-norms expressed in formula (13.7) are small. Now I formulate the definition of the properties we shall work with.

Definition of good tail behaviour for a class of decoupled U-statistics. *Let us have some measurable space $(X, \mathcal{X})$ and a probability measure μ on it. Let us consider some class $\mathcal{F}$ of functions $f(x_1, \dots, x_k)$ on the k-fold product $(X^k, \mathcal{X}^k)$ of the space $(X, \mathcal{X})$. Fix some positive integer $n \geq k$ and positive number $0 < \sigma \leq 1$, and take k independent copies $\xi_1^{(j)}, \dots, \xi_n^{(j)}$, $1 \leq j \leq k$, of a sequence of independent μ-distributed random variables $\xi_1, \dots, \xi_n$. Let us introduce with the help of these random variables the decoupled U-statistics $\bar{I}_{n,k}(f)$, $f \in \mathcal{F}$, defined in formula (10.5). Given some real number $T > 0$ we say that the set of decoupled U-statistics determined by the class of functions $\mathcal{F}$ has a good tail behaviour at level T (with parameters n and σ^2 which we* *fix in the sequel) if the inequality holds:*

$$P\left(\sup_{f\in\mathcal{F}} |n^{-k/2}\bar I_{n,k}(f)| \ge A n^{k/2}\sigma^{k+1}\right) \le \exp\left\{-A^{1/2k} n\sigma^2\right\} \quad \textit{for all } A > T. \qquad (13.3)$$

We shall also introduce the following property:

Definition of good tail behaviour for a class of integrals of decoupled U-statistics. *Let us have a product space $(X^k\times Y, \mathcal{X}^k\times\mathcal{Y})$ with some product measure $\mu^k\times\rho$, where $(X^k,\mathcal{X}^k,\mu^k)$ is the k-fold product of some probability space $(X,\mathcal{X},\mu)$, and $(Y,\mathcal{Y},\rho)$ is some other probability space. Fix some positive integer $n\ge k$ and positive number $\sigma>0$, and consider some class $\mathcal{F}$ of functions $f(x_1,\dots,x_k,y)$ on the product space $(X^k\times Y,\mathcal{X}^k\times\mathcal{Y},\mu^k\times\rho)$. Take k independent copies $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, of a sequence of independent, μ-distributed random variables $\xi_1,\dots,\xi_n$. For all $f\in\mathcal{F}$ and $y\in Y$ let us define the decoupled U-statistics $\bar I_{n,k}(f,y)=\bar I_{n,k}(f_y)$ by means of these random variables $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, the kernel function $f_y(x_1,\dots,x_k)=f(x_1,\dots,x_k,y)$ and formula (10.5). Define with the help of these U-statistics $\bar I_{n,k}(f,y)$ the random integrals*

$$H_{n,k}(f)=\int \bar I_{n,k}(f,y)^2\rho(\,dy), \quad f\in\mathcal{F}. \qquad (13.4)$$

Choose some real number $T>0$. We say that the set of random integrals $H_{n,k}(f)$, $f\in\mathcal{F}$, have a good tail behaviour at level T (with parameters n and σ^2 which we fix in the sequel) if

$$P\left(\sup_{f\in\mathcal{F}} n^{-k}H_{n,k}(f)\ge A^2n^k\sigma^{2k+2}\right)\le \exp\left\{-A^{1/(2k+1)}n\sigma^2\right\} \quad \textit{for } A>T. \qquad (13.5)$$

Now I formulate those two inductive statements in Propositions 13.2 and 13.3 which imply that the above introduced properties of good tail behaviour for a class of decoupled U-statistics and good tail behaviour for a class of integrals of decoupled U-statistics hold under fairly general conditions. Proposition 12.2$'$ can be obtained as a relatively simple consequence of these results.

Proposition 13.2. *Let us fix a positive integer $n\ge k$, a real number $0<\sigma\le 2^{-(k+1)}$ and a probability measure μ on a measurable space $(X,\mathcal{X})$ together with a countable L_2-dense class $\mathcal{F}$ of canonical kernel functions $f=f(x_1,\dots,x_k)$ (with respect to the measure μ) on the k-fold product space $(X^k,\mathcal{X}^k)$ which has exponent $L\ge1$ and parameter D. Let us also assume that all functions $f\in\mathcal{F}$ satisfy the conditions $\sup\limits_{x_j\in X, 1\le j\le k}|f(x_1,\dots,x_k)|\le 2^{-(k+1)}$, $\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\le\sigma^2$, and $n\sigma^2>K(L+\beta)\log n$ with an appropriately chosen fixed number $K=K(k)$ with $\beta=\max\left(\frac{\log D}{\log n},0\right)$.*

There exists some real number $A_0 = A_0(k) > 1$ such that if all classes of functions $\mathcal{F}$ satisfying the above conditions the sets of decoupled U-statistics determined $\bar{I}_{n,k}(f)$, $f \in \mathcal{F}$, have a good tail behaviour at level $T^{4/3}$ for some $T \ge A_0$, then they also have a good tail behaviour at level T.

Proposition 13.3. *Fix some positive integer $n \ge k$ and real number $0 < \sigma \le 2^{-(k+1)}$, and let us have a product space $(X^k \times Y, \mathcal{X}^k \times \mathcal{Y})$ with some product measure $\mu^k \times \rho$, where $(X^k, \mathcal{X}^k, \mu^k)$ is the k-fold product of some probability space $(X, \mathcal{X}, \mu)$, and $(Y, \mathcal{Y}, \rho)$ is some other probability space. Let us have a countable L_2-dense class $\mathcal{F}$ of canonical functions $f(x_1, \dots, x_k, y)$ on the product space $(X^k \times Y, \mathcal{X}^k \times \mathcal{Y}, \mu^k \times \rho)$ with some exponent $L \ge 1$ and parameter D. Let us also assume that the functions $f \in \mathcal{F}$ satisfy the conditions*

$$\sup_{x_j \in X, 1 \le j \le k, y \in Y} |f(x_1, \dots, x_k, y)| \le 2^{-(k+1)} \tag{13.6}$$

and

$$\int f^2(x_1, \dots, x_k, y) \mu(\,dx_1) \dots \mu(\,dx_k) \rho(\,dy) \le \sigma^2 \quad \text{for all } f \in \mathcal{F}. \tag{13.7}$$

Let the inequality $n\sigma^2 > K(L + \beta) \log n$ hold with a sufficiently large, appropriately chosen number $K = K(k)$ and $\beta = \max\left(\frac{\log D}{\log n}, 0\right)$.

There exists some number $A_0 = A_0(k) > 1$ such that if for all classes of functions $\mathcal{F}$ which satisfy the above conditions the random integrals $H_{n,k}(f)$, $f \in \mathcal{F}$, defined in (13.4) have a good tail behaviour at level $T^{(2k+1)/2k}$ with some $T \ge A_0$, then they also have a good tail behaviour at level T.

Remark: In the conditions of Proposition 13.3 the notion of canonical functions appeared in a slightly more general form than it was defined in formula (8.8). We say that a function $f(x_1, \dots, x_k, y)$ on the product space $(X^k \times Y, \mathcal{X}^k \times \mathcal{Y}, \mu^k \times \rho)$ is canonical if

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

and

$$\int f(x_1, \dots, x_k, y) \rho(\,dy) = 0 \quad \text{for all } x_j \in X, \; 1 \le j \le k.$$

It is not difficult to deduce Proposition 12.2′ from Proposition 13.2. Indeed, let us observe that the set of decoupled U-statistics determined by a class of functions $\mathcal{F}$ satisfying the conditions of Proposition 13.2 has a good tail-behaviour at level $T_0 = \sigma^{-(k+1)}$, since under the conditions of this Proposition the probability at the left-hand side of (13.3) equals zero for $A > \sigma^{-(k+1)}$. Then we get from Proposition 13.2 by induction with respect to the number j, that this set of decoupled U-statistics has a good tail-behaviour also for all $T \ge T_0^{(3/4)^j} = \sigma^{-(k+1)(3/4)^j}$ for $j = 0, 1, 2, \dots$ if $\sigma^{-(k+1)(3/4)^j} \ge A_0$. (Observe that $\sigma < 1$ under the conditions of Proposition 13.2, since $\sigma^2 \le 2^{-2(k+1)}$ in this case.) This implies that if a class of functions $\mathcal{F}$ satisfies the conditions of Proposition 13.2, then the set of decoupled U-statistics determined by this class of functions has a good tail-behaviour at level $T = A_0^{4/3}$, i.e. at a level which depends only on the order k of the decoupled U-statistics. This result implies Proposition 12.2′, only we have to apply it not directly for the class of functions $\mathcal{F}$ appearing in it, but these functions have to be multiplied by a sufficiently small positive number depending only on k.

Similarly to the above argument an inductive procedure yields a corollary of Proposition 13.3 formulated below. Actually, we shall need this corollary of Proposition 13.3.

Corollary of Proposition 13.3. *If the class of functions $\mathcal{F}$ satisfies the conditions of Proposition 13.3, then there exists a constant $\bar{A}_0 = \bar{A}_0(k) > 0$ depending only on k such that the class of integrals $H_{n,k}(f)$, $f \in \mathcal{F}$ defined in formula (13.4) have a good tail behaviour at level $\bar{A}_0$.*

The main difficulty in the proof of Proposition 13.2 appears as we try to apply the symmetrization procedure corresponding to Lemma 7.2 in the one-variate case. This difficulty can be overcome by means of Proposition 13.3, more precisely its corollary. It helps us to estimate the conditional variances of the decoupled U-statistics we have to handle in the proof of Proposition 13.2. The proof of Propositions 13.2 and 13.3 apply similar arguments, and they will be proved simultaneously. These results will be proved by means of the following inductive procedure. First Propositions 13.2 and then Proposition 13.3 are proved for $k = 1$. If Propositions 13.2 and 13.3 are already proved for all $k' < k$ for some number k, then first we prove Proposition 13.2 and then Proposition 13.3 for this number k.

14. A symmetrization argument

The proof of Propositions 13.2 and 13.3 apply similar ideas to the proof of Proposition 6.2, but here some additional technical difficulties have to be overcome. As a first step we prove two results formulated in Lemma 14.1A and 14.1B. They can be considered as a symmetrization argument analogous to Lemma 7.2 applied in the proof of Propositions 6.2. Lemma 14.1A will be needed in the proof of Proposition 13.2 and Lemma 14.1B in the proof of Proposition 13.3. This section contains the proof of these results.

Lemma 14.1A is a natural multivariate version of Lemma 7.2. In this result the probability we want to estimate in Proposition 13.2 is bounded by means of the distribution of the supremum of homogeneous polynomials of Rademacher functions of order k (the order of the decoupled U-statistic we investigate), and such an expression can be investigated similarly to the proof of Proposition 6.2 by means of the multi-dimensional version of Hoeffding's inequality given in Theorem 12.3. The case of Lemma 14.1B is more complicated. The probability we want to investigate in Proposition 13.3 will be bounded by the distribution of the supremum of some random variables $\bar{W}(f)$, $f \in \mathcal{F}$, which will be defined in formula (14.8). The expressions $\bar{W}(f)$ are squares of random polynomials of Rademacher functions. It is useful to study them more closely. This will be done in the proof of corollary of Lemma 14.1B which yields a more appropriate bound for the expression we want to estimate in Proposition 13.3. We shall apply this corollary in the sequel.

The proof of Lemmas 14.1A and 14.1B is similar to that of Lemma 7.2. First we introduce an independent copy $\bar\xi_n^{(j)},\dots,\bar\xi_n^{(j)}$ of the k sequences $\xi_n^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, and construct with their help some appropriate expressions which have the same distribution as the randomized sums we shall work with in the proof of Lemmas 14.1A and 14.1B. This statement will be formulated and proved in Lemmas 14.2A and 14.2B. These results enable us to reduce the problems we are interested in to some simpler questions which can be studied with the help of Lemmas 14.3A and 14.3B. In Lemma 14.3A the conditional variance of a random variable is estimated under some appropriate conditions. This estimate together with the generalized form of the Symmetrization Lemma enable us to prove Lemma 14.1A. Lemma 14.1B can be proved similarly, but here we need an estimate about the conditional distribution of a more complicated expression. This estimate can be proved with the help of Lemma 14.3B. In Lemma 14.3B the conditional expectation of the absolute value of an appropriate expression is bounded.

Now we formulate the main results of this section.

Lemma 14.1A. *Let $\mathcal{F}$ be a class of functions on the space $(X^k,\mathcal{X}^k)$ which satisfies the conditions of Proposition 13.2 with some probability measure μ. Let us have k independent copies $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, of a sequence of independent μ distributed random variables $\xi_1,\dots,\xi_n$, and a sequence of independent random variables $\varepsilon=(\varepsilon_1,\dots,\varepsilon_n)$, $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac12$, $1\le l\le n$, which is independent also of the random sequences $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$. Consider the decoupled U-statistics $\bar I_{n,k}(f)$, $f\in\mathcal{F}$, defined with the help of these random variables by formula (10.5) together with their randomized version*

$$\bar I^{\varepsilon}_{n,k}(f)=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'} \text{ if } j\ne j'}}\varepsilon_{l_1}\cdots\varepsilon_{l_k}f\left(\xi^{(1)}_{l_1},\dots,\xi^{(k)}_{l_k}\right),\qquad f\in\mathcal{F}. \tag{14.1}$$

Then there exists some constant $A_0=A_0(k)>0$ such that the inequality

$$P\left(\sup_{f\in\mathcal{F}}n^{-k/2}\left|\bar I_{n,k}(f)\right|>An^{k/2}\sigma^{k+1}\right)<2^{k+1}P\left(\sup_{f\in\mathcal{F}}\left|\bar I^{\varepsilon}_{n,k}(f)\right|>2^{-(k+1)}An^k\sigma^{k+1}\right)+2^kn^{k-1}e^{-A^{1/(2k-1)}n\sigma^2/k} \tag{14.2}$$

holds for all $A\ge A_0$.

To formulate Lemma 14.1B first we have to introduce some new quantities. We introduce them, because we want to adapt the symmetrization argument of Lemma 11.5 to the case when we work with a function $f(x_1,\dots,x_k,y)$ depending on a parameter y, and we have to introduce some new notions in this new situation. Some of the quantities introduced below will be used somewhat later. The quantities $\bar I^V_{n,k}(f,y)$ introduced in

(14.3) will depend on the sets $V\subset\{1,\dots,k\}$, and they are the natural adaptations of the inner sum terms in formula (14.9). Such expressions are needed when we want to formulate that version of the symmetrization result of Lemma 11.5 which is needed in the proof of Proposition 13.3. Their randomizations $\bar I_{n,k}^{(V,\varepsilon)}(f,y)$, introduced in formula (14.6), correspond to the inner sum terms in formula (11.9′). We also introduce the integrals of these expressions in formulas (14.4) and (14.7).

Let us consider a class $\mathcal F$ of functions $f(x_1,\dots,x_k,y)\in\mathcal F$ on a space $(X^k\times Y,\mathcal X^k\times\mathcal Y,\mu^k\times\rho)$ which satisfies the conditions of Proposition 13.3. Let us take $2k$ independent copies $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $\bar\xi_1^{(j)},\dots,\bar\xi_n^{(j)}$, $1\le j\le k$, of a sequence of independent μ distributed random variables $\xi_1,\dots,\xi_k$ together with a sequence of independent random variables $(\varepsilon_1,\dots,\varepsilon_n)$, $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac12$, $1\le l\le n$, which is also independent of the previous random sequences. Let us introduce the notation $\xi_l^{(j,1)}=\xi_l^{(j)}$ and $\xi_l^{(j,-1)}=\bar\xi_l^{(j)}$, $1\le l\le n$, $1\le j\le k$. For all subsets $V\subset\{1,\dots,k\}$ of the set $\{1,\dots,k\}$ let $|V|$ denote the cardinality of this set, and define for all functions $f(x_1,\dots,x_k,y)\in\mathcal F$ and $V\subset\{1,\dots,k\}$ the decoupled U-statistics

$$\bar I_{n,k}^V(f,y)=\frac1{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if }j\ne j'}} f\left(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)},y\right), \tag{14.3}$$

where $\delta_j=\pm1$, $1\le j\le k$, $\delta_j=1$ if $j\in V$, and $\delta_j=-1$ if $j\notin V$, together with the random variables

$$H_{n,k}^V(f)=\int \bar I_{n,k}^V(f,y)^2\rho(\,dy),\quad f\in\mathcal F. \tag{14.4}$$

Put

$$\bar I_{n,k}(f,y)=\bar I_{n,k}^{\{1,\dots,k\}}(f,y),\quad H_{n,k}(f)=H_{n,k}^{\{1,\dots,k\}}(f), \tag{14.5}$$

i.e. $\bar I_{n,k}(f,y)$ and $H_{n,k}(f)$ are the random variables $\bar I_{n,k}^V(f,y)$ and $H_{n,k}^V(f)$ with $V=\{1,\dots,k\}$ which means that these expressions are defined with the help of the random variables $\xi_l^{(j,1)}$, $1\le j\le k$, $1\le l\le n$.

Let us also define the 'randomized version' of the random variables $\bar I_{n,k}^V(f,y)$ and $H_{n,k}^V(f)$ as

$$\bar I_{n,k}^{(V,\varepsilon)}(f,y)=\frac1{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if }j\ne j'}} \varepsilon_{l_1}\cdots\varepsilon_{l_k} f\left(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)},y\right),\quad f\in\mathcal F, \tag{14.6}$$

and

$$H_{n,k}^{(V,\varepsilon)}(f)=\int \bar I_{n,k}^{(V,\varepsilon)}(f,y)^2\rho(\,dy),\quad f\in\mathcal F, \tag{14.7}$$

where $\delta_j=1$ if $j\in V$, and $\delta_j=-1$ if $j\in\{1,\dots,k\}\setminus V$.

Let us also introduce the random variables

$$\bar W(f)=\int\left[\sum_{V\subset\{1,\dots,k\}}(-1)^{|V|}\bar I_{n,k}^{(V,\varepsilon)}(f,y)\right]^2\rho(\,dy),\quad f\in\mathcal F \tag{14.8}$$

With the help of the above notations we can formulate Lemma 14.1B.

Lemma 14.1B. *Let $\mathcal{F}$ be a set of functions on $(X^k \times Y, \mathcal{X}^k \times \mathcal{Y})$ which satisfies the conditions of Proposition 13.3 with some probability measure $\mu^k \times \rho$. Let us have $2k$ independent copies $\xi_1^{j,\pm 1}, \dots, \xi_n^{j,\pm 1}$, $1 \le j \le k$, of a sequence of independent μ distributed random variables $\xi_1, \dots, \xi_n$ together with a sequence of independent random variables $\varepsilon_1, \dots, \varepsilon_n$, $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = \frac{1}{2}$, $1 \le j \le n$, which is independent also of the previously considered sequences.*

Then there exists some constant $A_0 = A_0(k) > 0$ such that if the integrals $H_{n,k}(f)$, $f \in \mathcal{F}$, determined by this class of functions $\mathcal{F}$ have a good tail behaviour at level $T^{(2k+1)/2k}$ for some $T \ge A_0$, (this property was defined in Section 13 in the definition of good tail behaviour for a class of integrals of decoupled U-statistics), then the inequality

$$P\left(\sup_{f\in\mathcal{F}} |H_{n,k}(f)| > A^2 n^{2k} \sigma^{2(k+1)}\right) < 2P\left(\sup_{f\in\mathcal{F}} \left|\bar{W}(f)\right| > \frac{A^2}{2} n^{2k} \sigma^{2(k+1)}\right) + 2^{2k+1} n^{k-1} e^{-A^{1/2k} n\sigma^2/k} \tag{14.9}$$

holds with the random variables $H_{n,k}(f)$ introduced in the second identity of relation (14.5) and with $\bar{W}(f)$ defined in formula(14.8) for all $A \ge T$.

We formulate a corollary of Lemma 14.1B which can be better applied than the original lemma. The inconvenience in Lemma 14.B arises, because at the right-hand side of formula (14.9) we have a probability depending on $\sup\limits_{f\in\mathcal{F}} |\bar{W}(f)|$, and $\bar{W}(f)$ is a too complicated expression. It equals the integral of the square of homogeneous polynomials of Rademacher functions (with random coefficients) depending on a parameter y with respect to this parameter. We have to understand better the structure of $\bar{W}(f)$. Hence we shall rewrite it by means of relations (14.10) and (14.11) in a somewhat complicated, but more explicit form. These formulas enable us to find such a corollary of Lemma 14.B which is more appropriate for us. To work out the details first we introduce some diagrams.

Let $\mathcal{G} = \mathcal{G}(k)$ denote the set of all diagrams consisting of two rows, such that each row is the set $\{1, \dots, k\}$, and the diagrams of $\mathcal{G}$ contain some edges $\{(j_1, j_1') \dots, (j_s, j_s')\}$, $0 \le s \le k$, connecting some point (vertex) of the first row with some point (vertex) of the second row. The vertices $j_1, \dots, j_s$ which are end points of some edge in the first row are all different, and the same relation holds also for the vertices $j_1', \dots, j_s'$ in the second row. Given some diagram $G \in \mathcal{G}$ let $e(G) = \{(j_1, j_1') \dots, (j_s, j_s')\}$ denote the set of its edges, and let $v_1(G) = \{j_1, \dots, j_s\}$ be the set of those vertices in the first row and $v_2(G) = \{j_1', \dots, j_s'\}$ the set of those vertices in the second row of the diagram G from which an edge of G starts.

Given some diagram $G \in \mathcal{G}$ and two sets $V_1, V_2 \subset \{1, \dots, k\}$, we define with the help of the random variables $\xi_1^{(j,1)}, \dots, \xi_n^{(j,1)}$, $\xi_1^{(j,-1)}, \dots, \xi_n^{(j,-1)}$, $1 \le j \le k$, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ taking part in the definition of the random variables $\bar{W}(f)$ the following random variables $H_{n,k}(f|G,V_1,V_2)$:

$$H_{n,k}(f|G,V_1,V_2)=\sum_{\substack{l_1,\dots,l_k,\,l'_1,\dots,l'_k\\ 1\le l_j\le n,\,l_j\ne l_{j'}\text{ if } j\ne j',\,1\le j,j'\le k,\\ 1\le l'_j\le n,\,l'_j\ne l'_{j'}\text{ if } j\ne j',\,1\le j,j'\le k,\\ l_j=l'_{j'}\text{ if }(j,j')\in e(G),\,l_j\ne l'_{j'}\text{ if }(j,j')\notin e(G)}}\prod_{j\in\{1,\dots,k\}\setminus v_1(G)}\varepsilon_{l_j}\prod_{j\in\{1,\dots,k\}\setminus v_2(G)}\varepsilon_{l'_j}$$
$$\frac{1}{k!^2}\int f(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)},y)f(\xi_{l'_1}^{(1,\bar\delta_1)},\dots,\xi_{l'_k}^{(k,\bar\delta_k)},y)\rho(\,dy) \tag{14.10}$$

where $\delta_j=1$ if $j\in V_1$, $\delta_j=-1$ if $j\notin V_1$, and $\bar\delta_j=1$ if $j\in V_2$, $\bar\delta_j=-1$ if $j\notin V_2$. (Let us observe that if the graph G contains s edges, then the product of the ε-s in (14.10) contains $2(k-s)$ terms and the number of terms in the sum (14.10) is of order n^{2k-s}.) As the Corollary of Proposition 14.1B will indicate in the proof of Proposition 13.3 the expression $H_{n,k}(f|G,V_1,V_2)$ has to be estimated. This can be done by means of Theorem 12.3, the multivariate version of Hoeffding's inequality. But the estimate we get in such a way has to be rewritten in a form appropriate for our inductive procedure. This will be done in the next section.

We shall prove that the identity

$$\bar W(f)=\sum_{G\in\mathcal G,\,V_1,V_2\subset\{1,\dots,k\}}(-1)^{|V_1|+|V_2|}H_{n,k}(f|G,V_1,V_2) \tag{14.11}$$

holds.

To prove this identity let us write first

$$\bar W(f)=\sum_{V_1,V_2\subset\{1,\dots,k\}}(-1)^{|V_1|+|V_2|}\int \bar I_{n,k}^{(V_1,\varepsilon)}(f,y)\bar I_{n,k}^{(V_2,\varepsilon)}(f,y)\rho(\,dy).$$

Then let us express the products $\bar I_{n,k}^{(V_1,\varepsilon)}(f,y)\bar I_{n,k}^{(V_2,\varepsilon)}(f,y)$ by means of formula (14.6). Let us rewrite this product as a sum of products of the form $\frac{1}{k!^2}\prod\limits_{j=1}^k\varepsilon_{l_j}f(\cdots)\prod\limits_{j=1}^k\varepsilon_{l'_j}f(\cdots)$ and let us define the following partition of the terms in this sum. The elements of this partition are indexed by the diagrams $G\in\mathcal G$, and if we take a diagram $G\in\mathcal G$ with the set of edges $e(G)=\{(j_1,j'_1),\dots,(j_s,j'_s)\}$, then the term of this sum determined by the indices $l_1,\dots,l_k,l'_1,\dots,l'_k$ belongs to the element of the partition indexed by this diagram G if and only if $l_{j_u}=l'_{j'_u}$ for all $1\le u\le s$, and no more numbers between the indices $l_1,\dots,l_k,l'_1\dots,l'_k$ may agree. Since $\varepsilon_{l_{j_u}}\varepsilon_{l'_{j'_u}}=1$ for all $1\le u\le s$ and all other ε_{l_j} and $\varepsilon_{l'_j}$ are different for a term of the sum in the element of the partition indexed by the diagram G we get by integrating the product $\bar I_{n,k}^{(V_1,\varepsilon)}(f,y)\bar I_{n,k}^{(V_2,\varepsilon)}(f,y)$ with respect to the measure ρ that

$$\int \bar I_{n,k}^{(V_1,\varepsilon)}(f,y)\bar I_{n,k}^{(V_2,\varepsilon)}(f,y)\rho(\,dy)=\sum_{G\in\mathcal G}H_{n,k}(f|G,V_1,V_2)$$

for all $V_1,V_2\in\{1,\dots,k\}$. The last two relations imply formula (14.11).

Since the number of terms in the sum of formula (14.11) is less than $2^{4k}k!$, this relation implies that Lemma 14.1B has the following corollary:

Corollary of Lemma 14.1B. *Let a set of functions $\mathcal{F}$ satisfy the conditions of Proposition 13.3. Then there exists some constant $A_0=A_0(k)>0$ such that if the integrals $H_{n,k}(f)$, $f\in\mathcal{F}$, determined by this class of functions $\mathcal{F}$ have a good tail behaviour at level $T^{(2k+1)/2k}$ for some $T\ge A_0$, then the inequality*

$$
\begin{aligned}
&P\left(\sup_{f\in\mathcal{F}}|H_{n,k}(f)|>A^2n^k\sigma^{2(k+1)}\right)\\
&\qquad\le 2\sum_{G\in\mathcal{G},\,V_1,V_2\subset\{1,\dots,k\}} P\left(\sup_{f\in\mathcal{F}}|H_{n,k}(f|G,V_1,V_2)|>\frac{A^2}{2^{4k+1}k!}n^k\sigma^{2(k+1)}\right)\\
&\qquad\quad+2^{2k+1}n^{k-1}e^{-A^{1/2k}n\sigma^2/k}
\end{aligned}
\tag{14.12}
$$

holds with the random variables $H_{n,k}(f)$ and $H_{n,k}(f|G,V_1,V_2)$ defined in formulas (14.5) and (14.10) for all $A\ge T$.

In the proof of Lemmas 14.1A and 14.1B we apply the result of the following Lemmas 14.2A and 14.2B.

Lemma 14.2A. *Let us take $2k$ independent copies*

$$\xi_1^{(j,1)},\dots,\xi_n^{(j,1)}\quad\text{and}\quad \xi_1^{(j,-1)},\dots,\xi_n^{(j,-1)},\quad 1\le j\le k,$$

of a sequence of independent μ distributed random variables $\xi_1,\dots,\xi_k$ together with a sequence of independent random variables $(\varepsilon_1,\dots,\varepsilon_n)$, $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac12$, $1\le l\le n$, which is also independent of the previous sequences.

Let $\mathcal{F}$ be a class of functions which satisfies the conditions of Proposition 13.2. Introduce with the help of the above random variables for all sets $V\subset\{1,\dots,k\}$ and functions $f\in\mathcal{F}$ the decoupled U-statistic

$$\bar I^V_{n,k}(f)=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if }j\ne j'}} f\left(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)}\right) \tag{14.13}$$

and its 'randomized version'

$$\bar I^{(V,\varepsilon)}_{n,k}(f)=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if }j\ne j'}} \varepsilon_{l_1}\cdots\varepsilon_{l_k} f\left(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)}\right),\quad f\in\mathcal{F}, \tag{14.13$'$}$$

where $\delta_j=\pm1$, and $\delta_j=1$ if $j\in V$, and $\delta_j=-1$ if $j\in\{1,\dots,k\}\setminus V$.

Then the sets of random variables

$$S(f)=\sum_{V\subset\{1,\dots,k\}}(-1)^{|V|}\bar I^V_{n,k}(f),\quad f\in\mathcal F \tag{14.14}$$

and

$$\bar S(f)=\sum_{V\subset\{1,\dots,k\}}(-1)^{|V|}\bar I^{(V,\varepsilon)}_{n,k}(f),\quad f\in\mathcal F \tag{14.14$'$}$$

have the same joint distribution.

Lemma 14.2B. *Let us take $2k$ independent copies*

$$\xi_1^{(j,1)},\dots,\xi_n^{(j,1)}\quad\textit{and}\quad \xi_1^{(j,-1)},\dots,\xi_n^{(j,-1)},\quad 1\le j\le k,$$

of a sequence of independent μ distributed random variables $\xi_1,\dots,\xi_k$ together with a sequence of independent random variables $(\varepsilon_1,\dots,\varepsilon_n)$, $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac12$, $1\le l\le n$, which is also independent of the previous sequences. Let $\mathcal F$ be a class of functions of k variables satisfying the conditions of Proposition 13.3. For all functions $f\in\mathcal F$ and $V\in\{1,\dots,k\}$ consider the decoupled U-statistics $\bar I^V_{n,k}(f,y)$ defined by formula (14.3) with the help of the random variables $\xi_1^{(j,1)},\dots,\xi_n^{(j,1)}$ and $\xi_1^{(j,-1)},\dots,\xi_n^{(j,-1)}$, and define with their help the random variables

$$W(f)=\int\left[\sum_{V\subset\{1,\dots,k\}}(-1)^{|V|}\bar I^V_{n,k}(f,y)\right]^2\rho(\,dy),\quad f\in\mathcal F. \tag{14.15}$$

Then the random vectors $\{W(f)\colon f\in\mathcal F\}$ defined in (14.15) and $\{\bar W(f)\colon f\in\mathcal F\}$ defined in (14.8) have the same distribution.

Proof of Lemmas 14.2A and 14.2B. Lemma 14.2A agrees actually with the already proved result Lemma 11.5, only the notation is different. The proof of Lemma 14.2B is also similar to the proof of 11.5. We can state that even the following stronger statement holds. For any ±1 sequence $(u_1,\dots,u_n)$ of length n the conditional distribution of the random field $\bar W(f)$, $f\in\mathcal F$, under the condition $(\varepsilon_1,\dots,\varepsilon_n)=(u_1,\dots,u_n)$ agrees with the distribution of the random field $W(f)$, $f\in\mathcal F$. To see this relation let us first observe that the conditional distribution of the field $\bar W(f)$ under this condition agrees with the distribution of the random field we get by replacing the random variables ε_l by u_l for all $1\le l\le n$ in formulas (14.6) and (14.8). Besides, we get by replacing the vectors $(\xi_l^{(j,1)},\xi_l^{(j,-1)})$ by $(\xi_l^{(j,-1)},\xi_l^{(j,1)})$ for those indices (j,l) for which $u(l)=-1$ (independently of the value of the parameter j) and not modifying these vectors with coordinates (l,j) such that $u(l)=1$ a measure preserving transformation of the distribution of the random vector consisting of the random variables $(\xi_l^{(j,1)},\xi_l^{(j,-1)})$, $1\le l\le n$, $1\le j\le k$, and this implies that also the distribution of the field $W(f)$, $f\in\mathcal F$, agrees with the distribution of the field we obtain by carrying out the above transformation in the elements of the field $W(f)$, $f\in\mathcal F$. These facts imply Lemma 14.2B.

Now we formulate and prove Lemma 14.3A.

Lemma 14.3A. *Let us consider a class of functions $\mathcal{F}$ satisfying the conditions of Proposition 13.2 with parameter k, and the random variables $\bar I^V_{n,k}(f)$, $f\in\mathcal{F}$, $V\subset\{1,\dots,k\}$, defined in formula (14.1). Let $\mathcal{B}=\mathcal{B}(\xi_1^{(j,1)},\dots,\xi_n^{(j,1)};\ 1\le j\le k)$ denote the σ-algebra generated by the random variables $\xi_1^{(j,1)},\dots,\xi_n^{(j,1)}$, $1\le j\le k$, i.e. by the random sequences with second coordinate 1 in their upper index taking part in the definition of the random variables $\bar I^V_{n,k}(f)$. For all $V\in\{1,\dots,k\}$, $V\neq\{1,\dots,k\}$, there exists a number $A_0=A_0(k)>0$ such that the inequality*

$$P\left(\sup_{f\in\mathcal{F}} E\left(\bar I^V_{n,k}(f)^2\middle|\,\mathcal{B}\right)>2^{-(3k+3)}A^2n^{2k}\sigma^{2k+2}\right)<n^{k-1}e^{-A^{1/(2k-1)}n\sigma^2/k}. \tag{14.16}$$

holds for all $A\ge A_0$.

Proof of Lemma 14.3A. Let us first consider the case $V=\emptyset$. In this case we can write $E\left(\bar I^{\emptyset}_{n,k}(f)^2\middle|\,\mathcal{B}\right)=E\left(\bar I^{\emptyset}_{n,k}(f)^2\right)\le\frac{n^k}{k!}\sigma^2\le n^{2k}\sigma^{2k+2}$ for all $f\in\mathcal{F}$. In the above calculation we have exploited that the functions $f\in\mathcal{F}$ are canonical, and this implies certain orthogonalities, and also the inequality $n\sigma^2\ge1$ holds. The above relation implies that for $V=\emptyset$ the probability at the left-hand side of (14.16) equals zero if the number A_0 is chosen sufficiently large, i.e. the inequality (14.16) holds in this case.

To avoid some complications in the notation let us first restrict our attention to sets of the form $V=\{1,\dots,u\}$ with some $1\le u<k$, and prove relation (14.16) for such sets. For this goal let us introduce the random variables

$$\bar I^V_{n,k}(f,l_{u+1},\dots,l_k)=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,u\\ l_j\neq l_{j'}\text{ if }j\neq j'}} f\left(\xi_{l_1}^{(1,1)},\dots,\xi_{l_u}^{(u,1)},\xi_{l_{u+1}}^{(u+1,-1)},\dots,\xi_{l_k}^{(k,-1)}\right)$$

for all $f\in\mathcal{F}$, i.e. we fix the last $k-u$ coordinates $\xi_{l_{u+1}}^{(u+1,-1)}$, … , $\xi_{l_k}^{(k,-1)}$ of the random variable $\bar I^V_{n,k}(f)$ and sum up with respect the first u coordinates. Then we can write

$$\begin{aligned}E\left(\bar I^V_{n,k}(f)^2\middle|\,\mathcal{B}\right)&=E\left(\left(\sum_{\substack{1\le l_j\le n\ j=u+1,\dots,k\\ l_j\neq l_{j'}\text{ if }j\neq j'}}\bar I^V_{n,k}(f,l_{u+1},\dots,l_k)\right)^2\middle|\,\mathcal{B}\right)\\&=\sum_{\substack{1\le l_j\le n\ j=u+1,\dots,k\\ l_j\neq l_{j'}\text{ if }j\neq j'}}E\left(\bar I^V_{n,k}(f,l_{u+1},\dots,l_k)^2\middle|\,\mathcal{B}\right).\end{aligned} \tag{14.17}$$

The last relation follows from the identity

$$E\left(\bar I^V_{n,k}(f,l_{u+1},\dots,l_k)\bar I^V_{n,k}(f,l'_{u+1},\dots,l'_k)\middle|\,\mathcal{B}\right)=0$$

if $(l_{u+1},\dots,l_k)\neq(l'_{u+1},\dots,l'_k)$, which relation holds, since f is a canonical function.

It follows from relation (14.17) that

$$\left\{\omega\colon \sup_{f\in\mathcal F} E\left(\bar I^V_{n,k}(f)^2\,\middle|\,\mathcal B\right)(\omega)>2^{-(3k+3)}A^2n^{2k}\sigma^{2k+2}\right\}$$
$$\subset \bigcup_{\substack{1\le l_j\le n\ j=u+1,\dots,k\\ l_j\neq l_{j'}\text{ if } j\neq j'}}\left\{\omega\colon \sup_{f\in\mathcal F} E\left(\bar I^V_{n,k}(f,l_{u+1},\dots,l_k)^2\,\middle|\,\mathcal B\right)(\omega)>\frac{A^2n^{2k}\sigma^{2k+2}}{2^{(3k+3)}n^{k-u}}\right\}. \tag{14.18}$$

The probability of the events in the union at the right-hand side of (14.18) can be estimated with the help of the corollary of Proposition 13.3 with parameter $u<k$ instead of k. (We may assume that Proposition 13.3 holds for $u<k$.) We claim that this corollary yields that

$$P\left(\sup_{f\in\mathcal F} E\left(\bar I^V_{n,k}(f,l_{u+1},\dots,l_k)^2\,\middle|\,\mathcal B\right)>\frac{A^2n^{k+u}\sigma^{2k+2}}{2^{(3k+3)}}\right)\le e^{-A^{1/(2u+1)}(n-u)\sigma^2}. \tag{14.19}$$

Indeed, introduce the space $(Y,\mathcal Y,\rho)=(X^{k-u},\mathcal X^{k-u},\mu^{k-u})$, the $k-u$-fold power of the measure space $(X,\mathcal X,\mu)$, and for the sake of simpler notations write $y=(x_{u+1},\dots,x_k)$ for a point $y\in Y$. Let us consider a class of functions $f\in\mathcal F$ which satisfies the conditions of Proposition 13.2 and let us prove for it relation (14.16). Let us introduce for this goal the class of those function $\bar{\mathcal F}$ on the space $(X^u\times Y,\mathcal X^u\times\mathcal Y,\mu^u\times\rho)$ which can be written in the form $\bar f(x_1,\dots,x_u,y)=f(x_1,\dots,x_k)$ with $y=(x_{u+1},\dots,x_k)$ and some function $f(x_1,\dots,x_k)\in\mathcal F$. The class of functions $\bar{\mathcal F}$ satisfies the conditions of Proposition 13.3 with parameter $u<k$, hence we may apply by our inductive hypothesis the Corollary of Proposition 13.3 for this class of functions. We shall apply this Corollary for decoupled U-statistics with sample size $n-u$ which is defined with the u independent sequences of independent μ-distributed random variables we define as $\xi_l^{(j,1)}$, $1\le j\le u$, $l\in\{1,\dots,n\}\setminus\{l_{u+1},\dots,l_k\}$ where the set of numbers $\{l_{u+1},\dots,l_k\}$ is the set of indices appearing in formula (14.19). With such a choice we get that

$$P\left(\sup_{\bar f\in\bar{\mathcal F}}(n-u)^{-u}H_{n-u,u}(\bar f)\ge A^2(n-u)^u\sigma^{2u+2}\right)\le e^{-A^{1/(2u+1)}(n-u)\sigma^2}\qquad\text{for } A>A_0(u), \tag{14.20}$$

where

$$H_{n-u,u}(\bar f)=\int I_{n,u}(\bar f,y)^2\rho(\,dy)=\frac{k!}{u!}E\left(\bar I^V_{n,k}(f,l_{u+1},\dots,l_k)^2|\mathcal B\right) \tag{14.21}$$

with the function $f\in\mathcal F$ for which the identity $\bar f(x_1,\dots,x_u,y)=f(x_1,\dots,x_k)$ holds with $y=(x_{u+1},\dots,x_k)$.

It is not difficult to deduce formula (14.19) from relations (14.20) and (14.21). It is enough to replace the level $\frac{A^2 n^{k+u}\sigma^{2k+2}}{2^{(3k+3)}}$ in the probability at the left-hand side of (14.19) by $A^2(n-u)^{2u}\sigma^{2u+2} < \frac{A^2 n^{k+u}\sigma^{2k+2}}{2^{(3k+3)}}$. The last inequality really holds if the constant K in the condition $n\sigma^2 > K\log n$ in Proposition 13.2 is chosen sufficiently large.

Relations (14.18) and (14.19) imply that

$$P\left(\sup_{f\in\mathcal{F}} E\left(\bar{I}_{n,k}^V(f)^2\big|\,\mathcal{B}\right)(\omega) > 2^{-(3k+3)}A^2 n^{2k}\sigma^{2k+2}\right) \le n^{k-u}e^{-A^{1/(2u+1)}(n-u)\sigma^2}.$$

Since $e^{-A^{1/(2u+1)}(n-u)\sigma^2} \le e^{-A^{1/(2k-1)}n\sigma^2/k}$ if $u \le k-1$ and $n \ge k$ inequality (14.16) holds for a set V of the form $V=\{1,\dots,u\}$, $1\le u<k$.

The case of a general set $V\subset\{1,\dots,k\}$, $1\le |V|<k$, can be handled similarly, only the notation becomes more complicated. Moreover, the case of general sets V can be reduced to the case of sets of form we have already considered. Indeed, given some set $V\subset\{1,\dots,k\}$, $1\le|V|<k$, let us define a new class of function $\mathcal{F}_V$ we get by applying a rearrangement of the indices of the arguments $x_1,\dots,x_k$ of the functions $f\in\mathcal{F}$ in such a way that the arguments indexed by the set V are the first $|V|$ arguments of the functions $f_V\in\mathcal{F}_V$, and put $\bar{V}=\{1,\dots,|V|\}$. Then the class of functions $\mathcal{F}_V$ also satisfies the condition of Proposition 13.2, and we can get relation (14.16) with the set V by applying it for the set of function $\mathcal{F}_V$ and set $\bar{V}$.

Now we prove Lemma 14.1A. It will be proved with the help of Lemma 14.2A, the generalized symmetrization lemma 13.1 and Lemma 14.3A.

Proof of Lemma 14.1A. We show with the help of the generalized symmetrization lemma, Lemma 13.1, and Lemma 14.3A that

$$P\left(\sup_{f\in\mathcal{F}} n^{-k/2}\left|\bar{I}_{n,k}(f)\right| > An^{k/2}\sigma^{k+1}\right) < 2P\left(\sup_{f\in\mathcal{F}}|S(f)| > \frac{A}{2}n^k\sigma^{k+1}\right) + 2^k n^{k-1}e^{-A^{1/(2k-1)}n\sigma^2/k} \tag{14.22}$$

with the function $S(f)$ defined in (14.14). To prove relation (14.22) introduce the random variables $Z(f)=I_{n,k}^{\{1,\dots,k\}}(f)$ and $\bar{Z}(f) = -\sum\limits_{V\subset\{1,\dots,k\},\,V\neq\{1,\dots,k\}}(-1)^{|V|}\bar{I}_{n,k}^V(f)$ for all $f\in\mathcal{F}$, the σ-algebra $\mathcal{B}$ considered in Lemma 14.3A and the set

$$B=\bigcap_{\substack{V\subset\{1,\dots,k\}\\ V\neq\{1,\dots,k\}}}\left\{\omega\colon\ \sup_{f\in\mathcal{F}} E\left(\bar{I}_{n,k}^V(f)^2\big|\,\mathcal{B}\right)(\omega)\le 2^{-(3k+3)}A^2n^{2k}\sigma^{2k+2}\right\}.$$

Observe that $S(f)=Z(f)-\bar{Z}(f)$, $f\in\mathcal{F}$, $B\in\mathcal{B}$, and by Lemma 14.3A the inequality $1-P(B)\le 2^k n^{k-1}e^{-A^{1/(2k-1)}n\sigma^2/k}$ holds. Hence to prove relation (14.22) it is enough to apply Lemma 13.1 and to show that

$$P\left(|\bar{Z}(f)| > \frac{A}{2}n^k\sigma^{k+1}|\mathcal{B}\right)(\omega)\le\frac{1}{2}\quad\text{for all } f\in\mathcal{F}\quad\text{if } \omega\in\mathcal{B}. \tag{14.23}$$

But $P\left(\bar I_{n,k}^{|V|}(f)|>2^{-(k+1)}An^k\sigma^{k+1}|\mathcal B\right)(\omega)\le 2^{-(k+1)}$ for all functions $f\in\mathcal F$ and sets $V\subset\{1,\dots,k\}$, $V\neq\{1,\dots,k\}$, if $\omega\in B$ by the 'conditional Chebishev inequality', hence relations (14.23) and (14.22) hold.

Lemma 14.1A follows from relation (14.22), Lemma 14.2A and the observation that the random vectors $\{\bar I_{n,k}^{(V,\varepsilon)}(f)\}$, $f\in\mathcal F$, defined in (14.13′) have the same distribution for all $V\in\{1,\dots,k\}$ as the random vector $\bar I_{n,k}^{\varepsilon}(f)$, defined in formula (14.1). Hence

$$P\left(\sup_{f\in\mathcal F}|S(f)|>\frac A2 n^k\sigma^{k+1}\right)\le 2^kP\left(\sup_{f\in\mathcal F}\left|\bar I_{n,k}^{\varepsilon}(f)\right|>2^{-(k+1)}An^k\sigma^{k+1}\right).$$

Lemma 14.1A is proved.

Lemma 14.1B will be proved with the help of the following version Lemma 14.3B of Lemma 14.3A.

Lemma 14.3B. *Let us consider a class of functions $\mathcal F$ satisfying the conditions of Proposition 13.3 and the random variables $\bar I_{n,k}^V(f,y)$, $f\in\mathcal F$, $V\subset\{1,\dots,k\}$, defined in formula (14.3). Let $\mathcal B=\mathcal B(\xi_1^{(j,1)},\dots,\xi_n^{(j,1)};\ 1\le j\le k)$ denote the σ-algebra generated by the random variables $\xi_1^{(j,1)},\dots,\xi_n^{(j,1)}$, $1\le j\le k$, i.e. by the random variables with second argument 1 in their upper index taking part in the definition of the random variables $\bar I_{n,k}^V(f,y)$ and $H_{n,k}^V(f)$ introduced in formulas (14.3) and (14.4).*

a) For all $V\in\{1,\dots,k\}$, $V\neq\{1,\dots,k\}$, there exists a number $A_0=A_0(k)>0$ such that the inequality

$$P\left(\sup_{f\in\mathcal F}E(H_{n,k}^V(f)|\mathcal B)>2^{-(4k+4)}A^{(2k-1)/k}n^{2k}\sigma^{2k+2}\right)<n^{k-1}e^{-A^{1/2k}n\sigma^2/k}. \tag{14.24}$$

holds for all $A\ge A_0$.

b) Given two subsets $V_1,V_1\subset\{1,\dots,k\}$ of the set $\{1,\dots,k\}$ define the integrals of random expressions

$$H_{n,k}^{(V_1,V_2)}(f)=\int|\bar I_{n,k}^{V_1}(f,y)\bar I_{n,k}^{V_2}(f,y)|\rho(\,dy),\quad f\in\mathcal F, \tag{14.25}$$

with the help of the functions $\bar I_{n,k}^V(f,y)$ defined in (14.3). If at least one of the sets V_1 and V_2 is not the set $\{1,\dots,k\}$, then there exists some number $A_0=A_0(k)>0$ such that if the integrals $H_{n,k}(f)$, $f\in\mathcal F$, determined by this class of functions $\mathcal F$ have a good tail behaviour at level $T^{(2k+1)/2k}$ for some $T\ge A_0$, then the inequality

$$P\left(\sup_{f\in\mathcal F}E(H_{n,k}^{(V_1,V_2)}(f)|\mathcal B)>2^{-(2k+2)}A^2n^{2k}\sigma^{2k+2}\right)<2n^{k-1}e^{-A^{1/2k}n\sigma^2/k} \tag{14.26}$$

holds for all $A \ge T$.

Proof of Lemma 14.3B. Part a) of Lemma 14.3B can be proved in almost the same way as Lemma 14.3A. Hence I only briefly explain the main step of the proof. In the case $V=\emptyset$ $E(H^V_{n,k}(f)|\mathcal{B})=E(H^V_{n,k}(f))$, hence it is enough to show that $E(H^V_{n,k}(f))\le \frac{n^k\sigma^2}{k!}\le\frac{n^{2k}\sigma^{2k+2}}{k!}$ for all $f\in\mathcal{F}$ under the conditions of Proposition 13.3. (Here we exploit in particular that the functions of the class $\mathcal{F}$ are canonical.) The case of a general set V, $V\ne\emptyset$ can be reduced to the case $V=\{1,\dots,u\}$ with some $1\le u<k$.

Given a set $V=\{1,\dots,u\}$ let us define the random variables

$$\bar I^V_{n,k}(f,l_{u+1},\dots,l_k,y)=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,u\\ l_j\ne l_{j'} \text{ if } j\ne j'}} f\left(\xi^{(1,1)}_{l_1},\dots,\xi^{(u,1)}_{l_u},\xi^{(u+1,-1)}_{l_{u+1}},\dots,\xi^{(k,-1)}_{l_k},y\right)$$

for all $f\in\mathcal{F}$. We can show by exploiting the canonical property of the functions $f\in\mathcal{F}$ that

$$E\left(\bar H^V_{n,k}(f)^2\big|\,\mathcal{B}\right)=\sum_{\substack{1\le l_j\le n\ j=u+1,\dots,k\\ l_j\ne l_{j'} \text{ if } j\ne j'}}\int E\left(\bar I^V_{n,k}(f,l_{u+1},\dots,l_k,y)^2\big|\,\mathcal{B}\right)\rho(\,dy),$$

and the proof of part a) of Lemma 14.3B can be reduced to the inequality

$$P\left(\sup_{f\in\mathcal{F}}\int E\left(\bar I^V_{n,k}(f,l_{u+1},\dots,l_k,y)^2\rho(\,dy)\big|\,\mathcal{B}\right)>\frac{A^{(2k-1)/k}n^{k+u}\sigma^{2k+2}}{2^{(4k+4)}}\right)\le e^{-A^{(2k-1)/2(2u+1)k}(n-u)\sigma^2}.$$

This inequality can be proved, similarly to relation (14.19) in the proof of Lemma 14.3A with the help of the Corollary of Proposition 13.3. Only here we have to work in the space $(X^u\times\bar Y,\mathcal{X}^u\times\bar{\mathcal{Y}},\mu^u\times\bar\rho)$ where $\bar Y=X^{k-u}\times Y$, $\bar{\mathcal{Y}}=\mathcal{X}^{k-u}\times\mathcal{Y}$, $\bar\rho=\mu^{k-u}\times\rho$ with the class of function $\mathcal{F}$ so that we identify a function $f(x_1,\dots,x_k,y)\in\mathcal{F}$ with $f(x_1,\dots,x_u,\bar y)=f(x_1,\dots,x_k,y)$ so that $\bar y=(x_{u+1},\dots,x_k,y)$. I omit the details.

Part b) of Lemma 14.3B will be proved with the help of Part a) and the inequality

$$\sup_{f\in\mathcal{F}}E(H^{(V_1,V_2)}_{n,k}(f)|\mathcal{B})\le\left(\sup_{f\in\mathcal{F}}E(H^{V_1}_{n,k}(f)|\mathcal{B})\right)^{1/2}\left(\sup_{f\in\mathcal{F}}E(H^{V_2}_{n,k}(f)|\mathcal{B})\right)^{1/2}$$

which follows from the Schwarz inequality applied for integrals with respect to conditional distributions. Let us assume that $V_1\ne\{1,\dots,k\}$. The last inequality implies that

$$\begin{aligned}P&\left(\sup_{f\in\mathcal{F}}E(H^{(V_1,V_2)}_{n,k}(f)|\mathcal{B})>2^{-(2k+2)}A^2n^{2k}\sigma^{2k+2}\right)\\
&\le P\left(\sup_{f\in\mathcal{F}}E(H^{V_1}_{n,k}(f)|\mathcal{B})>2^{-(4k+4)}A^{(2k-1)/k}n^{2k}\sigma^{2k+2}\right)\\
&\qquad+P\left(\sup_{f\in\mathcal{F}}E(H^{V_2}_{n,k}(f)|\mathcal{B})>A^{(2k+1)/k}n^{2k}\sigma^{2k+2}\right)\end{aligned}$$

Hence the estimate (14.24) together with the inequality

$$P\left(\sup_{f\in\mathcal{F}} E(H^{V_2}_{n,k}(f)|\mathcal{B}) > A^{(2k+1)/k}n^{2k}\sigma^{2k+2}\right) \le n^{k-1}e^{A^{1/2k}n\sigma^2} \tag{14.27}$$

imply relation (14.26). Relation 14.27 follows from Part a) of Lemma 14.3B if $V_2 \neq \{1,\dots,n\}$ and $A \ge A_0$ with a sufficiently large number A_0 (in this case the level $A^{(2k+1)/k}n^{2k}\sigma^{2k+2}$ can be replaced by the larger number $2^{-(4k+2)}A^{(2k-1)/k}n^{2k}\sigma^{2k+2}$ in the probability of formula (14.27)) and from the conditions of Part b) of Lemma 14.3B if $V_2 = \{1,\dots,k\}$. Indeed, in this case we may apply the estimate (13.5) for this probability, since $A^{(2k+1)/2k} \ge T^{(2k+1)/2k}$, and this implies relation (14.27).

Now we turn to the proof of Lemma 14.1B.

Proof of Lemma 14.1B. By Lemma 14.2B it is enough to prove that relation (14.9) holds if the random variables $\bar W(f)$ are replaced in it by the random variables $W(f)$ defined in formula (14.15). We shall prove this by applying the generalized form of the symmetrization lemma, Lemma 13.1 with the choice of $Z(f) = H^{(\bar V,\bar V)}_{n,k}(f)$, $\bar V = \{1,\dots,k\}$, $\bar Z(f) = W(f) - Z(f)$, $f \in \mathcal{F}$, $\mathcal{B} = \mathcal{B}(\xi_1^{(j,1)},\dots,\xi_n^{(j,1)};\ 1 \le j \le k)$, and the set

$$B = \bigcap_{\substack{(V_1,V_2)\colon\ V_j\in\{1,\dots,k\},\ j=1,2\\ V_1\neq\{1,\dots,k\}\text{ or }V_2\neq\{1,\dots,k\}}} \left\{\omega\colon\ \sup_{f\in\mathcal{F}} E(H^{(V_1,V_2)}_{n,k}(f)|\mathcal{B})(\omega) \le 2^{-(2k+2)}A^2n^{2k}\sigma^{2k+2}\right\}.$$

By Lemma 14.3B the inequality $1-P(B) \le 2^{k+1}n^{k-1}e^{A^{1/2k}n\sigma^2/k}$ holds, and to prove Lemma 14.1B with the help of Lemma 13.1 it is enough to show that

$$P\left(|\bar Z(f)| > \frac{A^2}{2}n^{2k}\sigma^{2(k+1)}\,\middle|\,\mathcal{B}\right)(\omega) \le \frac12 \quad \text{for all } f\in\mathcal{F} \text{ if } \omega\in B.$$

To prove this relation observe that

$$E(|\bar Z(f)||\mathcal{B}) \le \sum_{\substack{(V_1,V_2)\colon\ V_j\in\{1,\dots,k\},\ j=1,2\\ V_1\neq\{1,\dots,k\}\text{ or }V_2\neq\{1,\dots,k\}}} E(H^{(V_1,V_2)}_{n,k}(f)|\mathcal{B}) \le \frac{A^2}{4}n^{2k}\sigma^{2k+2} \quad \text{if } \omega\in B$$

for all $f \in \mathcal{F}$. Hence the 'conditional Markov inequality' implies that

$$P\left(|\bar Z(f)| > \frac{A^2}{2}n^{2k}\sigma^{2k+2}\,\middle|\,\mathcal{B}\right) \le \frac12 \quad \text{if } \omega\in B \quad \text{and } f\in\mathcal{F}.$$

Lemma 14.1B is proved.

15. The proof of the main result

In this section we prove Proposition 13.2 and thus complete the proof of the main result of this work, of Theorem 8.4 or of its equivalent version Theorem 8.2. Proposition 13.2 will be proved with the help of the symmetrization Lemma 14.1A. In the proof of this symmetrization lemma we have also applied the corollary of Proposition 13.3 (for orders $u<k$ if we want to prove Proposition 13.2 for decoupled U-statistics of order k.) Hence to complete the proof of Proposition 13.2 we also have to prove Proposition 13.3. This section contains the proof of both results. First we prove Proposition 13.2.

A.) The proof of Proposition 13.2.

The proof of Theorem 13.2 is similar to the proof of Proposition 6.2. It applies an induction procedure with respect to the parameter k. In the proof of Proposition 13.2 for parameter k we may assume that Propositions 13.2 and 13.3 hold for $u<k$. In the proof we want to give a good estimate on the probability $P\left(\sup\limits_{f\in\mathcal{F}}\left|\bar I^{\varepsilon}_{n,k}(f)\right|>2^{-(k+1)}An^k\sigma^{k+1}\right)$ appearing in the estimate (14.2) of Lemma 14.1A. To estimate this probability we introduce (using the notation of Proposition 13.2) the functions

$$S^2_{n,k}(f)(x^{(j)}_l,\,1\le l\le n,\,1\le j\le k)=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k,\\ l_j\ne l_{j'}\text{ if } j\ne j'}} f^2\left(x^{(1)}_{l_1},\dots,x^{(k)}_{l_k}\right),\quad f\in\mathcal{F}, \tag{15.1}$$

with $x^{(j)}_l\in X$, $1\le l\le n$, $1\le j\le k$. Then we estimate the probability we are interested in with the help of this quantity similarly to the argument applied in the solution of the corresponding problem in the proof of Proposition 6.2.

Fix some number $A>T$ and define the set $H\subset X^{kn}$

$$\begin{aligned}H=H(A)=\Big\{&(x^{(j)}_l,\,1\le l\le n,\,1\le j\le k)\colon\\ &\sup_{f\in\mathcal{F}}S^2_{n,k}(f)(x^{(j)}_l,\,1\le l\le n,\,1\le j\le k)>2^kA^{4/3}n^k\sigma^2\Big\}.\end{aligned} \tag{15.2}$$

We want to show that

$$P(\{\omega\colon(\xi^{(j)}_l(\omega),\,1\le j\le n,\,1\le j\le k)\in H\})\le 2^ke^{-A^{2/3k}n\sigma^2}\quad\text{if } A\ge T. \tag{15.3}$$

Relation (15.3) will be proved by means of the Hoeffding decomposition (Theorem 9.1) of the U-statistics with kernel functions $f^2(x_1,\dots,x_k)$, $f\in\mathcal{F}$, and by the estimation of the sum this decomposition yields. More explicitly, write (applying formula (9.2) in Theorem 9.1)

$$f^2(x_1,\dots,x_k)=\sum_{V\subset\{1,\dots,k\}}f_V(x_j,j\in V) \tag{15.4}$$

with $f_V(x_j, j\in V)=\prod\limits_{j\notin V} P_j \prod\limits_{j\in V} Q_j f^2(x_1,\dots,x_k)$, where P_j is the projection defined in formula (9.1) and $Q_j=I-P_j$ is also the same operator as the operator Q_j in formula (9.2).

The functions f_V appearing in formula (15.4) are canonical (with respect to the measure μ), and the identity $S^2_{n,k}(f)(\xi_l^{(j)}\ 1\le l\le n, 1\le j\le k)=\bar I_{n,k}(f^2)$ holds for all $f\in\mathcal F$. By applying the Hoeffding decomposition (15.4) for each term $f^2(\xi_{l_1}^{(1)}\dots,\xi_{l_k}^{(k)})$ in the expression $I_{n,k}(f^2)$ we get that

$$
\begin{aligned}
&P\left(\sup_{f\in\mathcal F} S^2_{n,k}(f)(\xi_l^{(j)},\ 1\le l\le n,\ 1\le j\le k)>2^kA^{4/3}n^k\sigma^2\right)\\
&\qquad\le \sum_{V\subset\{1,\dots,k\}} P\left(\frac{|V|!}{k!}\sup_{f\in\mathcal F} n^{k-|V|}|\bar I_{n,|V|}(f_V)|>A^{4/3}n^k\sigma^2\right)
\end{aligned}
\tag{15.5}
$$

with the functions f_V in (15.4). We want to give a good estimate for all terms in the sum at the right-hand side in (15.5). For this goal first we show that the classes of functions $\{f_V\colon f\in\mathcal F\}$ satisfy the conditions of Proposition 13.2 for all $V\subset\{1,\dots,k\}$.

The functions f_V are canonical for all $V\subset\{1,\dots,k\}$. It follows from the conditions of Proposition 13.2 that $|f^2(x_1,\dots,x_k)|\le 2^{-2(k+1)}$ and

$$\int f^4(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\le 2^{-(k+1)}\sigma^2.$$

Hence relations (9.4) and (9.4′) of Theorem 9.2 imply that $\left|\sup\limits_{x_j\in X, j\in V} f_V(x_j, j\in V)\right|\le 2^{-(k+2)}\le 2^{-(k+1)}$ for all $V\subset\{1,\dots,k\}$ and $\int f_V^2(x_j,j\in V)\prod\limits_{j\in V}\mu(\,dx_j)\le 2^{-(k+1)}\sigma^2\le\sigma^2$ for all $V\subset\{1,\dots,k\}$. Finally, to check that the class of functions $\mathcal F_V=\{f_V\colon f\in\mathcal F\}$ is L_2-dense with exponent L and parameter D observe that for all probability measures ρ on $(X^k,\mathcal X^k)$ and pairs of functions $f,g\in\mathcal F$ $\int(f^2-g^2)^2\,d\rho\le 2^{-2k}\int(f-g)^2\,d\rho$. This implies that if $\{f_1,\dots,f_m\}$, $m\le D\varepsilon^{-L}$, is an ε-dense subset of $\mathcal F$ in the space $L_2(X^k,\mathcal X^k,\rho)$, then the set of functions $\{2^kf_1^2,\dots,2^kf_m^2\}$ is an ε-dense subset of the class of functions $\mathcal F'=\{2^kf^2\colon f\in\mathcal F\}$, hence $\mathcal F'$ is also an L_2-dense class of functions with exponent L and parameter D. Then by Theorem 9.2 the class of functions $\mathcal F_V$ is also L_2-dense with exponent L and parameter D for all sets $V\subset\{1,\dots,k\}$.

For $V=\emptyset$, the function f_V is constant, $f_V=\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\le\sigma^2$ holds, and $I_{|V|}(f_{|V|})|=f_V\le\sigma^2$. Therefore the term corresponding to $V=\emptyset$ in the sum at the right-hand side of (13.5) equals zero if $A_0\ge1$ in the conditions of Proposition 13.2. I claim that the terms corresponding to sets V, $1\le|V|\le k$, in these sums satisfy the inequality

$$P\left(\sup_{f\in\mathcal F}|\bar I_{n,|V|}(f_V)|>A^{4/3}n^{|V|}\sigma^2\right)$$

$$\le P\left(\sup_{f\in\mathcal F}|\bar I_{n,|V|}(f_V)|>A^{4/3}\frac{k!}{|V|!}n^{|V|}\sigma^{|V|+1}\right)\le e^{-A^{2/3k}n\sigma^2}\quad\text{if } 1\le|V|\le k. \tag{15.6}$$

The first inequality in (15.6) holds, since $\sigma^{|V|+1}\le\sigma^2$ for $|V|\ge1$, the second inequality follows from the inductive hypothesis if $|V|<k$, since it yields the upper bound $e^{-(A^{4/3}k!/|V|!)^{1/2|V|}n\sigma^2}\le e^{-A^{2/3k}n\sigma^2}$ if $A_0=A_0(k)$ in Proposition 13.2 is sufficiently large, and in the case $V=\{1,\dots,k\}$ it follows from the inequality $A\ge T$ and the assumption that U-statistics determined by a class of functions satisfying the conditions of Proposition 13.2 have a good tail behaviour at level $T^{4/3}$. Relations (15.5) and (15.6) together with the estimate in the case $V=\emptyset$ imply formula (15.3).

By conditioning the probability $P\left(\left|\bar I^\varepsilon_{n,k}(f)\right|>2^{-(k+2)}An^{k/2}\sigma^{k+1}\right)$ with respect to the random variables $\xi_l^{(j)}$, $1\le l\le n$, $1\le j\le k$ we get with the help of the multivariate version of Hoeffding's inequality (Theorem 12.3) that

$$\begin{aligned}
&P\left(\left|\bar I^\varepsilon_{n,k}(f)\right|>2^{-(k+2)}An^k\sigma^{k+1}\,\middle|\,\xi_l^{(j)}(\omega)=x_l^{(j)},1\le l\le n,1\le j\le k\right)\\
&\qquad\le C\exp\left\{-B\left(\frac{A^2n^{2k}\sigma^{2(k+1)}}{2^{2k+4}S^2_{n,k}(x_l^{(j)},1\le l\le n,\,1\le j\le k)/k!}\right)^{1/k}\right\}\\
&\qquad\le Ce^{-2^{-3-4/k}BA^{2/3k}k!n\sigma^2}\quad\text{for all } f\in\mathcal F\quad\text{if } (x_l^{(j)},\,1\le l\le n,\,1\le j\le k)\notin H.
\end{aligned}\tag{15.7}$$

Given some points $x_l^{(j)}\in X$, $1\le l\le n$, $1\le j\le k$, define the probability measures $\rho_j=\rho_{j,(x_l^{(j)},1\le l\le n)}$, $1\le j\le k$, uniformly distributed on the set $x_l^{(j)}$, $1\le l\le n$, i.e. let $\rho_j(x_l^{(j)})=\frac1n$, $1\le l\le n$. Let us also define the product $\rho=\rho_1\times\cdots\times\rho_k$ of these measures. If f is a function on $(X^k,\mathcal X^k)$ such that $\int f^2\,d\rho\le\delta^2$ with some $\delta>0$, then

$$|f(x_{l_j}^{(j)},\,1\le j\le k)|\le\delta n^{k/2}\quad\text{for all vectors } (l_1,\dots,l_k),\ 1\le l_j\le n,\ 1\le j\le k,$$

and this implies that $P\left(\left|\bar I^\varepsilon_{n,k}(f)\right|>\delta n^{3k/2}\,\middle|\,\xi_l^{(j)}=x_l^{(j)},\,1\le l\le n,\,1\le j\le k\right)=0$ for such a function f. Take the numbers $\bar\delta=An^{-k/2}2^{-(k+2)}\sigma^{k+1}$ and $\delta=2^{-(k+2)}n^{-k-1/2}\le\bar\delta$. (The inequality $\delta\le\bar\delta$ holds, since $A\ge A_0\ge1$, and $\sigma\ge n^{-1/2}$.) Choose a δ-dense set $\{f_1,\dots,f_m\}$ in the $L_2(X^k,\mathcal X^k,\rho)$ space with $m\le D\delta^{-L}\le2^{(k+2)L}n^{\beta+(k+1)L/2}$ elements. Then the above estimates, the δ-dense property of the set of functions $\{f_1,\dots,f_m\}$ in $L_2(X^k,\mathcal X^k,\rho)$ and formula (15.7) imply that

$$\begin{aligned}
&P\left(\sup_{f\in\mathcal F}\left|\bar I^\varepsilon_{n,k}(f)\right|>2^{-(k+1)}An^k\sigma^{k+1}\,\middle|\,\xi_l^{(j)}(\omega)=x_l^{(j)},1\le l\le n,1\le j\le k\right)\\
&\quad\le\sum_{s=1}^m P\left(\left|\bar I^\varepsilon_{n,k}(f_s)\right|>2^{-(k+2)}An^k\sigma^{k+1}\,\middle|\,\xi_l^{(j)}(\omega)=x_l^{(j)},1\le l\le n,1\le j\le k\right)\\
&\quad\le C2^{(k+2)L}n^{\beta+(k+1)L/2}e^{-2^{-3-4/k}BA^{2/3k}nk!\sigma^2}\quad\text{if } \{x_l^{(j)},\,1\le l\le n,\,1\le j\le k\}\notin H.
\end{aligned}\tag{15.8}$$

Relations (15.3) and (15.8) imply that

$$P\left(\sup_{f\in\mathcal{F}}\left|\bar I^{\varepsilon}_{n,k}(f)\right| > 2^{-(k+1)}An^k\sigma^{k+1}\right) \le C2^{(k+2)L}n^{\beta+(k+1)L/2}e^{-2^{-3-4/k}BA^{2/3k}n\sigma^2}+2^ke^{-A^{2/3k}k!n\sigma^2} \quad \text{if } A\ge T. \tag{15.9}$$

Proposition 13.2 follows from the estimates (14.2) and (15.9) if the constant A_0 together with the constant K in the condition $n\sigma^2\ge K(L+\beta)\log n$ are chosen sufficiently large. In this case these estimates yield an upper bound less than $e^{-A^{1/2k}n\sigma^2}$ for the probability at the left-hand side of (13.3).

Now we turn to the proof of Proposition 13.3.

B.) The proof of Proposition 13.3.

Because of formula (14.12) in the corollary of Lemma 14.1B to prove Proposition 13.3 i.e. inequality (13.5) it is enough to choose the parameter A_0 in Proposition 13.3 for which $A>T\ge A_0$ sufficiently large and to show that

$$P\left(\sup_{f\in\mathcal{F}}|H_{n,k}(f|G,V_1,V_2)|>\frac{A^2}{2^{4k+1}k!}n^{2k}\sigma^{2(k+1)}\right)\le 2^{k+1}e^{-A^{1/2k}n\sigma^2} \quad \text{for all } G\in\mathcal{G} \quad \text{and} \quad V_1,V_2\in\{1,\dots,k\} \quad \text{if } A\ge A_0 \tag{15.10}$$

with the random variables $H_{n,k}(f|G,V_1,V_2)$ defined in formula (14.10). Let us first prove formula (15.10) in the case when $|e(G)|=k$, i.e. when all vertices of the diagram G are end-points of some edge, and the expression $H_{n,k}(f|G,V_1,V_2)$ contains no 'symmetryzing term' ε_j. In this case we apply a special argument to prove relation (15.10).

If G is such a diagram for which $|e(G)|=k$, then we can show by means of the Schwarz inequality that

$$|H_{n,k}(f|G,V_1,V_2)|\le \frac{1}{k!}\left(\sum_{\substack{l_1,\dots,l_k,\,1\le l_j\le n,\\ l_j\ne l_{j'} \text{ if } j\ne j'}}\int f^2(\xi^{(1),\delta_1)}_{l_1},\dots,\xi^{(k,\delta_k)}_{l_k},y)\rho(\,dy)\right)^{1/2} \frac{1}{k!}\left(\sum_{\substack{l_1,\dots,l_k,1\le l_j\le n,\\ l_j\ne l_{j'} \text{ if } j\ne j'}}\int f^2(\xi^{(1,\bar\delta_1)}_{l_1},\dots,\xi^{(k,\bar\delta_k)}_{l_k},y)\rho(\,dy)\right)^{1/2}, \tag{15.11}$$

where $\delta_j=1$ if $j\in V_1$, $\delta_j=-1$ if $j\notin V_1$, and $\bar\delta_j=1$ if $j\in V_2$, $\bar\delta_j=-1$ if $j\notin V_2$. Indeed, in this case the sum of integrals in (14.10) can be rewritten in a natural way as the integral of the product of two functions on the product space $(I_n^k\times Y,\mathcal{I}_n^k\times\mathcal{Y},\lambda_n^k\times\rho)$, where $I_n=\{1,\dots,n\}$, $\mathcal{I}_n$ is the σ-algebra of all subsets of I_n, and λ_n is the counting measure on $\mathcal{I}_n$. Then the Schwarz inequality for this product yields formula (15.11). (Observe that the coordinates $l_1,\dots,l_k$ determine the coordinates $l'_1,\dots,l'_k$ in the summation (14.10) if $|e(G)|=k$.)

By formula (15.11)

$$
\begin{aligned}
&\left\{\omega\colon \sup_{f\in\mathcal{F}}|H_{n,k}(f|G,V_1,V_2)(\omega)|>\frac{A^2}{2^{4k+1}k!}n^{2k}\sigma^{2(k+1)}\right\}\\
&\quad\subset\left\{\omega\colon \sup_{f\in\mathcal{F}}\sum_{\substack{l_1,\dots,l_k,1\le l_j\le n,\\ l_j\ne l_{j'} \text{ if } j\ne j'}}\int f^2(\xi^{(1,\delta_1)}_{l_1}(\omega),\dots,\xi^{(k,\delta_k)}_{l_k}(\omega),y)\rho(\,dy)>\frac{A^2n^{2k}\sigma^{2(k+1)}k!}{2^{4k+1}}\right\}\\
&\quad\cup\left\{\omega\colon \sup_{f\in\mathcal{F}}\sum_{\substack{l_1,\dots,l_k,1\le l_j\le n,\\ l_j\ne l_{j'} \text{ if } j\ne j'}}\int f^2(\xi^{(1,\bar\delta_1)}_{l_1}(\omega),\dots,\xi^{(k,\bar\delta_k)}_{l_k}(\omega),y)\rho(\,dy)>\frac{A^2n^{2k}\sigma^{2(k+1)}k!}{2^{4k+1}}\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
&P\left(\sup_{f\in\mathcal{F}}|H_{n,k}(f|G,V_1,V_2)|>\frac{A^2}{2^{4k+1}k!}n^{2k}\sigma^{2(k+1)}\right)\\
&\qquad\le 2P\left(\sup_{f\in\mathcal{F}}\frac{1}{k!}\sum_{\substack{l_1,\dots,l_k,1\le l_j\le n,\\ l_j\ne l_{j'} \text{ if } j\ne j'}}h_f(\xi^{(1,1)}_{l_1},\dots,\xi^{(k,1)}_{l_k})>\frac{A^2n^{2k}\sigma^{2(k+1)}}{2^{4k+1}}\right)
\end{aligned}
\tag{15.12}
$$

with the functions $h_f(x_1,\dots,x_k)=\int f^2(x_1,\dots,x_k,y)\rho(\,dy)$, $f\in\mathcal{F}$. (In this upper bound we could get rid of the terms δ_j and $\bar\delta_j$, i.e. on the dependence of the expression $H_{n,k}(f|G,V_1,V_2)$ on the sets V_1 and V_2, since the probability of the events in the previous formula do not depend on these terms.)

I claim that

$$
P\left(\sup_{f\in\mathcal{F}}|\bar I_{n,k}(h_f)|\ge An^k\sigma^2\right)\le 2^k e^{-A^{1/2k}n\sigma^2}\quad\text{for } A\ge A_0 \tag{15.13}
$$

if the constants A_0 and K are chosen sufficiently large in Proposition 13.3. Relation (15.13) together with the relation $A\frac{n^{2k}\sigma^{2(k+1)}}{2^{4k+1}}\ge n^k\sigma^2$ imply that the probability at the right-hand side of (15.12) can be bounded by $2^{k+1}e^{-A^{1/2k}n\sigma^2}$, and the estimate (15.10) holds in the case $|e(G)|=k$.

Relation (15.13) is similar to relation (15.5), and the proof of the latter formula helps to carry out the proof in the present case. Indeed, it follows from the conditions of Proposition 13.3 that $0\le\int h_f(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\le\sigma^2$, and it is not difficult to check thatsup $|h_f(x_1,\dots,x_k)|\le 2^{-2(k+1)}$, and the class of functions $\mathcal{H}=\{2^k h_f,\ f\in\mathcal{F}\}$ is an L_2-dense class with exponent L and parameter D. This means that by applying the Hoeffding decomposition of the functions h_f, $f\in\mathcal{F}$, similarly to formula (15.4) we get such sets of functions $(h_f)_V$, $f\in\mathcal{F}$, for all $V\subset\{1,\dots,k\}$ which satisfy the conditions of Proposition 13.2. Hence a natural adaptation of the estimate given for the expression at the right-hand side of (15.5) yields the proof of formula (15.13). We only have to replace $S_{n,k}(f)$ by $I_{n,k}(h_f)$, $I_{n,|V|}(f_V)$ by $I_{n,|V|}((h_f)_V)$ and the levels $2^kA^{4/3}n^k\sigma^2$ and $A^{4/3}n^k\sigma^2$ by $An^k\sigma^2$ and $2^{-k}An^k\sigma^2$. Let us observe that each term of the upper bound we get in such a way can be directly bounded, since our inductive hypothesis the result of Proposition 13.2 holds also for k.

In the case $e(G)<k$ formula (15.10) will be proved with the help of the multivariate version of Hoeffding's inequality, Theorem 12.3. In the proof of this case an expression, analogous to $S^2_{n,k}(f)$ defined in formula (15.1) will be introduced and estimated for all sets $V_1,V_2\subset\{1,\dots,k\}$ and diagrams $G\in\mathcal{G}$ such that $|e(G)|<k$. To define this expression first some notations will be introduced.

Let us consider the set $J_0(G)=J_0(G,k,n)$,

$$\begin{aligned}J_0(G)=\{(l_1,\dots,l_k,l'_1,\dots,l'_k)\colon\ &1\le l_j,l'_j\le n,\ 1\le j\le k,\ l_j\ne l_{j'} \text{ if } j\ne j',\\ &l'_j\ne l'_{j'} \text{ if } j\ne j',\ \ l_j=l'_{j'} \text{ if } (j,j')\in e(G),\ l_j\ne l'_{j'} \text{ if } (j,j')\notin e(G)\},\end{aligned}$$

the set of those sequences $(l_1,\dots,l_k,l'_1,\dots,l'_k)$ which appear as indices in the summation in formula (14.10). Let us introduce a partition of $J_0(G)$ appropriate for our purposes.

For this aim let us first define the sets $M_1=M_1(G)=\{j(1),\dots,j(k-|e(G)|)\}=\{1,\dots,k\}\setminus v_1(G)$, $j(1)<\dots<j(k-|e(G)|)$, and $M_2=M_2(G)=\{\bar\jmath(1),\dots,\bar\jmath(k-|e(G|)\}=\{1,\dots,k\}\setminus v_2(G)$, $\bar\jmath(1)<\dots<\bar\jmath(k-|e(G|)$, the sets of those vertices of the first and second row of the diagram G in increasing order from which no edges start. Let us also introduce the set $V(G)=V(G,n,k)$,

$$\begin{aligned}V(G)=\{(l_{j(1)},\dots,l_{j(k-|e(G)|)},l'_{\bar\jmath(1)},\dots,l'_{\bar\jmath(k-|e(G)|)})\colon\ &1\le l_{j(p)},l'_{\bar\jmath(p)}\le n,\\ &1\le p\le k-|e(G)|,\ l_{j(p)}\ne l_{j(p')},\ l'_{\bar\jmath(p)}\ne l'_{\bar\jmath(p')} \text{ if } p\ne p',\ 1\le p,p'\le k-|e(G)|,\\ &l_{j(p)}\ne l'_{\bar\jmath(p')},\ 1\le p,p'\le k-|e(G)|\}.\end{aligned}$$

The set $V(G)$ consists of such vectors which can be obtained as the restriction of some vector $(l_1,\dots,l_k,l'_1,\dots,l'_k)\in J_0(G)$ to the coordinates indexed by the elements of the set $M_1\cup M_2$. The elements of $V(G)$ can be characterized as such vectors, whose coordinates indexed by the set $M_1\cup M_2$, take different integer values between 1 and n. Given a vector $v\in V(G)$ put $v=(v_1,v_2)$, and let $v_1=\{v(r),\ 1\le r\le k-|e(G)|\}$, and $v_2=\{\bar v(r),\ 1\le r\le k-|e(G)|\}$, denote the set of coordinates of v indexed by the elements of the set M_1 and M_2 respectively. For all vectors $v\in V(G)$ define the set

$$\begin{aligned}E_G(v)=\{(l_1,\dots,l_k,l'_1,\dots,l'_k)\colon\ &1\le l_j\le n,\ 1\le l'_{\bar\jmath}\le n,\ \text{for } 1\le j,\bar\jmath\le k,\\ &l_j\ne l_{j'} \text{ if } j\ne j',\ l'_{\bar\jmath}\ne l'_{\bar\jmath'} \text{ if } \bar\jmath\ne\bar\jmath',\\ &l_j=l'_{\bar\jmath} \text{ if } (j,\bar\jmath)\in e(G) \text{ and } l_j\ne l'_{\bar\jmath} \text{ if } (j,\bar\jmath)\notin e(G),\\ &l_{j(r)}=v(r),\ l'_{\bar\jmath(r)}=\bar v(r),\ 1\le r\le k-|e(G)|\},\quad v\in V(G),\end{aligned}$$

where $\{j(1),\dots,j(k-|e(G)|)\}=M_1$, $\{\bar\jmath(1),\dots,\bar\jmath(k-|e(G)|)\}=M_2$, $v=(v^{(1)},v^{(2)})$ with $v^{(1)}=(v(1),\dots,v(k-|e(G)|))$ and $v^{(2)}=(\bar v(1),\dots,\bar v(k-|e(G)|))$ in the last line of this definition. The elements $\ell=(l_1,\dots,l_k,l'_1,\dots,l'_k)$ of the set $E_G(v)$ for some $v\in V(G)$ can be characterized in the following way: For $j\in M_1$ the coordinate l_j agrees with the corresponding element of $v^{(1)}$, for $\bar\jmath\in M_2$ the coordinate $l'_{\bar\jmath}$ agrees with the corresponding element of $v^{(2)}$. The indices of the remaining coordinates of ℓ can be partitioned into pairs $(j_s,\bar\jmath_s)$, $1\le s\le |e(G)|$ in such a way that $(j_s,\bar\jmath_s)\in e(G)$. The identity $l_{j_s}=l'_{\bar\jmath_s}$ holds for all these pairs, and these values $l_{j_s}=l'_{\bar\jmath_s}$ must be different for different indices s. Otherwise, they can be chosen freely in the set $\{1,\dots,n\}\setminus\{v^{(1)},v^{(2)}\}$.

The sets $E_G(v)$, $v\in V(G)$, constitute a partition of the set $J_0(G)$, and we can rewrite with their help the random variables $H_{n,k}(f|G,V_1,V_2)$ defined in (14.10) as

$$H_{n,k}(f|G,V_1,V_2)=\sum_{v=(l_{j(1)},\dots,l_{j(k-|e(G)|)},l'_{\bar\jmath(1)},\dots,l'_{\bar\jmath(k-|e(G)|)})\in V(G)}\prod_{s=1}^{k-|e(G)|}\varepsilon_{l_{j(s)}}\prod_{s=1}^{k-|e(G)|}\varepsilon_{l'_{\bar\jmath(s)}}$$
$$\sum_{(l_1,\dots,l_k,l'_1\dots,l'_k)\in E_G(v)}\frac{1}{k!^2}\int f(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)},y)f(\xi_{l'_1}^{(1,\bar\delta_1)},\dots,\xi_{l'_k}^{(k,\bar\delta_k)},y)\rho(\,dy) \tag{15.14}$$

where $\delta_j=1$ if $j\in V_1$, $\delta_j=-1$ if $j\notin V_1$, and $\bar\delta_j=1$ if $j\in V_2$, $\bar\delta_j=-1$ if $j\notin V_2$.

The inequality

$$P\left(S^2(\mathcal{F}|G,V_1,V_2)>A^{8/3}n^{2k}\sigma^4\right)\le 2^{k+1}e^{-A^{2/3k}n\sigma^2}\quad\text{if } A\ge A_0 \text{ and } e(G)<k \tag{15.15}$$

will be proved for the random variable

$$S^2(\mathcal{F}|G,V_1,V_2)=\sup_{f\in\mathcal{F}}\frac{1}{k!^2}\sum_{v\in V(G)}\left(\sum_{(l_1,\dots,l_k,l'_1,\dots,l'_k)\in E_G(v)}\int f(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)},y)\right.$$
$$\left. f(\xi_{l'_1}^{(1,\bar\delta_1)},\dots,\xi_{l'_k}^{(k,\bar\delta_k)},y)\rho(\,dy)\right)^2, \tag{15.15$'$}$$

where $\delta_j=1$ if $j\in V_1$, $\delta_j=-1$ if $j\notin V_1$, and $\bar\delta_j=1$ if $j\in V_2$, $\bar\delta_j=-1$ if $j\notin V_2$. The random variable $S^2(\mathcal{F}|G,V_1,V_2)$ defined in (15.15$'$) plays a similar role in the proof of Proposition 13.3 as the random variable $\sup\limits_{f\in\mathcal{F}}S^2_{n,k}(f)$ in the proof of Proposition 13.2, where $S^2_{n,k}(f)$ was defined in formula (15.1).

To prove formula (15.15) let us first fix some $v\in V(G)$ and let us observe that, similarly to the proof of relation (15.11), the Schwarz inequality implies the relation

$$\left(\sum_{(l_1,\dots,l_k,l'_1,\dots,l'_k)\in E_G(v)}\int f(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)},y)f(\xi_{l'_1}^{(1,\bar\delta_1)},\dots,\xi_{l'_k}^{(k,\bar\delta_k)},y)\rho(\,dy)\right)^2$$

$$\leq \left(\sum_{(l_1,\dots,l_k,l'_1,\dots,l'_k)\in E_G(v)} \int f^2(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)},y)\rho(\,dy)\right)$$
$$\left(\sum_{(\bar l_1,\dots,\bar l_k,\bar l'_1,\dots,\bar l'_k)\in E_G(v)} \int f^2(\xi_{\bar l'_1}^{(1,\bar\delta_1)},\dots,\xi_{\bar l'_k}^{(k,\bar\delta_k)},y)\rho(\,dy)\right)$$

for all $v\in V(G)$. Summing up these inequalities for all $v\in V(G)$ we get that

$$\begin{aligned}
S^2(\mathcal{F}|G,V_1,V_2) &\leq \sup_{f\in\mathcal{F}} \sum_{v\in V(G)} \frac{1}{k!}\left(\sum_{(l_1,\dots,l_k,l'_1,\dots,l'_k)\in E_G(v)} \int f^2(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)},y)\rho(\,dy)\right)\\
&\qquad \frac{1}{k!}\left(\sum_{(\bar l_1,\dots,\bar l_k,\bar l'_1,\dots,\bar l'_k)\in E_G(v)} \int f^2(\xi_{\bar l'_1}^{(1,\bar\delta_1)},\dots,\xi_{\bar l'_k}^{(k,\bar\delta_k)},y)\rho(\,dy)\right) \qquad (15.16)\\
&\leq \sup_{f\in\mathcal{F}} \frac{1}{k!}\left(\sum_{\substack{(l_1,\dots,l_k),\,1\leq l_j\leq n,\,1\leq j\leq k,\\ l_j\neq l_{j'} \text{ if } j\neq j'}} \int f^2(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)},y)\rho(\,dy)\right)\\
&\qquad \sup_{f\in\mathcal{F}} \frac{1}{k!}\left(\sum_{\substack{(\bar l_1,\dots,\bar l_k),\,1\leq \bar l_j\leq n,\,1\leq j\leq k,\\ \bar l_j\neq \bar l_{j'} \text{ if } j\neq j'}} \int f^2(\xi_{\bar l_1}^{(1,\bar\delta_1)},\dots,\xi_{\bar l_k}^{(k,\bar\delta_k)},y)\rho(\,dy)\right)
\end{aligned}$$

To check the second inequality of formula (15.16) let us first observe that it can be reduced to the simpler relation, where the expression $\sup\limits_{f\in\mathcal{F}}$ is omitted at each place. The simplified inequality we get after the omission of the expressions sup can be checked by carrying out the term by term multiplication between the products of sums appearing in (15.16). We get at both sides of the inequality sums consisting of terms of the form

$$\frac{1}{k!^2}\int f^2(\xi_{l_1}^{(1,\delta_1)},\dots,\xi_{l_k}^{(k,\delta_k)},y)\rho(\,dy)\int f^2(\xi_{\bar l_1}^{(1,\bar\delta_1)},\dots,\xi_{\bar l_k}^{(k,\bar\delta_k)},y)\rho(\,dy) \qquad (15.17)$$

and we have to check that if a term of this form appears in the middle term of the simplified version formula of (15.16), then it appears with coefficient 1, and it also appears at the right-hand side of this formula. To see this observe that each term of the form (15.17) which appears in the middle term determines uniquely the index $v=(v_1,v_2)\in V(G)$ in the outer sum in the middle term for which the product of the inner sums yields this term. Indeed, the coordinates of this vector $v=(v_1,v_2)$ (which depends only on the indices in $M_1\cup M_2$) is such that v_1 agrees with the coordinates of the vector $l=(l_1,\dots,l_k)$ at indices in M_1 and v_2 agrees with the coordinates of $(\bar l_1,\dots,\bar l_k)$ at indices in M_2. Besides, all terms of the form (15.17) which appear at the left-hand side also appear at the right-hand of this expression.

Relation (15.16) implies that

$$P(S^2(\mathcal{F}|G,V_1,V_2)) > A^{8/3}n^{2k}\sigma^4) \le 2P\left(\sup_{f\in\mathcal{F}} \bar I_{n,k}(h_f) > A^{4/3}n^k\sigma^2\right)$$

with $h_f(x_1,\dots,x_k) = \int f^2(x_1,\dots,x_k,y)\rho(\,dy)$. (Here we exploited that in the last formula $S^2(\mathcal{F}|G,V_1,V_2)$ is bounded by the product of two random variables whose distributions do not depend on the sets V_1 and V_2.) Thus to prove inequality (15.15) it is enough to show that

$$2P\left(\sup_{f\in\mathcal{F}} \bar I_{n,k}(h_f) > A^{4/3}n^k\sigma^2\right) \le 2^{k+1}e^{-A^{2/3k}n\sigma^2} \quad \text{if } A\ge A_0. \tag{15.18}$$

Actually formula (15.18) follows from the already proven formula (15.13), only the parameter A has to be replaced by $A^{4/3}$ in it.

With the help of relation (15.15) the proof of Proposition 13.3 can be completed similarly to that of Proposition 13.2. It follows from the generalized version of Hoeffding's inequality Theorem 12.3 and the definition of the random variable $H_{n,k}(f|G,V_1,V_2)$ given in the form (15.14) that

$$\begin{aligned}
P&\left(|H_{n,k}(f|G,V_1,V_2)| > \frac{A^2}{2^{4k+2}k!}n^{2k}\sigma^{2(k+1)}\,\middle|\, \xi_l^{j,\pm1},\, 1\le l\le n,\, 1\le j\le k\right)(\omega)\\
&\le Ce^{-B2^{-(4+2/k)}A^{2/3k}n\sigma^2} \quad \text{if} \quad S^2(\mathcal{F}|G,V_1,V_2)(\omega)\le A^{8/3}n^{2k}\sigma^4 \quad \text{for all } f\in\mathcal{F},\\
&\text{and } G\in\mathcal{G} \text{ such that } |e(G)|<k, \text{ and } V_1,V_2\in\{1,\dots,k\} \quad \text{if } A\ge A_0.
\end{aligned} \tag{15.19}$$

Indeed, in this case the conditional probability considered in (15.19) can be bounded by $C\exp\left\{-B\left(\frac{A^4n^{4k}\sigma^{4(k+1)}}{2^{8k+4}(k!)^2S^2(\mathcal{F}|G,V_1,V_2)/(k!)^2}\right)^{1/2j}\right\} \le C\exp\left\{-B\left(\frac{A^{4/3}n^{2k}\sigma^{4k}}{2^{8k+4}}\right)^{1/2j}\right\}$, where $2j=2k-2|e(G)|$, the number of vertices of the diagram G from which no edges start. Since $j\le k$, $n\sigma^2\ge 1$, and also $\frac{A^{4/3}}{2^{8k+4}}\ge 1$ if A_0 is chosen sufficiently large the above calculation implies relation (15.19).

Let us show that also the inequality

$$\begin{aligned}
P&\left(\sup_{f\in\mathcal{F}}|H_{n,k}(f|G,V_1,V_2)| > \frac{A^2}{2^{4k+1}k!}n^{2k}\sigma^{2(k+1)}\,\middle|\, \xi_l^{j,\pm1},\, 1\le l\le n,\, 1\le j\le k\right)(\omega)\\
&\le Cn^{(3k+1)L/2+\beta}e^{-BA^{2/3k}n\sigma^2/2^{(4+2/k)}(k!)^{1/k}} \quad \text{if } S^2(\mathcal{F}|G,V_1,V_2))(\omega)\le A^{8/3}n^{2k}\sigma^4\\
&\qquad \text{for all } G\in\mathcal{G} \text{ such that } |e(G)|<k, \text{ and } V_1,V_2\in\{1,\dots,k\} \quad \text{if } A\ge A_0
\end{aligned} \tag{15.20}$$

holds.

To prove formula (15.20) let us fix an elementary event $\omega\in\Omega$ which satisfies the relation $S^2(\mathcal{F}|G,V_1,V_2)(\omega)\le A^{8/3}n^{2k}\sigma^4$, two sets $V_1,V_2\subset\{1,\dots,k\}$, and a diagram

G such that $|e(G)|<k$, consider the points $x_l^{(j,\pm1)}=x_l^{(j,\pm1)}(\omega)=\xi_l^{(j,\pm1)}(\omega)$, $1\le l\le n$, $1\le j\le k$, and introduce with their help the following probability measures: For all $1\le j\le k$ define the probability measures $\nu_j^{(1)}$ which are uniformly distributed on the points $x_l^{(j,\delta_j)}$, $1\le l\le n$, and $\nu_j^{(2)}$ which are uniformly distributed on the points $x_l^{(j,\bar\delta_j)}$, $1\le l\le n$, i.e. let $\nu_j^{(1)}\left(\{x_l^{(j,\delta_j)}\}\right)=\frac1n$ and $\nu_j^{(2)}\left(\{x_l^{(j,\bar\delta_j)}\}\right)=\frac1n$, $1\le l\le n$, $1\le j\le k$, where $\delta_j=1$ if $j\in V_1$, $\delta_j=-1$ if $j\notin V_1$, and similarly $\bar\delta_j=1$ if $j\in V_2$ and $\bar\delta_j=-1$ if $j\notin V_2$. Let us consider the product measures $\alpha_1=\nu_1^{(1)}\times\cdots\times\nu_k^{(1)}\times\rho$, $\alpha_2=\nu_1^{(2)}\times\cdots\times\nu_k^{(2)}\times\rho$ on the product space $(X^k\times Y,\mathcal X^k\times\mathcal Y)$, where ρ is that probability measure on $(Y,\mathcal Y)$ which appears in Proposition 13.3, and define the measure $\alpha=\frac{\alpha_1+\alpha_2}{2}$. Given two functions $f\in\mathcal F$ and $g\in\mathcal F$ we give an upper bound for $|H_{n,k}(f|G,V_1,V_2)(\omega)-H_{n,k}(g|G,V_1,V_2)(\omega)|$ if $\int(f-g)^2\,d\alpha\le\delta^2$ with some $\delta>0$. (This bound does not depend on the 'randomizing terms' $\varepsilon_l(\omega)$ in the definition of the random variable $H_{n,k}(\cdot|G,V_1,V_2)$.)

In this case $\int(f-g)^2\,d\alpha_j\le2\delta^2$, and

$$\int|f(x_{l_1}^{(1,\delta_1)},\dots,x_{l_k}^{(k,\delta_k)},y)-g(x_{l_1}^{(1,\delta_1)},\dots,x_{l_k}^{(k,\delta_k)},y)|^2\rho(\,dy)\le2\delta^2n^k,$$
$$\int|f(x_{l_1}^{(1,\delta_1)},\dots,x_{l_k}^{(k,\delta_k)},y)-g(x_{l_1}^{(1,\delta_1)},\dots,x_{l_k}^{(k,\delta_k)},y)|\rho(\,dy)\le\sqrt2\delta n^{k/2}$$

for all $1\le l\le k$, and $1\le l_j\le n$, and the same result holds if all δ_j is replaced by $\bar\delta_j$, $1\le j\le k$. Since $|f|\le1$, $|g|\le1$ if $f,g\in\mathcal F$, the condition $\int(f-g)^2\,d\alpha\le\delta^2$ implies that

$$\begin{aligned}\int|f(\xi_{l_1}^{(1,\delta_1)}(\omega),\dots,\xi_{l_k}^{(k,\delta_k)}(\omega),y)f(\xi_{l_1'}^{(1,\bar\delta_1)}(\omega),\dots,\xi_{l_k'}^{(k,\bar\delta_k)}(\omega),y)\rho(\,dy)\\-g(\xi_{l_1}^{(1,\delta_1)}(\omega),\dots,\xi_{l_k}^{(k,\delta_k)}(\omega),y)g(\xi_{l_1'}^{(1,\bar\delta_1)}(\omega),\dots,\xi_{l_k'}^{(k,\bar\delta_k)}(\omega),y)\rho(\,dy)|\le2\sqrt2\delta n^{k/2}\end{aligned}$$

for all vectors $(l_1,\dots,l_k,l_1',\dots,l_k')$ which appear as an index in the summation in (14.10). Hence

$$|H_{n,k}(f|G,V_1,V_2)(\omega)-H_{n,k}(g|G,V_1,V_2)(\omega)|\le2\sqrt2\delta n^{5k/2}$$

if $f,g\in\mathcal F$, $\int(f-g)^2\,d\alpha<\delta^2$ and the originally fixed $\omega\in\Omega$ is considered. (The measure α is defined by means of this ω.)

Put $\bar\delta=\frac{A^2n^{-k/2}\sigma^{2(k+1)}}{2^{(4k+7/2)}k!}$, and $\delta=n^{-(3k+1)/2}\le\bar\delta$ (the inequality $\delta\le\bar\delta$ holds, since $\sigma\ge\frac1{\sqrt n}$ and we may assume that $A\ge A_0$ is sufficiently large), choose a δ-dense subset $\{f_1,\dots,f_m\}\subset\mathcal F$ in the $L_2(X^k\times Y,\mathcal X^k\times\mathcal Y,\alpha)$ space with $m\le D\delta^{-L}\le n^{(3k+1)L/2+\beta}$ elements. Relation (15.19) for these functions together with the above estimates yield formula (15.20).

It follows from relations (15.15) and (15.20) that

$$\begin{aligned}P\left(\sup_{f\in\mathcal F}|H_{n,k}(f|G,V_1,V_2)|>\frac{A^2}{2^{4k+1}k!}n^{2k}\sigma^{2(k+1)}\right)\le2^{k+1}e^{-A^{2/3k}n\sigma^2}\\+Cn^{(3k+1)L/2+\beta}e^{-B2^{-(4+2/k)}A^{2/3k}n\sigma^2}\quad\text{if } A\ge A_0\end{aligned}$$

for all $V_1, V_2 \subset \{1,\dots,k\}$ also in the case $|e(G)| \le k-1$. This inequality implies that relation (15.10) holds also in this case if the constants A_0 and K are chosen sufficiently large in Proposition 13.3. Proposition 13.3 is proved.

16. The improvement of some results in Section 8

In this section I present an improvement of Theorems 8.3 and 8.5. I shall explain the picture behind these results and the some ideas of the proofs. But the detailed proofs which are based on some results called the diagram formulas for the products of multiple Wiener–Itô integrals and degenerate U-statistics are omitted. These diagram formulas present an identity which enables us to express the product of Wiener–Itô integrals or degenerate U-statistics as a sum of such objects. I omitted the proofs because they heavily depend on the diagram formula, a technique not discussed in this work. The interested reader can find the detailed proofs in my papers [21] and [22].

The main result discussed in this section is the following

Theorem 16.1. *Let $\xi_1,\dots,\xi_n$ be a sequence of iid. random variables on a space $(X,\mathcal{X})$ with some distribution μ. Let us consider a function $f(x_1,\dots,x_k)$ canonical with respect to the measure μ on the space $(X^k,\mathcal{X}^k)$ which satisfies conditions (8.1) and (8.2) with some $0<\sigma^2\le 1$ together with the degenerate U-statistic $I_{n,k}(f)$ with this kernel function. There exist some constants $A=A(k)>0$ and $B=B(k)>0$ depending only on the order k of the U-statistic $I_{n,k}(f)$ such that*

$$P(k!n^{-k/2}|I_{n,k}(f)|>u) \le A\exp\left\{-\frac{u^{2/k}}{2\sigma^{2/k}\left(1+B\left(un^{-k/2}\sigma^{-(k+1)}\right)^{1/k}\right)}\right\} \tag{16.1}$$

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

Theorem 16.1 states in particular that if $0<u\le \varepsilon n^{k/2}\sigma^{k+1}$ with a sufficiently small $\varepsilon>0$, then $P(k!n^{-k/2}|I_{n,k}(f)|>u)\le A\exp\left\{-\frac{1-C\varepsilon^{1/k}}{2}\left(\frac{u}{\sigma}\right)^{2/k}\right\}$ with some universal constants $A>0$ and $C>0$ depending only on the order k of the U-statistic $I_{n,k}(f)$. This result is very similar to Theorem 8.3. Both theorems yield an estimate on the probability $P(k!n^{-k/2}|I_{n,k}(f)|>u)$ for $0\le u\le n^{k/2}\sigma^{k+1}$, but in the present result we also get a good estimate on the constant α in formula (8.9) for $0\le u\le \varepsilon n^{k/2}\sigma^{k+1}$. At first sight this additional result does not seem an essential improvement, but actually it expresses an important property of the estimate (16.1). To understand this it is worth while to compare Theorem 16.1 with Bernstein's inequality formulated in Theorem 3.1.

Theorem 3.1 implies the estimate

$$P(n^{-1/2}|I_{n,1}(f)|>u)\le 2e^{-Cu^2/\sigma^2} \quad \text{if } 0\le u\le n\sigma^2 \tag{16.2}$$

for the degenerate U-statistic $I_{n,1}(f)$ of order 1 with a kernel function f, (i.e. for a sum of iid. random variables $Ef(\xi_1)=0$) if the relations $\sup|f(x)|\le 1$ and $Ef(\xi_j)=0$ and $Ef^2(\xi_j)\le\sigma^2$ hold. Besides, relation (16.2) also holds with a constant of the form

$C=\frac{1-O(\varepsilon)}{2}$ if $0\le u\le \varepsilon n\sigma^2$. On the other hand, Example 3.2 shows an example (formulated with a different normalization) with a function f and a sequence of iid. random variables $\xi_1,\xi_2,\dots$ satisfying the above conditions such that

$$P(n^{-1/2}I_{n,1}(f)>u)\ge A\exp\left\{-B\left(\frac{u}{\sigma}\right)^2\cdot\frac{\sqrt{n}\sigma^2}{u}\log\frac{u}{\sqrt{n}\sigma^2}\right\}$$

if $u\gg n\sigma^2$. This means that in the special case $k=1$ the probability $P(n^{-1/2}|I_{n,1}(f)|>u)$ has a Gaussian type estimate for $0\le u\le \text{const.}\, n\sigma^2$, and such an estimate does not hold for $u\gg n\sigma^2$. Besides, in the smaller interval $0\le u\le\varepsilon n\sigma^2$ we can say more. In this case the relation (16.2) holds with such a constant C which almost agrees with the number $\frac12$, i.e. the upper bound we get for $k=1$ almost agrees with the quantity suggested by a formal application of the central limit theorem.

I want to explain that Theorem 16.1 states a similar result for degenerate U-statistics of any order $k\ge1$. To understand this let us first recall that a sequence of normalized degenerate U-statics $n^{-k/2}I_{n,k}(f)$, $n=1,2,\dots$, defined with the help of a sequence of iid. random variables $\xi_1,\xi_2,\dots$ taking values on some measurable space $(X,\mathcal{X})$ with distribution μ and a function $f(x_1,\dots,x_k)$ of k variables canonical with respect to μ and such that

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

has a limit distribution as $n\to\infty$. Moreover, this limit can be expressed explicitly as the distribution of the Wiener–Itô integral

$$Z_{\mu,k}(f)=\frac{1}{k!}\int f(x_1,\dots,x_k)\mu_W(\,dx_1)\dots\mu_W(dx_k), \tag{16.3}$$

where μ_W is the white noise with counting measure μ, i.e. $\mu_W(A)$, $A\in\mathcal{X}$, is a Gaussian field indexed by the measurable subsets of the space X such that $E\mu_W(A)=0$ and $E\mu_W(A)\mu_W(B)=\mu(A\cap B)$ for all $A,B\in\mathcal{X}$. (The definition of Wiener–Itô integrals can be found e.g. in [17].) Hence it is natural to expect that in the estimates about the distribution of degenerate U-statistics the distributions of Wiener–Itô integrals play a role similar to the Gaussian distributions in the case $k=1$. Therefore we are interested in good estimates on the distribution of Wiener–Itô integrals. The next result supplies such an estimate. As Theorem 16.1 was an improvement of Theorem 8.3, the next result is an improvement of the first estimate in Theorem 8.5 presented in formula (8.11).

Theorem 16.2. *Let us consider a σ-finite measure μ on a measurable space together with a white noise μ_W with counting measure μ. Let us have a real-valued function $f(x_1,\dots,x_k)$ on the space $(X^k,\mathcal{X}^k)$ which satisfies relation (8.2). Take the random integral $Z_{\mu,k}(f)$ introduced in formula (16.3). This random integral satisfies the inequality*

$$P(k!|Z_{\mu,k}(f)|>u)\le C\exp\left\{-\frac12\left(\frac{u}{\sigma}\right)^{2/k}\right\}\quad \textit{for all } u>0 \tag{16.4}$$

with an appropriate constant $C=C(k)>0$ depending only on the multiplicity k of the integral.

In Theorem 16.2 we gave only an upper bound for the distribution of Wiener–Itô integrals. The following example shows that there are cases when this estimate is essentially sharp.

Example 16.3. *Let us have a σ-finite measure μ on some measure space $(X,\mathcal{X})$ together with a white noise μ_W on $(X,\mathcal{X})$ with counting measure μ. Let $f_0(x)$ be a real valued function on $(X,\mathcal{X})$ such that $\int f_0(x)^2\mu(\,dx)=1$, and take the function $f(x_1,\dots,x_k)=\sigma f_0(x_1)\cdots f_0(x_k)$ with some number $\sigma>0$ and the Wiener–Itô integral $Z_{\mu,k}(f)$ introduced in formula (16.3).*

Then the relation $\int f(x_1,\dots,x_k)^2\,\mu(\,dx_1)\dots\mu(\,dx_k)=\sigma^2$ holds, and the random integral $Z_{\mu,k}(f)$ satisfies the inequality

$$P(k!|Z_{\mu,k}(f)|>u)\ge\frac{\bar C}{\left(\frac u\sigma\right)^{1/k}+1}\exp\left\{-\frac12\left(\frac u\sigma\right)^{2/k}\right\}\quad\textit{for all } u>0 \tag{16.5}$$

with some constant $\bar C>0$.

Proof of the statement of Example 16.3. We may restrict our attention to the case $k\ge2$. Itô's formula (see e.g. [17]) states that the random variable $k!\bar Z_{\mu,k}(f)$ can be expressed as $k!Z_{\mu,k}(f)=\sigma H_k\left(\int f_0(x)\mu_W(\,dx)\right)=\sigma H_k(\eta)$, where $H_k(x)$ is the k-th Hermite polynomial with leading coefficient 1, and $\eta=\int f_0(x)\mu_W(\,dx)$ is a standard normal random variable. Hence we get by exploiting that the coefficient of x^{k-1} in the polynomial $H_k(x)$ is zero that $P(k!|Z_{\mu,k}(f)|>u)=P(|H_k(\eta)|\ge\frac u\sigma)\ge P\left(|\eta^k|-D|\eta^{k-2}|>\frac u\sigma\right)$ with a sufficiently large constant $D>0$ if $\frac u\sigma>1$. There exist such positive constants A and B that $P\left(|\eta^k|-D|\eta^{k-2}|>\frac u\sigma\right)\ge P\left(|\eta^k|>\frac u\sigma+A\left(\frac u\sigma\right)^{(k-2)/k}\right)$ if $\frac u\sigma>B$.

Hence

$$P(k!|Z_{\mu,k}(f)|>u)\ge P\left(|\eta|>\left(\frac u\sigma\right)^{1/k}\left(1+A\left(\frac u\sigma\right)^{-2/k}\right)\right)\ge\frac{\bar C\exp\left\{-\frac12\left(\frac u\sigma\right)^{2/k}\right\}}{\left(\frac u\sigma\right)^{1/k}+1}$$

with an appropriate $\bar C>0$ if $\frac u\sigma>B$. Since $P(k!|Z_{\mu,k}(f)|>0)>0$, the above inequality also holds for $0\le\frac u\sigma\le B$ if the constant $\bar C>0$ is chosen sufficiently small. This means that relation (16.5) holds.

Let us remark that if $f(x_1,\dots,x_k)=\sigma f_0(x_1)\dots f_0(x_k)$ is a function on the space $(X^k,\mathcal{X}^k)$ such that $\int f_0(x)\mu(\,dx)=0$, $\int f_0^2(x)\mu(\,dx)=1$, $\sup|f_0(x)|\le1$, $0<\sigma\le1$, and we have a sequence of iid. random variables, $\xi_1,\xi_2,\dots$ with distribution μ, then the U-statistics $I_{n,k}(f)$, $n=1,2,\dots$, are degenerate, and they satisfy the conditions of Theorem 16.1. Besides, they converge in distribution to the Wiener–Itô integral $Z_{\mu,k}(f)$ as $n\to\infty$ which satisfies the conditions of example (16.3). Hence the U-statistics $I_{n,k}(f)$ satisfy relation (16.1), and also the inequality

$$\lim_{n\to\infty}P(k!n^{-k/2}|I_{n,k}(f)|>u)\ge\frac{\bar C\exp\left\{-\frac12\left(\frac u\sigma\right)^{2/k}\right\}}{\left(\frac u\sigma\right)^{1/k}+1}$$

holds with an appropriate $\bar{C} > 0$ if $\frac{u}{\sigma} > B$. This means that for not too large values of u, more explicitly if $u \le \varepsilon n^{k/2}\sigma^{k+1}$ with a small number $\varepsilon > 0$, the estimate given in Theorem 16.1 is essentially sharp. Let me also remark that Example 8.6 shows such a degenerate U-statistic of order $k = 2$ for which an estimate similar to that of Theorem 8.3 cannot hold for $n \gg n^{k/2}\sigma^{k+1}$. We have presented such an example only for $k = 2$, but similar examples can be given for all $k \ge 1$.

This means that Theorem 16.1 shows a similar picture about the distribution of degenerate U-statistics of order k for all $k \ge 1$ as Bernstein's inequality shows in the case $k = 1$. We have a good estimate on the distribution $P(n^{-k/2}I_{n,k}(f) > u)$ of a degenerate U-statistic with a kernel function f satisfying relations (8.1) and (8.2) in the domain $0 \le u \le n^{k/2}\sigma^{k+1}$. Such an estimate is already proved in Theorem 8.3, but Theorem 16.1 says more in an interval of the form $0 \le u \le \varepsilon n^{k/2}\sigma^{k+1}$ with a small $\varepsilon > 0$. The limit theorems about degenerate U-statistics give an upper bound for the coefficient α in the exponent of formula (8.9) in Theorem 8.3, and Theorem 16.1 states that the estimate (8.9) holds with an almost as good coefficient α in the interval $0 \le u \le \varepsilon n^{k/2}\sigma^{k+1}$ as this upper bound suggests.

The proof of the above results are based, similarly to the proof of Theorems 8.3 and 8.5, on some good estimates on high moments of degenerate U-statistics $I_{n,k}(f)$ and of Wiener–Itô integrals $Z_{n,k}(f)$. The result of Theorem 16.2 follows from the following

Proposition 16.4. *Let the conditions of Theorem 16.2 be satisfied for a multiple Wiener–Itô integral $Z_{\mu,k}(f)$ of order k. Then, with the notations of Theorem 16.2, the inequality*

$$E\left(k!|Z_{\mu,k}(f)|\right)^{2M} \le 1\cdot 3\cdot 5\cdots(2kM-1)\sigma^{2M} \quad \text{for all } M = 1,2,\dots \tag{16.6}$$

holds.

By the Stirling formula Proposition 16.4 implies that

$$E(k!|Z_{\mu,k}(f)|)^{2M} \le \frac{(2kM)!}{2^{kM}(kM)!}\sigma^{2M} \le A\left(\frac{2}{e}\right)^{kM}(kM)^{kM}\sigma^{2M} \tag{16.7}$$

for all numbers $A > \sqrt{2}$ if $M \ge M_0 = M_0(A)$. The following Proposition 16.5 states a similar, but weaker inequality for the moments of normalized degenerate U-statistics.

Proposition 16.5. *Let us consider a degenerate U-statistic $I_{n,k}(f)$ of order k with sample size n and with a kernel function f satisfying relations (8.1) and (8.2) with some $0 < \sigma^2 \le 1$. Fix a positive number $\eta > 0$. There exists some universal constants $A = A(k) > \sqrt{2}$, $C = C(k) > 0$ and $M_0 = M_0(k) \ge 1$ depending only on the order of the U-statistic $I_{n,k}(f)$ such that*

$$E\left(n^{-k/2}k!I_{n,k}(f)\right)^{2M} \le A\,(1+C\sqrt{\eta})^{2kM}\left(\frac{2}{e}\right)^{kM}(kM)^{kM}\,\sigma^{2M} \tag{16.8}$$

for all integers M such that $kM_0 \le kM \le \eta n\sigma^2$.

The constant $C = C(k)$ in formula (2.3) can be chosen e.g. as $C = 2\sqrt{2}$ which does not depend on the order k of the U-statistic $I_{n,k}(f)$.

Let us remark that formula (16.6) can be reformulated as $E(k!|Z_{\mu,k}(f)|)^{2M} \le E(\sigma\eta^k)^{2M}$, where η is a standard normal random variable. Theorem 16.2 states that the tail distribution of $k!|Z_{\mu,k}(f)|$ satisfies an estimate similar to that of $\sigma|\eta|^k$. This follows simply from Proposition 16.4 and the Markov inequality $P(k!|Z_{\mu,k}(f)| > u) \le \frac{E(k!|Z_{\mu,k}(f)|)^{2M}}{u^{2M}}$ with an appropriate choice of the parameter M.

Proposition 16.5 states that in the case $M_0 \le M \le \varepsilon n\sigma^2$ the inequality

$$E\left(n^{-k/2}k!I_{n,k}(f)\right)^{2M} \le E((1+\beta(\varepsilon))\sigma\eta^k)^{2M}$$

holds with a standard normal random variable η and a function $\beta(\varepsilon)$, $0 \le \varepsilon \le 1$, such that $\beta(\varepsilon) \to 0$ if $\varepsilon \to 0$, and $\beta(\varepsilon) \le C$ with some universal constant $C = C(k)$ if $0 \le \varepsilon \le 1$. This means that certain high but not too high moments of $n^{-k/2}k!I_{n,k}(f)$ behave similarly to the moments of $k!Z_{\mu,k}(f)$. As a consequence, we can prove a similar, but slightly weaker estimate for the distribution of $n^{-k/2}k!I_{n,k}(f)$ as for the distribution of $k!Z_{\mu,k}(f)$. Actually this is done in the proof of Theorem 16.1.

Estimate (16.8) is very similar to the bound (10.1) formulated in Proposition (10.1). The main difference is that here we get the estimate

$$E\left(n^{-k/2}k!I_{n,k}(f)\right)^{2M} \le C^M(kM)^{kM}\sigma^{2M} \tag{16.9}$$

with a good constant C, at least if $M \le \varepsilon n\sigma^2$ with a small number $\varepsilon > 0$. The method of proof of Theorem 8.3 presented in this paper cannot yield such a good estimate. The main problem with this method is that it applies a symmetrization argument (this is done in the proof of the Marcinkiewicz–Zygmund inequality), in which we bound the moments of the random variable we are investigating by the moments of a random variable with constant times larger variance. Such a step in the proof does not allow to get the estimate (16.9) with a good constant $C > 0$.

On the other hand, the estimation of the moments of a degenerate U-statistics by means of the diagram formula yields a better estimate of the moments. The idea behind this approach is that in calculating the even moments $E\left(I_{n,k}(f)\right)^{2M}$ of a degenerate U-statistics by means of the diagram formula we have to work with some terms which also appear in the calculation of the moments $E(Z_{\mu,k}(f))^{2M}$ of the Wiener–Itô integral $Z_{\mu,k}(f)$, but we also have to handle some additional terms. It must be checked that the contribution of these additional terms is not too large. This is the case if $M \le n\sigma^2$ with $\sigma^2 = \int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)$, and an even better estimate can be given about the contribution of these terms if $M \ge \varepsilon n\sigma^2$ with a small $\varepsilon > 0$.

Let me finally remark that the above method can also give an improvement of the multivariate version of the Hoeffding inequality (Theorem 12.3). The proof of the following inequality can be found in [22].

Theorem 16.6. The multivariate version of Hoeffding's inequality. *Let* $\varepsilon_1, \dots,$ ε_n *be independent random variables,* $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = \frac{1}{2}$, $1 \le j \le n$. *Fix a positive integer* k, *and define the random variable*

$$Z = \sum_{\substack{(j_1,\dots,j_k)\colon\, 1\le j_l\le n \text{ for all } 1\le l\le k\\ j_l\ne j_{l'} \text{ if } l\ne l'}} a(j_1,\dots,j_k)\varepsilon_{j_1}\cdots\varepsilon_{j_k} \tag{16.10}$$

with the help of some real numbers $a(j_1,\dots,j_k)$ *which are given for all sets of indices such that* $1\le j_l\le n$, $1\le l\le k$, *and* $j_l\ne j_{l'}$ *if* $l\ne l'$. *Put*

$$S^2 = \sum_{\substack{(j_1,\dots,j_k)\colon\, 1\le j_l\le n \text{ for all } 1\le l\le k\\ j_l\ne j_{l'} \text{ if } l\ne l'}} a^2(j_1,\dots,j_k) \tag{16.11}$$

Then

$$P(k!|Z|>u) \le C\exp\left\{-\frac{1}{2}\left(\frac{u}{S}\right)^{2/k}\right\} \quad \text{for all } u\ge 0 \tag{16.12}$$

with some constant $C>0$ *depending only on the parameter* k.

We may assume that the coefficients $a(j_1,\dots,j_k)$ in formulas (16.10) and (16.11) are symmetric functions of their arguments, i.e. $a(j_1,\dots,j_k) = a(j_{\pi(1)},\dots,j_{\pi(k)})$ for all permutations $\pi\in\Pi_k$ of the set $\{1,\dots,k\}$. If these coefficients $a(j_1,\dots,j_k)$ do not have not this symmetry property, then we can replace them with their symmetrizations $a_{\text{Sym}}(j_1,\dots,j_k) = \frac{1}{k!}\sum\limits_{\pi\in\Pi_k} a(j_{\pi(1)},\dots,j_{\pi(k)})$. In such a way we do not modify the value of the random variable Z, and decrease the value of the number S^2. With such a choice of the coefficients we have $EZ=0$ and $\operatorname{Var} Z = k!S^2$.

The main advantage of this result with respect to Theorem 12.3 is that formula (16.12) holds with the right constant in the exponent at the right-hand side. The proof is based on good moment estimates of the random variable Z defined in (16.10). I formulate this result which may be interesting in itself.

Theorem 16.7 *The random variable* Z *defined in formula (16.10) satisfies the inequality*

$$EZ^{2M} \le 1\cdot 3\cdot 5\cdots(2kM-1)S^{2M} \quad \text{for all } M=1,2,\dots \tag{16.13}$$

with the constant S *defined in formula (16.11).*

It is worth while to compare formula (16.13) with the estimate that Borell's inequality yields for this problem. By applying Borell's inequality with the choice $q=2$ and $p=2M$ we get that $EZ^{2M} \le (2M-1)^{kM}E(Z^2)^M = (2M-1)^{kM}\left(k!S^2\right)M$. Since $(2M-1)^{2M} = (2M)^{kM}\left(1-\frac{1}{2M}\right)^{kM} \sim e^{-k/2}(2M)^{kM}$ for large values M, hence Borell's inequality yields the inequality $EZ^{2M} \le \text{const.}\,(2M)^{kM}S^{2M}\cdot(k!)^M$ for large exponents M. On the other hand, Theorem 16.7 together with the Stirling formula yield the estimate $EZ^{2M} \le \text{const.}\,(2M)^{kM}S^{2M}\cdot\left(\frac{k}{e}\right)^{kM}$. It can be seen that $k! > \left(\frac{k}{e}\right)^k$ for all $k\ge 1$. This means that Theorem 16.7 yields an improvement of the Borell's inequality in the special case discussed above. This estimate is only a special case of Borell's inequality, but this is its most important special case.

17. An overview of the results in this work

I discuss briefly the problems investigated in this work and recall some basic results related to them. I also give a list of works where they can be found. Besides, I discuss some background problems and results which may explain the motivation for the study presented here.

I met the main problem considered in this work when tried to adapt the method of proof of the central limit theorem for maximum-likelihood estimates to some more difficult questions about so-called non-parametric maximum likelihood estimate problems. The Kaplan–Meyer estimate for the empirical distribution function with the help of censored data investigated in the second section is such a problem. It is not a maximum-likelihood estimate in the classical sense, but it can be considered as a non-parametric maximum likelihood estimate. Indeed, since in the estimation of a distribution function with the help of censored data the class of possible candidates for being the distribution function we are looking for is too large, there is no dominating measure with respect to which all of them have a density function. As a consequence, the classical principle of the maximum-likelihood estimate cannot be applied in this case. A natural way to overcome this difficulty is to choose a smaller class of distribution functions, to compare the probability of the appearance of the sample we observe with respect to all distribution functions of this class and to choose that distribution function as our estimate for which this probability takes its maximum. The Kaplan–Meyer estimate can be found on the basis of this principle in the following way: Let us estimate the distribution function $F(x)$ of the censored data simultaneously with the distribution function $G(x)$ of the censoring data. (We have a sample of size n and know which sample elements are censored and which are censoring data.) Let us consider the class of such pairs of estimates $(F_n(x), G_n(x))$ of the pair $(F(x), G(x))$ for which the distribution function $F_n(x)$ is concentrated in the censored sample points and the distribution function $G_n(x)$ is concentrated in the censoring sample points; more precisely, let us also assume that if the largest sample point is a censored point, then the distribution function $G_n(x)$ of the censoring data takes still another value which is larger than any sample point, and if it is a censoring point then the distribution function $F_n(x)$ of the censored data takes still another value larger than any sample point. (This modification at the end of the definition is needed, since if the largest sample points is from the class of censored data, then the distribution $G(x)$ of the censoring data in this point must be strictly less than 1, and if it is from the class of censoring data, then the value of the distribution function $F(x)$ of the censored data must be strictly less than 1 in this point.) Let us take this class of pairs of distribution functions $(F_n(x), G_n(x))$, and let us choose that pair of distribution functions of this class as the (non-parametric maximum likelihood) estimate with respect to which our observation has the greatest probability.

The above extremal problem for the pairs of distribution functions $(F_n(x), G_n(x))$ can be solved explicitly, and it yields the estimate of $F_n(x)$ written down in formula (2.3). (The function $G_n(x)$ satisfies a similar relation, only the random variables X_j and Y_j and the events $\delta_j = 1$ and $\delta_j = 0$ have to be replaced in it.) Then, as I have indicated, a natural analog of the linearization procedure in the maximum likelihood estimate also works in this case, and there is only one really hard part of the proof. We need a good estimate on the distribution of the integral of a function of two variables with respect to the product of a normalized empirical measure with itself. Moreover, we also need a good estimate on the distribution of the supremum of a class of integrals, when the elements of an appropriate class of functions are integrated with respect to the above product measure. The main subject of this work is to solve the above problems in a more general setting, when not only two-fold, but also k-fold integrals are considered with arbitrary number $k \ge 1$.

The proof of this work for the limit behaviour of the Kaplan–Meyer estimate applied the explicit form of this estimate. It would be interesting to find such a modification of this proof which exploits that the Kaplan–Meyer estimate is the solution of an appropriate extremal problem. We may expect that such a proof can be generalized to a general result about the limit behaviour for a wide class of non-parametric maximum likelihood estimates. Such a consideration is behind the remark of Richard Gill I quoted at the end of Section 2. I hope that such a program can be realized, but at the present time I cannot do this.

A detailed proof together with a sharp estimate on the speed of convergence for the limit behaviour of the Kaplan–Meyer estimate based on the ideas presented in Section 2 is given in paper [24]. Paper [25] explains more about its background, and it also discusses the solution of some other non-parametric maximum likelihood problems. The results about multiple integrals with respect to a normalized empirical distribution function needed in these works were proved in [17]. The results of [18] are completely satisfactory for the study in [24], but they also have some drawbacks. They do not show that if the random integrals we are considering have small variances, then they satisfy better estimates. Besides, if we consider the supremum of random integrals of an appropriate class of functions, then these results can be applied only in very special cases. Moreover, the method of proof of [18] did not allow a real generalization of its results, hence I had to find a different approach when tried to generalize them.

I do not know of other works where the distribution of multiple random integrals with respect to a normalized empirical distribution is studied. On the other hand, there are some works where the distribution of (degenerate) U-statistics is investigated. The most important results obtained in this field are contained in the book of de la Peña and Giné *Decoupling, From Dependence to Independence* [6]. The problems about the behaviour of degenerate U-statistics and multiple integrals with respect to a normalized empirical distribution function are closely related, but the explanation of their relation is far from trivial. I return to this question later.

Even the study of the one-dimensional version of the problems studied here, i.e. the description of the behaviour of one-fold integrals or classes of one-fold integrals contains several hard problems which have to be investigated closely to have a good understanding of the subject. In the one-dimensional case it is fairly simple to prove that the problems about the behaviour of one-fold integrals with respect to a normalized empirical measure and about the behaviour of normalized sums of independent random variables are equivalent. I start this work with the description of the case of (classes of) one-fold integrals or of sums of independent random variables. This question has a fairly big literature. I would mention first of all the books *A course on empirical processes* [9], *Real Analysis and Probability* [10] and *Uniform Central Limit Theorems* [11] written by R. M. Dudley. These books contain a much more detailed description of the empirical processes than the present work together with a lot of interesting results.

The first problem studied here deals with the tail behaviour of sums of independent and bounded random variables with expectation zero. This question is considered in Section 3 where the proof of two already classical results, that of Bernstein's and Bennett's inequalities is explained. (These results are proved e.g. [4]). We are also interested in the question when these results give an estimate suggested by the central limit theorem. Bernstein's inequality provides such an estimate if the variance of the sum is not too small. (The results in Section 3 tell explicitly when this variance should be considered too small.) If the variance of the sum is too small, then Bennett's inequality provides a slight improvement of Bernstein's inequality. On the other hand, Example 3.2 shows that in the unpleasant case when this variance is too small Bennett's inequality is essentially sharp. I inserted this example to the text, because it may help to understand better the content of Bernstein's and Bennett's inequality. I have not found similar examples in the literature.

The estimate on the distribution of a sum of independent random variables if this sum has a small variance is weak because of the following reason. In this case the probability that the sum will be larger than a given value may be much larger than the (rather small) value suggested by the central limit theorem because of the appearance of some irregularities with relatively large probability. The hardest problems we have to cope with in the solution of the problems of this work are closely related to the weak estimates for sums of independent random variables if the variance of the sums are small and to the weak estimates in some similar problems. The weakness of these estimates imply that in the study of the problems we are interested in the method of proof for their Gaussian counterpart cannot be adapted completely, some new ideas are needed. We have overcome this difficulty by applying a symmetrization argument. The last result of Section 3, Hoeffding's inequality presented in Theorem 3.4 is an important ingredient of this symmetrization argument. It is also a classical result whose proof can be found for instance in [15].

In Section 4 I formulated the one-variate version of our main result about the supremum of the integrals of a class $\mathcal{F}$ of functions with respect to a normalized empirical measure together with an equivalent statement about the distribution of the supremum of a class of random sums $\sum\limits_{j=1}^{n} f(\xi_j)$ defined with the help of a sequence of i.i.d. random variables $\xi_1, \dots, \xi_n$ and a class of functions $f \in \mathcal{F}$ satisfying some appropriate conditions. These results are given in Theorems 4.1 and 4.1$'$. Also a Gaussian version of them is presented in Theorem 4.2 about the distribution of the supremum of a Gaussian field with some appropriate properties.

In the above mentioned results we have imposed the condition that the class of functions $\mathcal{F}$ or what is equivalent the set of random variables whose supremum we estimate is countable. In the proofs this condition is really exploited. On the other hand, in some important applications we also need results about the supremum of a possibly non-countable set of random variables. Hence I introduced the notion of countably approximable classes of random variables and proved that in the results of this work the condition about countability can be replaced by the weaker condition that the class of random variables whose supremum is taken is countably approximable. R. M. Dudley worked out a different method to handle the supremum of possibly non-countably many random variables, and generally his method is applied in the literature. The relation between these two methods deserves some discussion.

Let us first recall that if a class of random variables S_t, $t \in T$, indexed by some index set T is given, then a set A can be measurable with respect to the σ-algebra generated by the random variables S_t, $t \in T$, only if there exists a countable subset $T' = T'(A) \subset T$ such that the set A is also measurable with respect to the smaller σ-algebra generated by the random variable S_t, $t \in T'$. Besides, if the finite dimensional distributions of the random variables S_t, $t \in T$, are given, then by the results of classical measure theory also the probability of the events measurable with respect to the σ-algebra generated by these random variables S_t, $t \in T$, is determined. But there are rather few other events whose probabilities are determined by the finite dimensional distributions of the random variables S_t, $t \in T$. On the other hand, if T is a non-countable set, then the events $\left\{\sup\limits_{t\in T} S_t > u\right\}$ are not measurable with respect to the above σ-algebra, hence generally we cannot speak of their probabilities. To overcome this difficulty Dudley worked out a theory which enabled him to work also with outer measures. His theory is based on some rather deep results of the analysis. It can be found for instance in his book [11].

I restricted my attention to the case when after the completion of the probability measure P we can also speak of the real (and not only outer) probabilities $P\left(\sup\limits_{t\in T} S_t > u\right)$. I tried to find appropriate conditions under which these probabilities really exist. More explicitly, we are interested in the case when for all $u > 0$ there exists some set $A = A_u$ measurable with respect to the σ-algebra generated by the random variables S_t, $t \in T$, such that the symmetric difference of the sets A_u and $\left\{\sup\limits_{t\in T} S_t > u\right\}$ is contained in a set measurable with respect to the σ-algebra generated by the random variables S_t, $t \in T$, and has probability zero. In such a case we can define also the probability $P\left(\sup\limits_{t\in T} S_t > u\right)$ as $P(A_u)$. This approach led me to the definition of countable approximable classes of random variables. Its validity enables us to speak about the probability of the event that the supremum of the random variables we are interested in is larger than some fixed value. I also proved a simple but useful result in Lemma 4.3, which provides a condition for the validity of this property.

The problem we met here is not an abstract, technical difficulty. Indeed, the distribution of such a supremum can become different if we modify each random variable at a set of probability zero, although the joint distribution of the random variables we consider remains the same after such an operation. Hence, if we are interested in the supremum of a non-countable set of random variables with described joint distribution we have to describe more explicitly which version of this set of random variables we consider. It is natural to look for such an appropriate version of the random field S_t,

$t \in T$, whose 'trajectories' $S_t(\omega)$, $t \in T$, have nice properties for all elementary events $\omega \in \Omega$. Lemma 4.3 can be interpreted as a result in this spirit. The condition given for the countable approximability of a class of random variables at the end of this lemma can be considered as a smoothness condition about the 'trajectories' of the random field we consider. This approach shows some analogy to some important problems in the theory of stochastic process when a regular version of a stochastic process is considered and the smoothness properties of its trajectories are investigated.

In our problems the version of the set of random variables S_t, $t \in T$, we shall work with appears in a simple and natural way. In these problems we have finitely many random variables $\xi_1, \dots, \xi_n$ at the start, and all random variables $S_t(\omega)$, $t \in T$, we are considering can be defined individually for each ω as a functional of these random variables $\xi_1(\omega), \dots, \xi_n(\omega)$. We take the version of the random field $S_t(\omega)$, $t \in T$, we get in such a way and want to show that it is countably approximable. In Section 4 we have proved this property in an important model, probably in the most important model in possible applications we are interested in. In more complicated situations when our random variables are defined not as a functional of finitely many sample points, for instance in the case when we define our set of random variables by means of integrals with respect to a Gaussian field it is harder to find the right regular version of our sets of random variables. In this case the integrals we consider are defined only with probability 1, and we have to make some extra work to find their right version. At any rate, in the problems we are interested in our approach is satisfactory for our purposes, and it is simpler than that of Dudley; we do not have to follow his rather difficult technique. On the other hand, I must admit that I do not know the precise relation between the approach of this work and that of Dudley.

In Section 4 the notion of L_p-dense classes, $1 \le p < \infty$, is also introduced. The notion of L_2-dense classes plays an important role in the formulation Theorems 4.1 and 4.1$'$. The notion of L_2-dense classes can be considered as a version of the ε-entropy discussed at many places in the literature. On the other hand, there seems to be no unique definition of ε-entropy in the literature. I introduced the term of L_2-dense classes, because this seems to be the appropriate notion in the study of this work. To apply the results related to L_2-dense classes we also need some knowledge about how to check it in concrete models. For this goal I discussed here Vapnik–Červonenkis classes, a popular and important notion of modern probability theory. Several books and papers, (see e.g. the books [11], [28], [30] and the references in them) deal with this subject. An important result in this field is Sauer's lemma, (Lemma 5.2) which together with some other results, like Lemma 5.3 imply that the classes of sets or functions are in many several interesting models Vapnik–Červonenkis classes.

I put these results to the Appendix, partly because they can be found in the literature, partly because in our investigation Vapnik–Červonenkis classes play a different and less important role than at other places. In our discussion Vapnik–Červonenkis classes are applied to show that certain classes of functions are L_2-dense. A result of Dudley formulated in Lemma 5.2 implies that a Vapnik–Červonenkis class of functions with absolute value bounded by a fixed constant is an L_1, hence also an L_2-dense class of functions. The proof of this important result which seems to be less known even among experts of this subject than it should be is contained in the main text. Dudley's original result was formulated in the special case when the functions we consider are indicator functions of some sets, but its proof contained all important ideas needed in the proof of Lemma 5.2.

Theorem 4.2, which is the Gaussian counterpart of Theorems 4.1 and 4.1$'$ is proved in Section 6 by means of a natural and important technique, called the chaining argument. We apply an inductive procedure, during which an appropriate sequence of finite subsets of our set of random variables is defined, and try to give a good estimate on the supremum of these subsets of our random variables. The subsets we consider are denser and denser subsets of the original set of random variables, and if they are constructed in a clever way, then we get the result we want to prove by means of a limiting procedure. In such a way we get a relatively simple proof of Theorem 4.2, but this method is not strong enough to supply a complete proof of Theorem 4.1. The cause of the weakness of the method in this case is that we cannot give a good estimate on the probability that a sum of independent random variables is greater than a prescribed value if these random variables have too small variances. The chaining argument supplies a result much weaker than that what we want to prove under the conditions of Theorem 4.1. Lemma 6.1 contains the result the chaining argument yields under the conditions of Theorem 4.1. In Section 6 still another result, Lemma 6.2 is formulated, and it is also shown that Lemmas 6.1 and 6.2 together imply Theorem 4.1. The proof is not difficult, despite of some non-attractive details. We have to check that the parameters in Lemmas 6.1 and 6.2 can be fitted to each other.

Lemma 6.2 is proved in Section 7. It is based on a symmetrization argument. This proof applies the ideas of a paper of Kenneth Alexander [1], and although its presentation is essentially different of Alexander's approach, it can be considered as a version of his proof.

A similar problem should also be mentioned at this place. M. Talagrand wrote a series of papers about concentration inequalities, and this research was also continued by some other authors. I would mention the works of M. Ledoux [16] and P. Massart [26]. Concentration inequalities give a bound about the difference of the supremum of a set of appropriately defined random variables from its expected value; they express how strongly this supremum is concentrate around its expected value. Such results are closely related to Theorem 4.1, and the discussion of their relation deserves some attention. A typical concentration inequality is the following result of Talagrand [29].

Theorem 17.1. (Theorem of Talagrand.) *Consider n independent and identically distributed random variables $\xi_1,\dots,\xi_n$ with values in some measurable space $(X,\mathcal{X})$. Let $\mathcal{F}$ be some countable family of real-valued measurable functions of $(X,\mathcal{X})$ such that $\|f\|_\infty\le b<\infty$ for every $f\in\mathcal{F}$. Let $Z=\sup\limits_{f\in\mathcal{F}}\sum\limits_{i=1}^n f(\xi_i)$ and $v=E(\sup\limits_{f\in\mathcal{F}}\sum\limits_{i=1}^n f^2(\xi_i))$. Then for every positive number x,*

$$P(Z\ge EZ+x)\le K\exp\left\{-\frac{1}{K'}\frac{x}{b}\log\left(1+\frac{xb}{v}\right)\right\}$$

and
$$P(Z \ge EZ + x) \le K \exp\left\{-\frac{x^2}{2(c_1 v + c_2 bx)}\right\},$$
where K, K', c_1 and c_2 are universal positive constants. Moreover, the same inequalities hold when replacing Z by $-Z$.

Theorem 17.1 yields, similarly to Theorem 4.1, an estimate about the distribution of the supremum for a class of sums of independent random variables. It can be considered as a generalization of Bernstein's and Bennett's inequalities when the distribution of the supremum of partial sums is estimated. A remarkable feature of this result is that it assumes no condition about the structure of the class of functions $\mathcal{F}$ (like the condition of L_2-dense property of the class $\mathcal{F}$ imposed in Theorem 4.1.) On the other hand, the estimates in Theorem 17.1 contain the quantity $EZ = E\left(\sup\limits_{f\in\mathcal{F}} \sum\limits_{i=1}^{n} f(\xi_i)\right)$. Such an expectation of some supremum appears in all concentration inequalities. As a consequence, they are useful only if we can bound the expected value of an appropriate supremum. This is a hard question in the general case, and this is the reason why I preferred a direct proof of Theorem 4.1 without the application of concentration inequalities. Let me remark that the condition $u \ge \text{const.}\,\sigma \log^{1/2}\frac{2}{\sigma}$ with some appropriate constant which cannot be dropped from Theorem 4.1 is related to the fact that the expected value of the supremum of the normalized random sums considered in Theorem 4.1 has such a magnitude.

The main results of this work are presented in Section 8. Theorem 8.3 which contains an estimate about the distribution of a degenerate U-statistic was first proved in a paper of Giné and Arcones in [2], its equivalent version about the multiple integrals with respect to a normalized empirical measure formulated in Theorem 8.1 in my paper [19]. The equivalence of these two results is not self-evident. Later I proved an improved version of Theorem 8.3 in paper [21]. This result is formulated in Theorem 16.1, and it is also compared with Theorem 8.3. It is also explained that Theorem 16.1 could be considered the multivariate version of Bernstein's inequality with more right than Theorem 8.3. Here I omitted its proof which applies a technique (diagram formulas for the calculation of products of multiple random integrals or degenerate U-statistics) not discussed in this work. Here Theorem 8.3 was proved by means of a symmetrization argument. The explanation of such a proof was simpler in the present work, because it applies such methods which were worked out in the investigation of other problems. On the other hand, some arguments can be posed against such a proof. The application of symmetrization arguments in the proof of Theorem 8.3 also has some drawbacks. In certain problems, like the problem of Theorem 8.3, this method cannot supply a really sharp result. Some mathematicians working in this field seem not to be aware of this fact.

It may be interesting to mention that the problem of Theorem 8.3 has a natural generalization worth of a closer study. We can consider such generalized U-statistics in which the underlying random variables $\xi_1, \dots, \xi_n$ are independent, but they need not be identically distributed, and the U-statistic also may have a more general form. Namely, we can take a class of kernel functions $\mathbf{f} = \{f_{l_1,\dots,l_k}(x_1,\dots,x_k)\}$ on the space $(X^k, \mathcal{X}^k)$ with such an indexation that $1 \le l_j \le n$, $1 \le j \le k$, and $l_j \ne l_{j'}$ if $j \ne j'$, and define with the help of these independent random variables and class of kernel functions the generalized U-statistic

$$I_{n,k}(\mathbf{f}) = \sum_{\substack{1 \le l_j \le n,\ 1 \le j \le k \\ l_j \ne l_{j'} \text{ if } j \ne j'}} f_{l_1,\dots,l_k}(\xi_{l_1},\dots,\xi_{l_k}). \tag{17.1}$$

One can also naturally define generalized degenerate U-statistics. We call a generalized U-statistic degenerate if for all sets of indices $(l_1,\dots,l_k)$ in the sum (17.1) and for all $1 \le j \le k$

$$E(f_{l_1,\dots,l_k}(\xi_{l_1},\dots,\xi_{l_k})|\xi_{l_s},\ s \in \{1\dots,k\} \setminus \{j\}) \equiv 0.$$

Generalized degenerate U-statistics can be considered as the natural multivariate generalizations of sums of independent random variables, just as degenerate U-statistics are the natural multivariate generalizations of sums of iid. random variables. One would also try to generalize Theorem 8.3 to an estimation about the distribution of generalized degenerate U-statistics. One may hope that the method of proof of Theorem 8.3 can also be applied for the study of generalized degenerate U-statistics, just as the distribution of sums independent random variables can be investigated similarly to the sums of iid. random variables. Probably, the methods worked out for the study of the problems related to Theorem 8.3 are helpful, but in the study of generalized degenerate U-statistics first some special questions have to be clarified. We have to find the right form of the estimation about the distribution of a generalized degenerate U-statistic. In particular, it must be clarified which are the natural quantities by which we should express this estimate.

It is natural to expect that generalized degenerate U-statistics $I_{n,k}(\mathbf{f})$ of order k (without normalization) satisfy the inequality

$$P(|I_{n,k}(\mathbf{f})| > u) < A \exp\left\{-C\left(\frac{u}{V_n}\right)^{2/k}\right\} \tag{17.2}$$

with some universal constants $A = A(k) > 0$ and $C = C(k) > 0$ in a relatively large interval for the parameter u, where V_n^2 denotes the variance of $I_{n,k}(\mathbf{f})$. An essential problem is to find a relatively good constant C and to determine the interval $0 < u < D_n$, where the estimate (17.2) holds. Theorem 8.3 states that in the case of classical degenerate U-statistics (17.2) holds in the interval $[0, D_n]$ with $D_n = \text{const.}\, n^k\sigma^{k+1}$, where $\sigma^2 = Ef(\xi_1,\dots,\xi_k)^2$. For $k = 1$ this means that relation (1.9) holds in the interval $0 \le u \le V_n^2$. But it is not clear what corresponds in the case of generalized degenerate U-statistics to the right end-point $D_n = \text{const.}\, n^k\sigma^{k+1}$ of the interval where the estimate (17.2) should hold. (The variance of a degenerate U-statistic of order k is of order $n^k\sigma^2$.)

Theorems 8.2 and 8.4 yield an estimate about the supremum of (degenerate) U-statistics or of multiple random integrals with respect to a normalized empirical measure when the class of kernel functions in these U-statistics or random integrals satisfy some conditions. They were proved in my paper [20]. Earlier Arcones and Giné proved a weaker form of this result in paper [3]. The Gaussian version of Theorem 8.1 or 8.3 given in Theorem 8.5 was proved much earlier. My lecture note [17] also contains a proof of this result. The second statement of Theorem 8.5 about the supremum of Wiener–Itô integrals can be simply proved. Section 8 also contains an example which shows in particular that the probability $P\left(n^{-1}I_{n,2}(f)>u\right)$ can be bounded for a degenerate U-statistic $I_{n,2}(f)$ of order 2 by the estimate given in Theorem 8.3 only if $u\le \text{const.}\, n\sigma^3$, i.e. this condition of Theorem 8.3 (in the case $k=2$) cannot be dropped. Similar examples could be constructed for all $k\ge 1$. The paper of Arcones and Giné [2] contains another example explained by Talagrand to the authors which also has a similar consequence.

On the other hand, this example does not exclude the possibility to prove such a multi-dimensional version of Hoeffding's inequality Theorem 3.3 which provides a slight improvement of Theorems 8.1 and 8.3 similarly to the improvement of Bernstein's inequality provided by Hoeffding's inequality. Moreover, we can also expect such a strengthened form of Theorems 8.2 and 8.4 (or of Theorem 4.2 in the one-dimensional case) which takes into account the above improvements if the supremum of a nice class of random integrals or degenerated U-statistics is considered. There is a hope that some refinement of the methods of the present work would supply such results. However, here we did not study this problem.

Theorems 9.2 and 9.3 deal with the properties of degenerate U-statistics. This subject deserves special attention. Degenerate U-statistics can be considered as the multivariate version of sums of independent and identically distributed random variables with expectation zero. Similarly, if f is a canonical function with respect to a measure μ and put independent μ-distributed random variables into its arguments, then the random variables we get in such a way can be considered as the multivariate version of random variables with expectation zero. The background of several proofs about the behaviour of U-statistics can be better understood with the help of the above remark. I tried to explain for instance that the proof about the Hoeffding decomposition of U-statistics (Theorem 9.1) is actually a natural adaptation of the decomposition of a random variable to the sum of a random variable with expectation zero plus the expected value of the random variable.

Hoeffding's decomposition is a fairly well-known result which can be found for instance in the Appendix of [12]. Theorem 9.1 slightly differs from the formulation of Hoeffding's decomposition one usually meets in the literature. It can be exploited that a U-statistic does not change if we replace its kernel function by its symmetrized version. Besides, the value of the U-statistics $I_{n,|V|}(f_V)$ do not change if we replace the kernel function $f_V(x_{j_1},\dots,x_{j_{|V|}})$, $V=\{j_1,\dots,j_{|V|}\}$, by $f_V(x_1,\dots,x_{|V|})$ in the Hoeffding decomposition (9.3) of the U-statistic $I_{n,k}(f)$, and $f_V(x_1,\dots,x_{|V|})$ is also a canonical function. The above observations enable us to unify the contribution of all terms $I_{n,|V|}(f_V)$ with $|V|=l$ for some $0\le l\le k$ into one non-degenerate U-statistics of order l. Generally, the formula obtained in such a way is called the Hoeffding decomposition in the literature. Nevertheless, we have applied Theorem 9.1 in this work, because this form of the Hoeffding's decomposition was more convenient for us.

In our investigations it is important to know that if a function satisfies a good L_2-norm or L_∞-norm estimate, then the elements of its Hoeffding decomposition also have this property, and if a class of function is L_2-dense, then the same relation holds for the classes of functions in the Hoeffding decomposition of the functions in this class. This is the content of Propositions 9.2 and 9.3. The estimates on the L_2-norm given in formulas (9.7) and (9.8) are actually reformulations of some well-known facts about the properties of conditional expectations.

Theorem 9.4 enables us to reduce the estimates about multiple random integrals with respect to normalized empirical measures to estimates about degenerate U-statistics. Such random integrals are actually sums of U-statistics, and we can apply for each of these U-statistics the Hoeffding decomposition. Besides, as we consider multiple integrals with respect to a *normalized* empirical measure we can expect that a lot of cancellations appear during the calculation by which we express our random integral in the form of linear combination of degenerate U-statistics. We get such a representation which enables us to reduce the estimates we want to prove about multiple random integrals to analogous estimates about degenerate U-statistics. This is the main content of Theorem 9.4 which can be considered as an analog of the Hoeffding decomposition for multiple stochastic integrals with respect to normalized empirical measures. This representation of a multiple stochastic integral as a linear combination of degenerate U-statistics of different order also contains degenerate U-statistics of low order. But as a consequence of the cancellation effects these U-statistics are multiplied with small coefficients. The proof of Theorem 9.4 is based on a good "book-keeping" of the different contributions to the integral $J_{n,k}(f)$. An essential, although less spectacular step of this "book-keeping" procedure is to express the terms we are working with by means of the (signed) measures μ and $\mu^{(l)} - \mu$, i.e. the measures $\mu^{(l)}$ have to be replaced by their normalizations $\mu^{(l)} - \mu$. The calculations needed in the proof are quite natural, but unfortunately they contain some unpleasant and complicated technical details.

Theorem 9.4 also has the consequence that the second moment of the multiple random integral of a function with respect to a normalized empirical measure can be bounded by constant times the L_2-norm of the kernel function we integrate. The representation of the stochastic integrals given in Theorem 9.4 may also contain a non-zero constant term. This has the unexpected consequence that the expected value of a multiple random integral with respect to a normalized empirical measure can be non-zero. Our random integrals may show such an unusual behaviour because the numbers of sample points falling to disjoint sets are not independent random variables. But the dependence between such random variables is very weak, and the expected value of the random integrals we consider is sufficiently small.

From the pair of Theorems 8.1 and 8.3 I have proved only Theorem 8.3, since its proof is simpler, and by the results of Section 9 Theorem 8.1 follows from it. The proof of Theorem 8.3 is different from its original proof published in paper [2]. First a good estimate is presented about the moments of the degenerate U-statistics in Proposition 10.1. Theorem 8.3 can be deduced from this estimate. Actually the proof is different, first a version Theorem 8.3$'$ of Theorem 8.3 is proved, where an analogous estimate is proved for degenerate decoupled U-statistic. The adjective 'decoupled' refers to the fact that we put independent copies of a sequence of iid. random variables in different coordinates of the kernel function of the U-statistic. The study of decoupled U-statistics is a popular subject of some authors. In particular, the main subject of the book [6] is a comparison of the properties of U-statistics and decoupled U-statistics.

The study of decoupled U-statistics is simpler than that of usual U-statistics, because the arguments applied in the study of usual U-statistics can be applied for them, and they also satisfy a multivariate version of the Marcinkiewicz–Zygmund inequality. On the other hand, the Marcinkiewicz–Zygmund inequality does not hold for usual U-statistics, at least the proofs I know of do not work for them. We can prove with the help of the multivariate version of the Marcinkiewicz–Zygmund and Borell's inequality an estimate about the moments of degenerate U-statistics formulated in Proposition 10.1′. Proposition 8.3′ can be deduced from Proposition 10.1′, and by a result of de la Peña and Montgomery–Smith formulated in Theorem 10.4 Theorem 8.3′ implies Theorem 8.3. The results applied in the proof of Theorem 8.3 are proved in Section 11. Let me also remark that Proposition 10.1 is not proved in this text, since we chose such an approach where we do not need it. On the other hand, it follows from the results of this work and some other standard results about U-statistics not discussed in the present work.

I have mentioned the possibility of another proof of Theorem 8.3 on the basis of the methods of the theory of Wiener–Itô integrals to this problem. In [19] I gave a proof of Theorem 8.1 by means of the so-called diagram method. Let me also remark that the method of paper [21] which yields an improvement of Theorem 8.3 presented in Theorem 16.1 is actually a refinement of the method in [19]. Both in paper [19] and in the present work the main step of the proof consists of finding a good estimate on the moments of the random variables we are investigating. It is enough to estimate the moments of the type $M = 2^m$, where m is a positive integer. For $m = 1$ such an estimate is known, and we can get an estimate for $m > 1$ by means of a recursive procedure. A similar approach is applied in [19] and in the present work. The main difference between them is in the form of the recursive inequality between the moments of the random variables we work with and the way we prove them.

I found the result about the multivariate version of the Marcinkiewicz–Zygmund equation in the book [6], but the proof of the result given here is different. Only the proof about the upper estimate of the p-th moment of decoupled U-statistics is written down. There is also an estimate in the opposite direction, but such a result would be interesting for us only for the sake of some orientation. Theorem 10.4 was proved by de la Peña and Montgomery–Smith in their paper [7]. I formulated their result for separable Banach space valued random variables, just as they did it. Such a general formulation of the results is very popular in the literature, but here the discussion of Banach space valued random variables had a different cause. I also wanted to prove formula (10.8′), a result which is actually not contained in paper [7]. (Book [6] contains this result, but the proof is left to the reader.) The simplest way to get this statement was to prove the original result in Banach spaces, and to apply it in appropriate L_∞ spaces. Paper [7] also contains some kind of converse result of Theorem 10.4, but as we do not need it I omitted its discussion.

This work contains the proof of de la Peña and Montgomery–Smith for Theorem 10.4, but I have explained it in my own style. In particular, I worked out some details where the author gave only a very short explanation. This proof is given in the Appendix.

The proof of Borell's inequality is closely related to that of Nelson's inequality. Edward Nelson published the inequality named after him in his paper [27]. He also showed that the general inequality presented in Appendix C can be reduced to the inequality given in formula (C1) or in Proposition C2 of this work. This reduction follows actually from our Theorem 11.2. However, this observation did not help him to find a proof, and finally he gave a proof without its application. Borell's inequality can also be reduced to a one-dimensional statement formulated in Theorem 11.3. This seems to be a simple inequality, but its proof is surprisingly hard. Actually in this paper it is enough to prove this inequality in the special case $q = 2$ and $p = 2k$, $k = 1, 2, \dots$. Actually, as I mentioned in Theorem 16.6, Borell's inequality can be proved in this special case with better constants. (See paper [22].)

In the proof of Theorem 11.3 I have followed the paper of Leonhard Gross *Logarithmic Sobolev inequalities* [13]. Gross has worked out a general theory and he could prove both Nelson's and Borell's inequality (more precisely an estimate which simply implies this result) with its help. Gross' method and results are interesting, because they are very useful in several parts of the mathematics. (See e.g [16] or [14].) Let me also remark that similar results and ideas also appeared in an earlier work of A. Bonami [5].

Gross introduced a so-called logarithmic Sobolev inequality related to Markov processes and showed that it implies another inequality, which is in the case of a Wiener process Nelson's inequality, while we can define such a simple Markov process for which the logarithmic Sobolev inequality corresponding to it yields the proof of Theorem 11.3. This Markov process is explicitly described in Section 11, and the logarithmic Sobolev inequality corresponding to it is also formulated and proved there. Actually Gross showed that each logarithmic Sobolev inequality is equivalent to the inequality he proved as its consequence. On the other hand, the proof of the logarithmic Sobolev inequalities is less difficult than a direct proof of the inequalities he has obtained as their consequence.

The name 'logarithmic Sobolev inequality' has the following explanation. Generally one calls 'Sobolev inequality' such inequalities where for some pairs of numbers $1 \le q < p < \infty$ we prove a bound on the L_p-norm of a function if we have an estimate on its L_q-norm together with the L_q-norm of some partial derivatives of this function. In the logarithmic Sobolev inequalities the integral of a function of the form $|f|^p \log |f|$ is bounded by means of the integral of $|f|^p$ and the integral of a differential type operator of this function f which is closely related to the infinitesimal operator of a Markov process.

The proof of Borell's inequality presented here is due to Leonhard Gross. We have also shown in the Appendix that from this estimate and the central limit theorem Nelson's inequality can be deduced. In this proof we have applied some basic facts about Wiener–Itô integrals which we did not discuss in detail. The most important results we have used here are the so-called Itô's formula for Wiener–Itô integrals and the diagram formula. All these results can be found in my lecture note [17]. Borell's inequality was applied in the proof of Theorem 8.3′. We also proved another result with its help which plays an important role in our study. This is the multivariate version of Hoeffding's inequality in Theorem 12.3. This result is a simple consequence of Borell's inequality, but I did not find it in the literature. Paper [22] contains an improved version of this estimates presented in Theorem 16.6.

Sections 12 — 15 deal with the proof of Theorem 8.4 about the tail-behaviour of the supremum of a class of degenerate U-statistics under appropriate conditions. This result was proved in my paper [20]. The proof of this result is similar to that of its one-variate version Theorem 4.1, but some additional difficulties have to be overcome. We have formulated some results in Propositions 12.1 and 12.2 which are the multivariate analogs of Propositions 6.1 and 6.2, and Theorem 8.4 can be proved as their consequence. Proposition 12.1 can be proved similarly to Proposition 6.1, and also the deduction of Theorem 8.4 from Propositions 12.1 and 12.2 is similar to the argument applied in the proof of Theorem 4.1.

The hard part of the problem is to prove Proposition 12.2. By means of the results of de la Peña and Montgomery–Smith it can be reduced to a version formulated in Proposition 12.2′, where degenerate U-statistics are replaced by degenerate decoupled U-statistics. This result is proved by means of a refinement of the argument of the proof of Proposition 6.2. The main difficulty appears as we want to find the multivariate analog of the symmetrization argument made by means of the Symmetrization Lemma, Lemma 7.1 and Lemma 7.2 in the one-variate case. In the proof of Theorem 4.1 we could carry out a symmetrization procedure by investigating the difference of two independent copies of the random sums we have considered. In the proof of Proposition 12.2′ a more sophisticated construction has to be applied. This construction actually appeared in the proof of Theorem 8.3, and Lemma 11.5 explains its most important properties.

In the proof of Proposition 12.2′ Lemma 7.1 is not sufficient for us in its original form. We need a generalization of this result, and this is given in Lemma 13.1. The proof of Lemma 13.1 is not hard. The real difficulty arises when we want to apply it in our case. Then as we want to check formula (13.1) we have to bound some non-trivial conditional probabilities. In the analog relation, in formula (7.1) of Lemma 7.1 it was enough to bound a usual probability, and this was simple. But as we want to adapt this method in the multivariate case we have to bound an appropriate conditional variance. This demands much more work, and the hardest new steps of the proof were introduced to overcome this difficulty.

Proposition 12.2′ was proved by means of an inductive procedure formulated in Proposition 13.2, which is the multivariate analog of Proposition 6.2. But because of the problems we meet in carrying out the symmetrization procedure the arguments of Proposition 7.2 are not sufficient in this case. Hence another statement is introduced in Proposition 13.3. Propositions 13.2 and 13.3 can be proved simultaneously by means of an appropriate inductive procedure. The proof is based on a refinement of the arguments in the proof of Proposition 6.2. We also have to exploit our knowledge about the properties of Hoeffding's decomposition.

Appendix A.

The proof of some results about Vapnik–Červonenkis classes

Proof of Theorem 5.1. (Sauer's lemma) Let $F_1,\dots,F_m$ be the subsets of cardinality k of the set $S_0(n)$, $m=\binom nk$. By the conditions of the theorem all sets F_j, $1\le j\le m$, have a "hidden" subset $H_j\subset F_j$ such that the class of sets $\mathcal{D}(S_0,F_j)=\{F_j\cap B;\, B\in\mathcal{D}(S_0)\}$ does not contain the set H_j. Let us denote by $\mathcal{C}_0=\mathcal{C}_0((F_1,H_1),\dots,(F_m,H_m))$ the class of subsets of $S_0(n)$ we get by taking first all subsets of $S_0(n)$ and then omitting all subsets of the form $H_j\cup G_j$ with some $G_j\subset S_0\setminus F_j$, $1\le j\le m$. The subsets omitted in the definition of $\mathcal{C}_0$ do not belong to $\mathcal{D}(S_0)$, thus $\mathcal{C}_0$ contains all elements of $\mathcal{D}(S_0)$, and it is enough to show that $\mathcal{C}_0$ contains no more than $\binom n0+\binom n1+\cdots+\binom n{k-1}$ subsets of $S_0(n)$. If $H_j=F_j$ for all "hidden" subsets H_j, $1\le j\le m$, then $\mathcal{C}_0$ contains the subsets of $S_0(n)$ with cardinality at most $k-1$, and we have to show that this is the extreme case.

Let us choose some element $s\in S_0$, and define similarly to the class $\mathcal{C}_0$ a new class $\mathcal{C}_1=\mathcal{C}_1((F_1,\bar H_1),\dots,(F_m,\bar H_m))$ with the difference that instead of the "hidden" subsets H_j of F_j taking part in the definition of $\mathcal{C}_0$ we work with the sets $\bar H_j$ we get by augmenting H_j with the element s if it is possible, i.e. in the definition of $\mathcal{C}_1$ H_j is replaced by $\bar H_j=(H_j\cup\{s\})\cap F_j$. Given a set $B\subset S_0$ we can say that $B\in\mathcal{C}_0$ if and only if $B\cap F_j\ne H_j$ for all $1\le j\le m$, and $B\in\mathcal{C}_1$ if and only if $B\cap F_j\ne\bar H_j$ for all $1\le j\le m$. We want to show that $\mathcal{C}_1$ has more elements than $\mathcal{C}_0$. Theorem 5.1 can be deduced from this statement, because by iterating this procedure for enlarging the "hidden" subsets H_j of the sets F_j for all $s\in S_0$ we get that the class $\mathcal{C}_0$ has the greatest cardinality in the case when $H_j=F_j$ for all $1\le j\le k$.

Let us define the map $T(B)=B\setminus\{s\}$ for all sets $B\subset S_0(n)$. We shall show that $T(\cdot)$ is an injection of $\mathcal{C}_0\setminus\mathcal{C}_1$ to $\mathcal{C}_1\setminus\mathcal{C}_0$. This implies that the cardinality of $\mathcal{C}_1$ is larger than that of $\mathcal{C}_0$ just as we claimed. To prove the above property of $T(\cdot)$ first we check that a) if $B\in\mathcal{C}_0\setminus\mathcal{C}_1$ then $s\in B$. This implies that different elements of $\mathcal{C}_0\setminus\mathcal{C}_1$ have different images under the map T. We also check that b) if $B\in\mathcal{C}_0\setminus\mathcal{C}_1$, then $T(B)\in\mathcal{C}_1\setminus\mathcal{C}_0$, i.e. b1) $T(B)\in\mathcal{C}_1$ and b2) $T(B)\notin\mathcal{C}_0$.

If $B\in\mathcal{C}_0\setminus\mathcal{C}_1$ then $B\cap F_j\ne H_j$ for all $1\le j\le m$, and $B\cap F_j=\bar H_j$ for some j. This means that $B\cap F_j\ne H_j$ and $B\cap F_j=\bar H_j$ for some index j. This is only possible if $s\notin H_j$, $s\in F_j$ and $s\in B$, i.e. property a) holds. Besides, $T(B)\cap F_j=\bar H_j\setminus\{s\}=H_j$ for such an index j which means that property b2) holds. To check property b1) we have to show that if $B\in\mathcal{C}_0\setminus\mathcal{C}_1$, then $(B\setminus\{s\})\cap F_j\ne\bar H_j$ for all $1\le j\le m$. This relation clearly holds for such indices j for which $s\in F_j$, since in this case $s\in\bar H_j$. If $s\notin F_j$, then the condition $B\in\mathcal{C}_0$ implies that $B\cap F_j\ne H_j$, and $\bar H_j=H_j$ and $B\cap F_j=(B\setminus\{s\})\cap F_j$ because of the relation $s\notin F_j$. These relations imply that $(B\setminus\{s\})\cap F_j\ne\bar H_j$ also in this case.

The proof of Theorem 5.3 Let us fix an arbitrary set $F=\{s_1,\dots,s_{k+1}\}$ of the set S, and consider the set of vectors $\mathcal{G}_k(F)=\{(g(s_1),\dots,g(s_{k+1}))\colon g\in\mathcal{G}_k\}$ of the $k+1$-dimensional space R^{k+1}. By the conditions of the Theorem $\mathcal{G}_k(F)$ is an at most k-dimensional subspace of R^{k+1}. Hence there exists a non-zero vector $a=(a_1,\dots,a_{k+1})$

such that $\sum\limits_{j=1}^{k+1} a_j g(s_j)=0$ for all $g\in\mathcal{G}_k$. We may assume that $A=A(a)=\{j\colon a_j<0, 1\le j\le k+1\}$ is a non-empty set by multiplying the vector a by -1 if it is necessary.

Thus we can write

$$\sum_{j\in A} a_j g(s_j)=\sum_{j\in\{1,\dots,k+1\}\setminus A}(-a_j)g(s_j), \qquad \text{for all } g\in\mathcal{G}_k. \tag{A1}$$

Put $B=\{s_j, j\in A\}$. Then $B\subset F$, and we claim that $B\ne\{g\colon g(s)\ge 0\}\cap F$ for all $g\in\mathcal{G}_k$. Indeed, if there were some $g\in\mathcal{G}_k$ such that $B=\{g\colon g(s)\ge 0\}\cap F$, then the left-hand side of the equation (A1) would be strictly positive and its right-hand side would be non-positive for this $g\in\mathcal{G}_k$, and this is a contradiction.

Thus Theorem 5.1 implies that for all subsets $S_0(n)$ of S with $n\ge k+1$ elements and the class of subsets $\mathcal{D}$ of S introduced in the formulation of Theorem 5.3 $S_0(n)\cap\mathcal{D}$ has at most $\binom{n}{0}+\binom{n}{1}+\cdots+\binom{n}{k}$ elements. Hence $\mathcal{D}$ is a Vapnik–Červonenkis class.

Appendix B. The proof of Theorem 10.3

(A result of de le Peña and Montgomery–Smith)

The proof of Theorem 10.3. We concentrate our efforts to prove relation (10.8). Formula (10.8′) can be obtained as a relatively simple consequence of this result. The proof of formula (10.8) will be made by means of an inductive procedure. To carry out it we have to formulate and prove our statement in a more general form where such generalized U-statistics are considered for which different kernel functions may appear in each term of the sum. More explicitly, let $\ell=\ell(n,k)$ denote the set of all sequences $l=(l_1,\dots,l_k)$ of length k such that $1\le l_j\le n$, $1\le j\le k$. Let us fix a class of functions $\{f_{l_1,\dots,l_k}(x_1,\dots,x_k), (l_1,\dots,l_k)\in\ell\}$ which map the space $(X^k,\mathcal{X}^k)$ to a separable Banach space B. Let us denote this class of functions by $f(\ell)$, and define similarly to the U-statistics and decoupled U-statistics the generalized U-statistics and generalized decoupled U-statistics by the formulas

$$I_{n,k}(f(\ell))=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'} \text{ if } j\ne j'}} f_{l_1,\dots,l_k}\left(\xi_{l_1},\dots,\xi_{l_k}\right)$$

and

$$\bar I_{n,k}(f)=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'} \text{ if } j\ne j'}} f_{l_1,\dots,l_k}\left(\xi^{(1)}_{l_1},\dots,\xi^{(k)}_{l_k}\right)$$

(with the same random variables ξ_l and $\xi_l^{(j)}$, $1\le l\le n$, $1\le j\le k$ as before.)

The following generalization of relation (10.8) will be proved.

$$P\left(\|I_{n,k}(f(\ell))\|>u\right)\le AP\left(\|\bar I_{n,k}(f(\ell))\|>\gamma u\right) \tag{10.8b}$$

with some constants $A = A(k)$ and $\gamma = \gamma(k)$ depending only on the order of these U-statistics.

To prove relation (10.8b) first we verify the following statement.

Let us take two independent copies $\xi_1^{(1)},\dots,\xi_n^{(1)}$ and $\xi_1^{(2)},\dots,\xi_n^{(2)}$ of our original sequence of random variables $\xi_1,\dots,\xi_n$ and introduce for all sets $V\subset\{1,\dots,k\}$ the function $\alpha_V(j)$, $1\le j\le k$, defined as $\alpha_V(j)=1$ if $j\in V$ and $\alpha_V(j)=2$ if $j\notin V$. Let us define with the help of these quantities the decoupled generalized U-statistics

$$I_{n,k,V}(f(\ell))=\frac{1}{k!}\sum_{\substack{1\le l_j\le n,\; j=1,\dots,k\\ l_j\ne l_{j'} \text{ if } j\ne j'}} f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(\alpha_V(1))},\dots,\xi_{l_k}^{(\alpha_V(k))}\right)\quad\text{for all } V\subset\{1,\dots,k\}. \tag{B1}$$

The following inequality will be proved: There are some constants $C_k>0$ and $D_k>0$ depending only on the order k of the generalized U-statistic $I_{n,k}(f(\ell))$ such that for all numbers $u>0$

$$P\left(\|I_{n,k}(f(\ell))\|>u\right)\le\sum_{V\subset\{1,\dots,k\},\,1\le|V|\le k-1} C_kP\left(D_k\|I_{n,k,V}(f(\ell))\|>u\right). \tag{B2}$$

Here $|V|$ denotes the cardinality of the set V, and the condition $1\le|V|\le k-1$ in the summation of formula (B2) means that we omit the sets $V=\emptyset$ and $V=\{1,\dots,k\}$ from the summation, i.e. the cases when either $\alpha_V(j)=1$ for all $1\le j\le k$ or $\alpha_V(j)=2$ for all $1\le j\le k$ are not considered in this sum. Formula (10.8b) can be deduced from formula (B2) by means of a relatively simple inductive argument. In the proof of formula (B2) we shall apply the following simple lemma.

Lemma B1. *Let ξ and η be two independent and identically distributed random variables taking values on a separable Banach space B. Then*

$$3P\left(|\xi+\eta|>\frac{2}{3}u\right)\ge P(|\xi|>u)\quad\textit{for all } u>0.$$

Proof of Lemma B1. Let ξ, η and ζ three independent, identically distributed random variables taking values in B. Then

$$\begin{aligned}3P\left(|\xi+\eta|>\frac{2}{3}u\right)&=P\left(|\xi+\eta|>\frac{2}{3}u\right)+P\left(|\xi+\zeta|>\frac{2}{3}u\right)+P\left(|-(\eta+\zeta)|>\frac{2}{3}u\right)\\&\ge P(|\xi+\eta+\xi+\zeta-\eta-\zeta|>2u)=P(|\xi|>u).\end{aligned}$$

To prove formula (B2) let us introduce the random variable

$$T_{n,k}(f(\ell))=\frac{1}{k!}\sum_{\substack{1\le l_j\le n, s_j=1 \text{ or } s_j=2,\; j=1,\dots,k\\ l_j\ne l_{j'} \text{ if } j\ne j'}} f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(s_1)},\dots,\xi_{l_k}^{(s_k)}\right)=\sum_{V\subset\{1,\dots,k\}}\bar I_{n,k,V}(f(\ell)), \tag{B3}$$

and observe that the random variables $I_{n,k}(f(\ell))$, $I_{n,k,\emptyset}(f(\ell))$ and $I_{n,k,\{1,\dots,k\}}(f(\ell))$ are identically distributed and the last two random variables are independent of each other. Hence Lemma B1 yields that

$$\begin{aligned}P(\|I_{n,k}(f(\ell))\|>u)&\le 3P\left(\|I_{n,k,\emptyset}(f(\ell))+I_{n,k,\{1,\dots,k\}}(f(\ell))\|>\frac23u\right)\\&=3P\left(\left\|T_{n,k}(f(\ell))-\sum_{V\colon\, V\subset\{1,\dots,k\},\,1\le|V|\le k-1}I_{n,k,|V|}(f(\ell))\right\|>\frac23u\right)\\&\le P(3\cdot2^{k-1}\|T_{n,k}(f(\ell))\|>u) \qquad\qquad (B4)\\&\quad+\sum_{V\colon\, V\subset\{1,\dots,k\},\,1\le|V|\le k-1}P(3\cdot2^{k-1}\|I_{n,k,|V|}(f(\ell))\|>u).\end{aligned}$$

To deduce relation (B2) from relation (B4) we need a good estimate on the probability $P(3\cdot2^{k-1}\|T_{n,k}(f(\ell))\|>u)$. We shall compare the distribution of $\|T_{n,k}(f(\ell))\|$ with that of $\|I_{n,k,V}(f(\ell))\|$ for an arbitrary set $V\subset\{1,\dots,k\}$ and get an estimate which is sufficient to prove relation (B2). To carry out this program first we prove the following lemmas.

Lemma B2. *Let us consider a sequence of independent random variables $\varepsilon_1,\dots,\varepsilon_n$, $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac12$, $1\le l\le n$, which is also independent of the random variables $\xi_1^{(1)},\dots,\xi_n^{(1)}$ and $\xi_1^{(2)},\dots,\xi_n^{(2)}$ appearing in the definition of the decoupled U-statistics $I_{n,k,V}(f(\ell))$ defined in formula (B1). Let us define with their help the sequences of random variables $\eta_1^{(1)},\dots,\eta_n^{(1)}$ and $\eta_1^{(2)},\dots,\eta_n^{(2)}$ whose elements $(\eta_l^{(1)},\eta_l^{(2)})=(\eta_l^{(1)}(\varepsilon_l),\eta_l^{(2)}(\varepsilon_l))$, $1\le l\le n$, are given as*

$$(\eta_l^{(1)}(\varepsilon_l),\eta_l^{(2)}(\varepsilon_l))=\left(\frac{1+\varepsilon_l}{2}\xi_l^{(1)}+\frac{1-\varepsilon_l}{2}\xi_l^{(2)},\frac{1-\varepsilon_l}{2}\xi_l^{(1)}+\frac{1+\varepsilon_l}{2}\xi_l^{(2)}\right),$$

i.e. let $(\eta_l^{(1)}(\varepsilon_l),\eta_l^{(2)}(\varepsilon_l))=(\xi_l^{(1)},\xi_l^{(2)})$ if $\varepsilon_l=1$, and $(\eta_l^{(1)}(\varepsilon_l),\eta_l^{(2)}(\varepsilon_l))=(\xi_l^{(2)},\xi_l^{(1)})$ if $\varepsilon_l=-1$, $1\le l\le n$. Then the joint distribution of the pair of sequences of random variables $\xi_1^{(1)},\dots,\xi_n^{(1)}$ and $\xi_1^{(2)},\dots,\xi_n^{(2)}$ agrees with that of the pair of sequences $\eta_1^{(1)},\dots,\eta_n^{(1)}$ and $\eta_1^{(2)},\dots,\eta_n^{(2)}$.

Let us fix some $V\subset\{1,\dots,k\}$, and introduce the random variable

$$\bar I_{n,k,V}(f(\ell))=\frac1{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if }j\ne j'}}f_{l_1,\dots,l_k}\left(\eta_{l_1}^{(\alpha_V(1))},\dots,\eta_{l_k}^{(\alpha_V(k))}\right),\qquad (B5)$$

where similarly to formula (B1) $\alpha_V(j)=1$ if $j\in V$, and $\alpha_V(j)=2$ if $j\notin V$. Then the identity

$$\begin{aligned}&2^k\bar I_{n,k,V}(f(\ell)) \qquad\qquad (B6)\\&\quad=\frac1{k!}\sum_{\substack{1\le l_j\le n,s_j=1\text{ or }s_j=2,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if }j\ne j'}}(1+\varepsilon^{(1)}_{l_1,s_1,V})\cdots(1+\varepsilon^{(k)}_{l_k,s_k,V})f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(s_1)},\dots,\xi_{l_k}^{(s_k)}\right)\end{aligned}$$

holds, where $\varepsilon^{(j)}_{l,1,V}=\varepsilon_l$, $\varepsilon^{(j)}_{l,2,V}=-\varepsilon_l$ *if* $j\in V$, *and* $\varepsilon^{(j)}_{l,1,V}=-\varepsilon_l$, $\varepsilon^{(j)}_{l,2,V}=\varepsilon_l$ *if* $j\notin V$, $1\le l\le n$.

In the proof of relation (B2) we need besides Lemma B2 another result given in Lemma B4. Before the formulation of Lemma B4 we present Lemma B3 whose result will be used in its proof.

Lemma B3. *Let* Z *be a random variable in a separable Banach space* B *with expectation zero, i.e. let* $E\kappa(Z)=0$ *for all* $\kappa\in B'$. *Then* $P(\|v+Z\|\ge\|x\|)\ge \inf\limits_{\kappa\in B'}\frac{(E\kappa(Z))^2}{4E\kappa(Z)^2}$ *for all* $v\in B$. *Here* B' *denotes the (Banach) space of all (bounded) linear transformations on* B *to the real line.*

Lemma B4. *Let us consider a sequence of independent random variables* $\varepsilon_1,\dots,\varepsilon_n$, $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac12$, $1\le l\le n$, *a polynomial of order* k *of these random variables with some coefficients* $a(l_1,\dots,l_s)$, $1\le s\le k$, $1\le l_s\le n$, *from some separable Banach space* B. *Let us assume that the coefficients of this polynomial satisfy the relation* $a(l_1,\dots,l_s)=0$ *if* $l_p=l_q$ *with some* $1\le p<q\le s$, *and the constant term is zero. The inequality*

$$P\left(\left\|v+\sum_{s=1}^{k}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,s\\ l_j\ne l_{j'}\text{ if }j\ne j'}}a(l_1,\dots,l_s)\varepsilon_{l_1}\cdots\varepsilon_{l_s}\right\|>\|v\|\right)\ge c_k \tag{B7}$$

holds for all $v\in B$ *with some constant* $c_k>0$ *depending only on the order* k *of this polynomial.*

The proof of Lemma B2. Let us consider the conditional joint distribution of the sequences of random variables $\eta^{(1)}_1,\dots,\eta^{(1)}_n$ and $\eta^{(2)}_1,\dots,\eta^{(2)}_n$ under the condition that the random vector $\varepsilon_1,\dots,\varepsilon_n$ takes the value of some prescribed ±1 series of length n. Observe that this conditional distribution agrees with the joint distribution of the sequences $\xi^{(1)}_1,\dots,\xi^{(1)}_n$ and $\xi^{(2)}_1,\dots,\xi^{(2)}_n$ for all possible conditions. This fact implies the statement about the joint distribution of the sequences $\eta^{(1)}_l,\eta^{(2)}_l$, $1\le l\le n$.

To prove identity (B6) let us fix a set $M\subset\{1,\dots,n\}$ and consider the case when $\varepsilon_l=1$ if $l\in M$ and $\varepsilon_l=-1$ if $l\notin M$. Observe that for all fixed sequences $1\le l_1,\dots,l_k\le n$, $l_j\ne l_{j'}$ if $j\ne j'$

$$f_{l_1,\dots,l_k}\left(\eta^{(\alpha_V(1))}_{l_1},\dots,\eta^{(\alpha_V(k))}_{l_k}\right)=f_{l_1,\dots,l_k}\left(\xi^{(\beta_{V,M}(1,l_1))}_{l_1},\dots,\xi^{(\beta_{V,M}(k,l_k))}_{l_k}\right),$$

where $\beta_{V,M}(j,l)=1$ if $j\in V$ and $l\in M$ or $j\notin V$ and $l\notin M$), and $\beta_{V,M}(j,l)=2$ otherwise. On the other hand,

$$\begin{aligned}\sum_{s_j=1\text{ or }s_j=2,\ j=1,\dots,k}&(1+\varepsilon^{(1)}_{l_1,s_1,V})\cdots(1+\varepsilon^{(k)}_{l_k,s_k,V})f_{l_1,\dots,l_k}\left(\xi^{(s_1)}_{l_1},\dots,\xi^{(s_k)}_{l_k}\right)\\ &=2^kf_{l_1,\dots,l_k}\left(\xi^{(\beta_{V,M}(1,l_1))}_{l_1},\dots,\xi^{(\beta_{V,M}(k,l_k))}_{l_k}\right),\end{aligned}$$

since the product $(1+\varepsilon^{(1)}_{l_1,s_1,V})\cdots(1+\varepsilon^{(k)}_{l_k,s_k,V})$ equals either zero or 2^k, and $\varepsilon^{(j)}_{l_j,s_j,V}=1$ if $\beta_{V,M}(j,l_j)=s_j$, and $\varepsilon^{(j)}_{l_j,s_j,V}=-1$ if $\beta_{V,M}(j,l_j)\neq s_j$.

Summing up these identities for all $1\le l_1,\dots,l_k\le n$ such that $l_j\neq l_{j'}$ if $j\neq j'$ we get identity (B6).

The proof of Lemma B3. Let us first observe that if ξ is a real valued random variable with zero expectation, then $P(\xi>0)\ge\frac{(E|\xi|)^2}{4E\xi^2}$ since $(E|\xi|)^2=4(E(\xi I(\{\xi>0\}))^2\le 4P(\xi>0)E\xi^2$ by the Schwarz inequality, where $I(A)$ denotes the indicator function of the set A.

Given some $v\in B$ let us choose a linear operator κ such that $\|\kappa\|=1$ and $\kappa(v)=\|v\|$. Such an operator exists by the Banach–Hahn theorem. Observe that $\{\omega\colon \|v+Z(\omega)\|\ge\|v\|\}\supset\{\omega\colon \kappa(v+Z(\omega))\ge\kappa(v)\}=\{\omega\colon\kappa(Z(\omega))\ge 0\}$. Besides, $E\kappa(Z)=0$. Hence we can apply the above proved inequality for $\xi=\kappa(Z)$, and it yields that $P(\|v+Z\|\ge\|v\|)\ge\frac{E\kappa(Z)^2}{4(E\kappa(Z))^2}$. Lemma B3 is proved.

Proof of Lemma B4. Take the class of random polynomials

$$Y=\sum_{s=1}^k \sum_{\substack{1\le l_j\le n,\ j=1,\dots,s\\ l_j\neq l_{j'} \text{ if } j\neq j'}} b(l_1,\dots,l_s)\varepsilon_{l_1}\cdots\varepsilon_{l_s},$$

where ε_l, $1\le l\le n$, are independent random variables with $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac12$, and the coefficients $b(l_1,\dots,l_s)$, $1\le s\le k$, are arbitrary real numbers. It is enough to show that there exists a constant c_k depending only on the order k of these polynomials such that the inequality

$$(E|Y|)^2\ge 4c_kEY^2. \tag{B8}$$

holds for all of these polynomials Y. Indeed, formula (B7) follows from relation (B8) and Lemma B3 with $c_k\ge\inf\limits_\kappa\frac{(E\kappa(Z))^2}{4E\kappa(Z)^2}$ if we apply them for the vector $v\in B$ in formula (B7) and

$$Z=\sum_{s=1}^k \sum_{\substack{1\le l_j\le n,\ j=1,\dots,s\\ l_j\neq l_{j'} \text{ if } j\neq j'}} a(l_1,\dots,l_s)\varepsilon_{l_1}\cdots\varepsilon_{l_s},$$

and the infimum is taken for all bounded linear operators κ on the Banach space B. But this inequality follows from relation (B8).

To prove relation (B8) first we compare the moments EY^2 and EY^4. Let us introduce the random variables

$$Y_s=\sum_{\substack{1\le l_j\le n,\ j=1,\dots,s\\ l_j\neq l_{j'} \text{ if } j\neq j'}} b(l_1,\dots,l_s)\varepsilon_{l_1}\cdots\varepsilon_{l_s}\quad 1\le s\le k,$$

and observe that because of Borell's inequality (Theorem 10.2) and the uncorrelatedness of the random variables Y_s, $1\le s\le k$,

$$EY^4=\left(\sum_{s=1}^k Y_s\right)^4\le k^3\sum_{s=1}^k EY_s^4\le k^3 3^{3k/2}\sum_{s=1}^k (EY_s^2)^2$$
$$\le k^3 3^{3k/2}\left(\sum_{s=1}^k EY_s^2\right)^2=k^3 3^{3k/2}(EY^2)^2.$$

This estimate together with the Hölder inequality yield that $EY^2=E(Y^4)^{1/3}|Y|^{2/3}\le (EY^4)^{1/3}(E|Y|)^{2/3}\le k3^{k/2}(EY^2)^{1/3}(E|Y|)^{2/3}$, i.e. $EY^2\le k^{3/2}3^{3k/4}(E|Y|)^2$, and relation (B8) holds with $4c_k=k^{-3/2}3^{-3k/4}$. Lemma B4 is proved.

Let us turn back to the estimation of the probability $P(3\cdot 2^{k-1}\|T_{n,k}(f)\|>u)$. Let us introduce the σ-algebra $\mathcal{F}=\mathcal{B}(\xi_l^{(1)},\xi_l^{(2)},\ 1\le l\le n)$ generated by the random variables $\xi_l^{(1)},\xi_l^{(2)}$, $1\le l\le n$, and fix some set $V\subset\{1,\dots,k\}$. We claim that there exists some constant $c_k>0$ that the random variable $\bar I_{n,k,V}(f(\ell))$ defined in formula (B5) satisfies the inequality

$$P\left(\|2^k\bar I_{n,k,V}(f(\ell))\|>\|T_{n,k}(f(\ell))\||\mathcal{F}\right)\ge c_k\quad\text{with probability 1.}\tag{B9}$$

Indeed, formula (B6) and the independence of the random sequences $\varepsilon_{l,V}$, $\xi_l^{(1)}$ and $\xi_l^{(2)}$, $1\le l\le n$ yield that

$$\begin{aligned}&P\left(\|2^k\bar I_{n,k,V}(f(\ell))\|>\|T_{n,k}(f(\ell))\||\mathcal{F}\right)\\&\qquad=P_{\varepsilon_V}\Bigg(\Bigg\|\frac{1}{k!}\sum_{\substack{1\le l_j\le n,s_j=1\text{ or }s_j=2,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if }j\ne j'}}(1+\varepsilon^{(1)}_{l_1,s_1,V})\cdots(1+\varepsilon^{(k)}_{l_k,s_k,V})f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(s_1)},\dots,\xi_{l_k}^{(s_k)}\right)\Bigg\|\\&\qquad\qquad>\|T_{n,k}(f(\ell))\|\Bigg),\end{aligned}\tag{B10}$$

where P_{ε_V} means that we fix the values of the random variables $\xi_l^{(1)}$, $\xi_l^{(2)}$, $1\le l\le n$ and take the probability with respect to the remaining random variables $\varepsilon^{(j)}_{l,s,V}$, $1\le j\le k$, $1\le l\le n$, and $s=1$ or $s=2$. Let us observe that the probability considered at the right-hand side of (B10) is a polynomial of order k of the random variables $\varepsilon_1,\dots,\varepsilon_n$. (The terms $\varepsilon^{(j)}_{l_j,s_j,V}$ taking part in it equal either ε_{l_j} or $-\varepsilon_{l_j}$ depending on the parameters j and s_j.) Besides, the constant term of this polynomial equals $T_{n,k}(f)$. Hence this probability can be bounded by means of Lemma B4, and this result yields relation (B9).

Relation (B9) implies that

$$
\begin{aligned}
&P(\|2^k\bar I_{n,k,V}(f(\ell))\|\ge 3\cdot 2^{k-1}u)\\
&\qquad \ge P(\|2^k\bar I_{n,k,V}(f(\ell))\|\ge \|T_{n,k}(f(\ell))\|, \|T_{n,k}(f(\ell))\|\ge 3\cdot 2^{k-1}u)\\
&\qquad =\int_{\{\omega\colon\ \|T_{n,k}(f(\ell))(\omega)\|\ge 3\cdot 2^{k-1}u\}} P\left(\|2^k\bar I_{n,k,V}(f(\ell))\|>\|T_{n,k}(f(\ell))\|\,|\mathcal F\right)\,dP\\
&\qquad \ge c_kP(\|T_{n,k}(f(\ell))\|\ge 3\cdot 2^{k-1}u)
\end{aligned}
$$

The last inequality with the choice of any set $V\subset\{1,\dots,k\}$, $1\le|V|\le k-1$, together with relation (B4) imply formula (B2).

To formulate the inductive hypothesis we need to prove formula (10.8b) with the help of relation (B2) first we introduce the following quantities. Let $\mathcal W=\mathcal W(k)$ denote the set of all partitions of the set $\{1,\dots,k\}$. Let us fix k independent copies $\xi_1^{(j)},\dots,\xi_n^{(j)}$, $1\le j\le k$, of the sequence of random variables $\xi_1,\dots,\xi_n$. Given a partition $W=(V_1,\dots,V_s)\in\mathcal W(k)$ let us introduce the function $s_W(j)$, $1\le j\le k$, which tells for all arguments j the index of that element of the partition W which contains the point j, i.e. the function $s_W(j)$, $1\le j\le k$, is defined by the relation $j\in V_{s_W(j)}$. Let us define (actually generalizing the notion introduced in formula (B1)) the notion of generalized decoupled U-statistics corresponding to a partition $W\in\mathcal W(k)$ as

$$
I_{n,k,W}(f(\ell))=\frac1{k!}\sum_{\substack{1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if } j\ne j'}} f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(s_W(1))},\dots,\xi_{l_k}^{(s_W(k))}\right)\qquad\text{for all } W\in\mathcal W(k).
$$

Given a partition $W=(V_1,\dots,V_s)$ let us call the number s of the elements of this partition the rank both of the partition W and of the generalized decoupled U-statistic $I_{n,k,W}(f(\ell))$.

Relation (10.8b) will be proved by induction with respect to the order k of the U-statistics. This induction assumption clearly holds for $k=1$, so when we prove it for k we may assume that it holds for all $k'<k$. We prove it by first showing the following statement. Fix the number k. For all numbers $2\le j\le k$ there exist some constants $C(k,j)>0$ and $\delta(k,j)>0$ such that for all generalized decoupled U-statistics $I_{n,k,W}(f(\ell))$ of order k

$$
\begin{aligned}
P(\|I_{n,k,W}(f(\ell))\|>u)\le C(k,j)P\left(\|\bar I_{n,k}(f(\ell))\|>\delta(k,j)u\right)\\
\text{for all } 2\le j\le k \text{ if the rank of } W \text{ equals } j.
\end{aligned}\tag{B11}
$$

(In relation (B11) we compare the distribution of some generalized decoupled U-statistics with that of the decoupled U-statistic $\bar I_{n,k}(f(\ell))$.) We shall prove this statement by means of a backward induction with respect to the rank j of the generalized decoupled U-statistics.

Relation (B11) clearly holds for $j=k$ with $C(k,k)=1$ and $\delta(k,k)=1$. To prove it for generalized decoupled U-statistics of rank $2\le j<k$ first we make the following

observation. If the rank j of the partition $W = (U_1, \dots, U_j)$ satisfies the relation $2 \le j \le k-1$, then it contains an element with cardinality strictly less than k and strictly greater than 1. For the sake of simpler notation let us assume that the element U_j of this partition is such an element, and $U_j = \{s, \dots, k\}$ with some $2 \le s \le k-1$. The investigation of general U-statistics of rank j, $2 \le j \le k-1$ can be reduced to this case by a reindexation of the random arguments in the U-statistics if it is necessary. Let us consider the partition $\bar W = (U_1, \dots, U_{j-1}, \{s\}, \dots, \{k\})$ and the generalized decoupled U-statistic $I_{n,k,\bar W}(f(\ell))$ corresponding to this partition $\bar W$. We show that our inductive hypothesis implies the inequality

$$P(\|I_{n,k,W}(f(\ell))\| > u) \le \bar A(k) P\left(\|I_{n,k,\bar W}(f(\ell))\| > \bar\gamma(k) u\right) \tag{B12}$$

with $\bar A(k) = \sup\limits_{j \le k-1} A(j)$, $\gamma(k) = \inf\limits_{j \le k-1} \gamma(j)$ if the rank j of W is such that $2 \le j \le k-1$.

To prove relation (B12) let us define the σ-algebra $\mathcal{F}$ generated by the random variables appearing in the first $s-1$ coordinates of these generalized U-statistics. We show that relation (10.8b) for U-statistics of order $k-s+1 \le k-1$ yields that $P(\|I_{n,k,W}(f(\ell))\| > u|\mathcal{F}) \le \bar A(k) P\left(\|I_{n,k,\bar W}(f(\ell))\| > \bar\gamma(k) u|\mathcal{F}\right)$ with probability 1. This inequality follows from our inductive hypothesis, since the conditional probabilities we compare here are generalized U-statistics and generalized decoupled U-statistics of order $k-s+1$ we get by putting substituting the (known) first $s-1$ coordinates in the generalized U-statistics $I_{n,k,W}(f(\ell))$ and $I_{n,k,\bar W}(f(\ell))$. Then taking expectation at both sides of this inequality we get relation (B12). As the rank of $\bar W$ is strictly greater than the rank of W relation (B12) together with our backward inductive assumption imply relation (B11) for all $2 \le j \le k$.

Inequality (10.8b) is a simple consequence of relations (B2) and (B11). Indeed, the probability $P(\|I_{n,k}(f(\ell))\| > u)$ is bounded in formula (B2) by such an expression, where some linear combination of the probabilities are considered that certain generalized decoupled U-statistics of order k and rank 2 are larger than uD_k^{-1}. Each of these terms can be bounded by means of relation (B11), and in such a way we get relation (10.8b).

We prove formula (10.8′) first in the simpler case when the supremum of finitely many functions is taken. Let us have M functions $f_1, \dots, f_M$, and to prove relation (10.8′) in this case let us apply formula (10.8) with the function $f = (f_1, \dots, f_M)$ taking values in the separable Banach space B_M consisting of the points $(v_1, \dots, v_M)$, $v_j \in B$, $1 \le j \le M$, with the norm $\|(v_1, \dots, v_M)\| = \sup\limits_{1 \le j \le m} \|v_j\|$. The application of formula (10.8) with this choice yields formula (10.8′) in this case. Let us emphasize that the constants appearing in this estimate do not depend on the number M. Since the distribution of the random variables $\sup\limits_{1 \le s \le M} \|I_{n,k}(f_s)\|$ converge to $\sup\limits_{1 \le s < \infty} \|I_{n,k}(f_s)\|$, the distribution of the random variables $\sup\limits_{1 \le s \le M} \left\|\bar I_{n,k}(f_s)\right\|$ converge to $\sup\limits_{1 \le s < \infty} \left\|\bar I_{n,k}(f_s)\right\|$ as $M \to \infty$, we get the proof of relation (10.8′) in the general case by taking limit $M \to \infty$ in this relation.

Appendix C.

Nelson's inequality and its application

In this part of the Appendix I formulate and prove Nelson's inequality and briefly indicate how it can be applied in the proof of Theorem 8.5, i.e. in the Gaussian counterpart of Theorems 8.3 and 8.4. As the latter problem does not belong to the main subject of the work, the detailed explanation of some background results I shall apply in the proof will be omitted. In particular, I do not discuss the basic results about the properties of multiple Wiener–Itô integrals. These results can be found for instance in my lecture note *Multiple Wiener–Itô integrals.*

There are several equivalent formulations of Nelson's inequality. First I present its terminologically simplest form. Before its formulation let me recall that the Hermite polynomials $H_k(x)$, $k = 0, 1, 2, \dots$, are those polynomials which constitute an orthogonal system with respect to the normal density function $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. To fix their normalization, let us make the agreement that $H_k(x)$ is a polynomial of order k, and the coefficient of its leading term x^k equals 1.

Theorem C1. (Nelson's inequality). *Let $(Y, \mathcal{Y}, \nu) = (R^\infty, \mathcal{B}^\infty, \nu^\infty)$ be the direct product of infinite many copies of the space $(R, \mathcal{B}, \lambda_\varphi)$, where R denotes the real line, $\mathcal{B}$ is the Borel σ-algebra on it, λ_φ is the measure determined by the standard normal distribution function, i.e. the probability measure which is absolutely continuous with respect to the Lebesgue measure with density function $\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$.*

Given a number $\gamma > 0$ introduce the operator $\mathbf{T}_\gamma$ on $(Y, \mathcal{Y})$ by defining it first on polynomials by the formula

$$\begin{aligned} \mathbf{T}_\gamma &\left(\sum c_{l_1, j_1, \dots, l_s, j_s} H_{l_1}(y_{j_1}) \cdots H_{l_s}(y_{j_s}) \right) \\ &= \sum \gamma^{l_1 + \cdots + l_s} c_{l_1, j_1, \dots, l_s, j_s} H_{l_1}(y_{j_1}) \cdots H_{l_s}(y_{j_s}), \end{aligned}$$

where all finite sums of the above form are considered, and $H_l(\cdot)$ denotes the Hermite polynomial of order l. Let us extend this linear operator to general functions on the space $(Y, \mathcal{Y})$ in the natural way.

Fix two numbers $1 < q \le p < \infty$ and a number $\gamma \le \sqrt{\frac{q-1}{p-1}}$. Then the operator $\mathbf{T}_\gamma$ defined above, considered as a linear operator from the space $L_q(Y, \mathcal{Y}, \nu)$ to $L_p(Y, \mathcal{Y}, \nu)$ is a contraction, i.e. $\|\mathbf{T}_\gamma(f)\|_p \le \|f\|_q$ for all functions $f \in L_q(Y, \mathcal{Y}, \nu)$.

By Theorem 11.2 Nelson's inequality can be reduced to the following one-dimensional inequality

$$\int_{-\infty}^{\infty} \left| \sum_{l=0}^{s} c_l \gamma^l H_l(x) \right|^p \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \le \left[\int_{-\infty}^{\infty} \left| \sum_{l=0}^{s} c_l H_l(x) \right|^q \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \right]^{p/q} \tag{C1}$$

for all finite polynomials $\sum\limits_{l=0}^{s} c_l H_l(x)$ if $1 < q \le p < \infty$, and $\gamma \le \sqrt{\frac{q-1}{p-1}}$.

We shall prove inequality (C1) in a seemingly more complicated equivalent form in the following Proposition C2. Proposition C2 can be proved by means of the hypercontractive inequality for Rademacher functions, the central limit theorem and by some basic results in the theory of multiple Wiener–Itô integrals.

Proposition C2. *Let us consider a Wiener process $W(t)$ on the interval $[0,1]$, and consider multiple Wiener–itô integrals with respect to it. The inequality*

$$E\left|\sum_{l=0}^s c_l\gamma^l \int W(\,dx_1)\cdots W(\,dx_l)\right|^p \le \left[E\left|\sum_{l=0}^s c_l \int W(\,dx_1)\cdots W(\,dx_l)\right|^q\right]^{p/q} \tag{C2}$$

holds for all numbers s and coefficients c_l, $0\le l\le s$ if $1<q\le p<\infty$, and $\gamma\le\sqrt{\frac{q-1}{p-1}}$.

Remark: Relations (C1) and (C2) are equivalent. To show this observe that by Itô's formula for multiple Wiener–Itô integrals $\int W(\,dx_1)\cdots W(\,dx_l)=H_l\left(\int W(\,dx)\right)$. Besides, the random variable $\xi=\int W(\,dx)=W(1)-W(0)$ has standard normal distribution, and formula (C2) can be rewritten with its help as

$$E\left|\sum_{l=0}^s c_l\gamma^l H_l(\xi)\right|^p \le \left[E\left|\sum_{l=0}^s c_l H_l(\xi)\right|^q\right]^{p/q}$$

which is clearly equivalent to relation (C1).

The proof of Proposition C2. First we want to show a version of formula (C2) where the multiple Wiener–Itô integrals are replaced by appropriate approximations of these integrals. For this goal let us consider m independent, normally distributed random variables $\xi_1,\dots,\xi_m$ with expectation zero and variance $\frac1m$. We shall prove with the help of the hypercontractive inequality for Rademacher functions and the central limit theorem (more precisely a slight generalization of it) the following inequality:

$$E\left|\sum_{l=0}^s c_l\gamma^l \sum_{\substack{1\le j_1,\dots,j_l\le m\\ j_u\ne j'_u \text{ if } u\ne u',\,1\le u,u'\le m}} \xi_{j_1}\cdots\xi_{j_l}\right|^p \le \left[E\left|\sum_{l=0}^s c_l \sum_{\substack{1\le j_1,\cdots,j_l\le m\\ j_u\ne j'_u \text{ if } u\ne u',\,1\le u,u'\le m}} \xi_{j_1}\dots\xi_{j_l}\right|^q\right]^{p/q} \tag{C3}$$

for all s and coefficients c_s, $1\le s\le l$, if $1<q\le p<\infty$, and $\gamma\le\sqrt{\frac{q-1}{p-1}}$.

To prove relation (C3) let us choose for all $n=1,2,\dots$ a sequence of independent random variables $\varepsilon_1,\dots,\varepsilon_{mn}$ such that $P(\varepsilon_j=1)=P(\varepsilon_j=-1)=\frac12$, $1\le j\le mn$, and define the random variables $Z_j^{(n)}=\frac{1}{\sqrt{mn}}\sum\limits_{k=(j-1)n+1}^{jn}\varepsilon_j$, $1\le j\le m$. The hypercontractive inequality for Rademacher functions implies that

$$E\left|\sum_{l=0}^s c_l\gamma^l \sum_{\substack{1\le j_1,\dots,j_l\le m\\ j_u\ne j'_u \text{ if } u\ne u',\,1\le u,u'\le m}} Z_{j_1}^{(n)}\cdots Z_{j_l}^{(n)}\right|^p \le \left[E\left|\sum_{l=0}^s c_l \sum_{\substack{1\le j_1,\dots,j_l\le m\\ j_u\ne j'_u \text{ if } u\ne u',\,1\le u,u'\le m}} Z_{j_1}^{(n)}\cdots Z_{j_l}^{(n)}\right|^q\right]^{p/q} \tag{C4}$$

By the central limit theorem the random vectors $(Z_1^{(n)},\dots,Z_m^{(n)})$ converge in distribution to the random vector $(\xi_1,\dots,\xi_m)$ as $n\to\infty$. This convergence in distribution also can be expressed as the relation

$$\lim_{n\to\infty} Ef(Z_1^{(n)},\dots,Z_m^{(n)}) = Ef(\xi_1,\dots,\xi_m) \tag{C5}$$

for all bounded and continuous functions f on R^m. Moreover, it can be proved that the distribution of the random vectors $Z_j^{(n)}$ converge to zero sufficiently fast at infinity, more explicitly for all $K>0$ there exists some constant $C=C(K)>0$ such that $P(|Z_m^{(n)}|>x)\le Cx^K$ for all $x\ge 1$ and $n=1,2,\dots$. This fact implies that relation (C5) also holds for continuous functions f such that $|f(x)|\le C(1+|x|)^K)$ with some constant $C>0$ and $K>0$, where $x=(x_1,\dots,x_m)$ and $|x|$ is the length of the vector x. This strengthened form of (C5) enables us to take the limit $n\to\infty$ in formula (C4) and to get relation (C3) in such a way.

Let us apply formula (C3) with the choice $\xi_j = W\left(\frac{j}{m}\right) - W\left(\frac{j-1}{m}\right)$, $1\le j\le m$. Observe that for all indices l the inner sums at both sides of this expression are approximative sums for the Wiener–Itô integral $\int W(\,dx_1)\cdots W(\,dx_l)$. Hence it is natural to expect that by applying the limiting procedure $m\to\infty$ in formula (C3) with the above choice of the random variables ξ_j we get relation (C2). This belief is correct, only its justification requires the application of some deeper results from the theory of Wiener–Itô integrals. We need some estimate which states that also the high moments of a Wiener–Itô integral with a small kernel function are small. We can apply the following result. If h is such a function in $[0,1]^l$ for which $\int |h(x_1,\dots,x_l)|^2\,dx_1\dots\,dx_l<\varepsilon$ with some $\varepsilon>0$, then also the inequality $E\left|\int h(x_1,\dots,x_l)W(\,dx_1)\dots W(\,dx_l)\right|^{2K}\le C(K,l)\varepsilon^K$ holds for all $K=1,2,\dots$ with some constant $C(K,l)$ depending only on K and l. But the proof of this estimate demands some deeper results about Wiener–Itô integrals. (In my lecture note about Wiener–Itô integrals this result is proved as a consequence of the so-called diagram formula.) By applying this limiting procedure we get the proof of (C2). In such a way we have proved Proposition C2 which, as we have shown, implies Theorem C1.

Now I formulate a version of Nelson's inequality presented in the language of Wiener–Itô integrals.

Theorem C3. *Let us fix a measurable space $(X,\mathcal{X})$ together with a countable non-atomic measure μ on it, and let Z_μ be an orthogonal Gaussian random measure with counting measure μ on $(X,\mathcal{X})$. (See the definition of counting measure before the formulation of Theorem (8.5).) For the sake of simplicity let us assume that the space $L_2(X,\mathcal{X},\mu)$ is separable.*

Let us have a sequence of measurable functions $f_k(x_1,\dots,x_k)$ on $(X^k,\mathcal{X}^k)$ of real constant c_k, $k=1,2,\dots$, and also a constant c_0 such that

$$c_0^2+\sum_{k=1}^\infty \frac{c_k^2}{k!}\int f_k^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)<\infty. \tag{C6}$$

Then

$$
\begin{aligned}
E&\left|c_0+\sum_{k=1}^{\infty}\gamma^k\frac{c_k}{k!}\int f_k(x_1,\dots,x_k)Z_\mu(\,dx_1)\dots Z_\mu(\,dx_k)\right|^p \\
&\le\left[E\left|c_0+\sum_{k=1}^{\infty}\frac{c_k}{k!}\int f_k(x_1,\dots,x_k)Z_\mu(\,dx_1)\dots Z_\mu(\,dx_k)\right|^q\right]^{p/q}.
\end{aligned}
\tag{C7}
$$

if $1<q\le p<\infty$, *and* $\gamma\le\sqrt{\frac{q-1}{p-1}}$.

Inequality (C7) means in particular that if the right-hand side is finite then the left-hand side is also finite.

The proof of Theorem C3. Theorem C3 will be proved as the consequence of Theorem C1 and Itô's formula for multiple Wiener–Itô integrals. Let us choose a complete orthonormal system $\psi_1(x),\psi_2(x),\cdots$ in the space $L_2(X,\mathcal{X},\mu)$, and define the random variables $\xi_n=\int\psi(x)Z_\mu(\,dx)$. Then $\xi_1,\xi_2,\dots$ is a sequence of independent random variables with standard normal distribution, and $U(\omega)=(\xi_1(\omega),\xi_2(\omega),\dots)$ is a measure preserving transformation of the probability space $(\Omega,\mathcal{A},P)$ where the orthonormal Gaussian random measure Z_μ is defined to the space $(Y,\mathcal{Y},\nu)$ introduced in the formulation of Theorem C1. We express the Wiener–Itô integrals

$$
V_k=\int f_k(x_1,\dots,x_k)Z_\mu(\,dx_1)\dots Z_\mu(\,dx_k),\quad 1\le k<\infty,
$$

by means of Itô's formula as a function of the Hermite polynomials of the random variables ξ_j, $1\le j<\infty$, and then we deduce Theorem C3 from Theorem C1 by means of the above introduced measure preserving transformation U.

To carry out this program let us expand the function $f_k(x_1,\dots,x_k)$ by means of the complete orthonormal system consisting of the products $\psi_{j_1}(x_1)\cdots\psi_{j_k}(x_k)$ $1\le j_s<\infty$, $j=1,\dots,k$, in the space $(X^k,\mathcal{X}^k,\mu^k)$. We can write

$$
f_k(x_1,\dots,x_k)=\sum_{s=1}^{k}\sum_{\substack{(j_1,\dots,j_s),\,(l_1,\dots,l_s)\\ j_u\ge1,\,l_u\ge1,\,1\le u\le s,\,j_1+\cdots+j_s=k\\ l_u\ne l_{u'}\text{ if } u\ne u',\,1\le u,u'\le s}} d_{j_1,\dots,j_s,l_1,\dots,l_s}F_{j_1,\dots,j_s,l_1,\dots,l_s}(x_1,\dots,x_k)
$$

with some appropriate coefficients $d_{j_1,\dots,j_s,l_1,\dots,l_s}$ and

$$
F_{j_1,\dots,j_s,l_1,\dots,l_s}(x_1,\dots,x_k)=\prod_{u=1}^{s}\psi_{l_u}(x_{J(u-1)+1})\psi_{l_u}(x_{J(u-1)+1})\cdots\psi_{l_u}(x_{J(u)}),
$$

where $J(0)=0$ and $J(u)=j_1+\cdots+j_u$, $1\le u\le s$.

By Itô's formula $\int F_{j_1,\dots,j_s,l_1,\dots,l_s}(x_1,\dots,x_k)Z_\mu(\,dx_1)\dots Z_\mu(\,dx_k) = \prod\limits_{u=1}^{s} H_{j_u}(\xi_{l_u})$, and $\int \gamma^k F_{j_1,\dots,j_s,l_1,\dots,l_s}(x_1,\dots,x_k)Z_\mu(\,dx_1)\dots Z_\mu(\,dx_k) = \prod\limits_{u=1}^{s} \gamma^{j_u} H_{j_u}(\xi_{l_u})$.

By summing up the above inequalities we get that

$$V_k = U^*\left(\sum_{s=1}^{k} \sum_{\substack{(j_1,\dots,j_s),\,(l_1,\dots,l_s)\\ j_u\ge1,\,l_u\ge1,\,1\le u\le s,\,j_1+\cdots+j_s=k}} \bar d_{j_1,\dots,j_s,l_1,\dots,l_s} H_{j_1}(y_{l_1})\cdots H_{j_s}(y_{l_s})\right)$$

and

$$\gamma^k V_k = U^*\left(\sum_{s=1}^{k} \sum_{\substack{(j_1,\dots,j_s),\,(l_1,\dots,l_s)\\ j_u\ge1,\,l_u\ge1,\,1\le u\le s,\,j_1+\cdots+j_s=k}} \bar d_{j_1,\dots,j_s,l_1,\dots,l_s} \gamma^{j_1} H_{j_1}(y_{l_1})\cdots \gamma^{j_s} H_{j_s}(y_{l_s})\right)$$

with some coefficients $\bar d_{j_1,\dots,j_s,l_1,\dots,l_s}$, where U^* denotes the operator from the space of functions on $(Y,\mathcal{Y},\nu)$ to the space of functions on $(\Omega,\mathcal{A},P)$ induced by the measure preserving transformation U. Summing up these identities for all $k=0,1,2,\dots$ and exploiting the measure preserving property of the transformation U we get that Theorem C1 implies Theorem C3. (Let me remark that condition (C6) was imposed only to guarantee that the infinite sum of the Wiener–Itô integrals we considered really exists.)

The proof of formula (8.11) in Theorem 8.5 is fairly simple with the help of Nelson's inequality.

The proof of formula (8.11). Let us observe that relation (B7) with $q=2$, $p=2M$ yields that for a k-fold Wiener–Itô integral

$$\begin{aligned} E&\left|\frac{1}{k!}\int f(x_1,\dots,x_k)Z_\mu(\,dx_1)\dots Z_\mu(\,dx_k)\right|^{2M} \\ &\le (2M-1)^{kM}\left[E\left|\frac{1}{k!}\int f(x_1,\dots,x_k)Z_\mu(\,dx_1)\dots Z_\mu(\,dx_k)\right|^2\right]^M \\ &= (2M-1)^{kM}\left(\frac{1}{k!}\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\right)^M \qquad \text{if } M\ge1. \end{aligned} \tag{C8}$$

In the last line of formula (C8) we exploited the following simple but basic relation of the theory of Wiener–Itô integrals:

$$E\left(\frac{1}{k!}\int f(x_1,\dots,x_k)Z_\mu(\,dx_1)\dots Z_\mu(\,dx_k)\right)^2 = \frac{1}{k!}\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k).$$

Relation (C8) and the Markov inequality imply that under the conditions of Theorem 8.5

$$P(|Z_{\mu,k}(f)| > u) \le \frac{EZ_{\mu,k}(f)^{2M}}{u^{2M}} \le \left(\frac{(2M)^k\sigma^2}{u^2}\right)^M$$

if $M \ge 1$. We get with the choice $M = \frac{1}{2e}\left(\frac{u}{\sigma}\right)^{2/k}$ that

$$P(|Z_{\mu,k}(f)| > u) \le \exp\left\{-\frac{k}{2e}\left(\frac{u}{\sigma}\right)^{2/k}\right\} \quad \text{if } u > (2e)^{k/2}\sigma.$$

By choosing a sufficiently large $A \ge 1$ at the right-hand side of this inequality we get that formula (8.11) holds for all $u \ge 0$.

The second inequality (8.12) of Theorem 8.5 can be proved in the same way as Theorem 4.1 in the one-dimensional case. No difficulty arises during the proof. The main point is that inequality (8.11) holds for all $u > 0$, hence the chaining argument applied in the proof of Theorem 4.1 supplies the proof also in this case. I omit the details.

Let me finally remark that Leonhard Gross in his paper *Logarithmic Sobolev inequalities* also gave a proof of the Nelson inequality by means of the hypercontractive inequality for Rademacher functions. He showed that the central limit theorem enables us to prove that the logarithmic Sobolev inequality holds not only for the Markov process considered in Section 11, but also for Wiener processes. This result together with the general theory he presents imply an inequality which is equivalent to our formula (C1).

References:

1.) Alexander, K. (1987) The central limit theorem for empirical processes over Vapnik–Červonenkis classes. *Ann. Probab.* **15**, 178–203
2.) Arcones, M. A. and Giné, E. (1993) Limit theorems for U-processes. *Ann. Probab.* **21**, 1494–1542
3.) Arcones, M. A. and Giné, E. (1994) U-processes indexed by Vapnik–Červonenkis classes of functions with application to asymptotics and bootstrap of U-statistics with estimated parameters. *Stoch. Proc. Appl.* **52**, 17–38
4.) Bennett, G. (1962) Probability inequality for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57**, 33-45
5.) Bonami, A. (1970) Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Ann. Inst. Fourier* **20**, 335–402
6.) de la Peña, V. H. and Giné, E. (1999) *Decoupling. From dependence to independence.* Springer series in statistics. Probability and its application. Springer Verlag, New York, Berlin, Heidelberg
7.) de la Peña, V. H. and Montgomery–Smith, S. (1995) Decoupling inequalities for the tail-probabilities of multivariate U-statistics. *Ann. Probab.*, **23**, 806–816
8.) Dudley, R. M. (1978) Central limit theorems for empirical measures. *Ann. Probab.* **6**, 899–929

9.) Dudley, R. M. (1984) A course on empirical processes. *Lecture Notes in Mathemematics* **1097**, 1–142 Springer Verlag, New York

10.) Dudley, R. M. (1989) *Real Analysis and Probability.* Wadsworth & Brooks, Pacific Grove, California

11.) Dudley, R. M. (1998) *Uniform Central Limit Theorems.* Cambridge University Press, Cambridge U.K.

12.) Dynkin, E. B. and Mandelbaum, A. (1983) Symmetric statistics, Poisson processes and multiple Wiener integrals. *Annals of Statistics* **11**, 739–745

13.) Gross, L. (1975) Logarithmic Sobolev inequalities. Amer. J. Math. **97**, 1061–1083

14.) Guionnet, A. and Zegarlinski, B. (2003) Lectures on Logarithmic Sobolev inequalities. *Lecture Notes in Mathematics* **1801** 1–134 2. Springer Verlag, New York

15.) Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Math. Society* **58**, 13–30

16.) Ledoux, M. (1996) On Talagrand deviation inequalities for product measures. *ESAIM: Probab. Statist.* **1.** 63–87. Available at http://www.emath./fr/ps/.

17.) Major, P. (1981) Multiple Wiener–Itô integrals. *Lecture Notes in Mathematics* **849**, Springer Verlag, Berlin, Heidelberg, New York,

18.) Major, P. (1988) On the tail behaviour of the distribution function of multiple stochastic integrals. *Probability Theory and Related Fields*, **78**, 419–435

19.) Major, P. (2005) An estimate about multiple stochastic integrals with respect to a normalized empirical measure. Submitted to *Studia Scientarum Mathematicarum Hungarica.*

20.) Major, P. (2005) An estimate on the maximum of a nice class of stochastic integrals. Submitted to *Probability Theory and Related Fields*,

21.) Major, P. (2005) On a multivariate version of Bernstein's inequality submitted to *Ann. Probab.*

22.) Major, P. (2005) A multivariate generalization of Hoeffding's inequality. Submitted to *Ann. Probab.*

23.) Major, P. (2005) On the tail behaviour of multiple random integrals and U-statistics, on the supremum of classes of such quantities, and some related questions. (An overview work submitted to xxx)

24.) Major, P. and Rejtő, L. (1988) Strong embedding of the distribution function under random censorship. *Annals of Statistics*, **16**, 1113–1132

25.) Major, P. and Rejtő, L. (1998) A note on nonparametric estimations. In the conference volume to the 65. birthday of Miklós Csörgő. 759–774

26.) Massart, P. (2000) About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28**, 863–884

27.) Nelson, E. (1973) The free Markov field. J. Functional Analysis **12**, 211–227

28.) Pollard, D. (1984) *Convergence of Stochastic Processes.* Springer Verlag, New York

29.) Talagrand, M. (1996) New concentration inequalities in product spaces. *Invent. Math.* **126**, 505–563

30.) Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory.* Springer Verlag, New York

Content

1. Introduction 1
2. Motivation of the investigation. Discussion of some problems ... 3
3. Some estimates about sums of independent random variables ... 10
4. On the supremum of a nice class of partial sums 15
5. Vapnik–Červonenkis classes and L_2-dense classes of functions .. 23
6. The proof of Theorems 4.1 and 4.2 on the supremum of random sums 26
7. The completion of the proof of Theorem 4.1 33
8. Formulation of the main results of this work 40
9. Some results about U-statistics 47
10. The proof of Theorem 8.3 about the distribution of U-statistics 60
11. Some useful basic results 69
12. Reduction of the main result in this work 79
13. The strategy of the proof for the main result of this paper 87
14. A symmetrization argument 92
15. The proof of the main result 105
16. The improvement of some results in Section 8. 115
17. An overview of the results in this work 121

Appendix A. The proof of some results about Vapnik–Červonenkis classes 134

Appendix B. The proof of Theorem 10.3. (A result of de le Peña and Montgomery–Smith) 135

Appendix C. Nelson's inequality and its application 143

References 148