TAIL-DISTRIBUTION ESTIMATES FOR MULTIPLE RANDOM INTEGRALS AND U-STATISTICS

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

1. Introduction.

First I briefly describe the main subject of this work. Fix a positive integer n, consider n independent and identically distributed random variables $\xi_1, \dots, \xi_n$ on a measurable space $(X, \mathcal{X})$ with some distribution μ and take their empirical distribution μ_n together with its normalization $\sqrt{n}(\mu_n - \mu)$. Beside this, 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 the k-th power of the normalized empirical measure $\sqrt{n}(\mu_n - \mu)$ on $(X^k, \mathcal{X}^k)$, and define the integral of the function f with respect to this signed product measure. This integral is a random variable, and we want to give a good estimate on its tail distribution. More precisely, we take the integrals not on the whole space, the diagonals $x_s = x_{s'}$, $1 \le s, s' \le k$, $s \ne s'$, of the space X^k are omitted 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 integrals of all functions in this class with respect to the k-fold direct product of our normalized empirical measure. Give a good estimate on the tail distribution of the supremum of these integrals.

It may be asked why the above problems deserve a closer study. I found them important, because they may help in solving some essential 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 similar but more difficult questions. 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 applying 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 these 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 was 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 maximum likelihood estimates. 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 above formulated problems about random integrals are interesting and non-trivial even in the special case $k = 1$. Their solution leads to some interesting and non-trivial generalization of the fundamental theorem of the mathematical statistics about the difference of the empirical and real distribution of a large sample.

These 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.

Let us try to get some feeling about what kind of results can be expected in these problems. For a 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 problems about multiple Gaussian integrals, called Wiener–Itô integrals in the literature. 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. Beside this, a similar estimate should hold for the supremum of multiple random integrals of a class of functions with respect to a normalized empirical measure or with respect to a Gaussian random measure 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 consideration supplies a fairly good description of the situation, but it does not take into account a very 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 very small, what turns out to be equivalent to a very small L_2-norm of the function we are integrating, then the behaviour of this integral is different from that of a multiple Gaussian integral with the same kernel function. 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. This case must be better understood, and some new methods have to be worked out to handle it.

The precise formulation of the results will be given in the main part of the work. Beside their proof I also tried 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. The results about Gaussian random variables and their non-linear functionals, in particular multiple integrals with respect to a Gaussian field, have a most important role in the study of the present work. Hence they will be discussed in detail together with some counterparts about multiple random integrals with respect to a normalized empirical measure and some results about U-statistics.

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, the diagram formula about the product of Wiener–Itô integrals, their counterpart about the product of degenerate U-statistics, etc. I tried to give such an exposition where different parts of the problem are explained independently of each other, and they can be understood in themselves.

An earlier version of 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 problems of the probability theory. I try to give a good picture about the main ideas, but I 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. The following problem is considered. 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.

The maximum likelihood method suggests the following approach. Choose that value $\hat{\vartheta}_n = \hat{\vartheta}_n(\xi_1, \dots, \xi_n)$ as the estimate of the parameter ϑ_0 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 yields 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

$$\sum_{k=1}^{n} \frac{\partial}{\partial\vartheta} \log f(\xi_k, \hat{\vartheta}_n) = \sum_{k=1}^{n} \frac{\frac{\partial}{\partial\vartheta} f(\xi_k, \vartheta_0)}{f(\xi_k, \vartheta_0)}$$
$$+ (\hat{\vartheta}_n - \vartheta_0) \left(\sum_{k=1}^{n} \left(\frac{\frac{\partial^2}{\partial\vartheta^2} f(\xi_k, \vartheta_0)}{f(\xi_k, \vartheta_0)} - \frac{\left(\frac{\partial}{\partial\vartheta} f(\xi_k, \vartheta_0)\right)^2}{f^2(\xi_k, \bar{\vartheta}_0)} \right) \right) + O\left(n(\hat{\vartheta}_n - \vartheta_0)^2 \right)$$
$$= \sum_{k=1}^{n} \left(\eta_k + \zeta_k(\hat{\vartheta}_n - \vartheta_0) \right) + O\left(n(\hat{\vartheta}_n - \vartheta_0)^2 \right), \tag{2.2}$$

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 a differentiation of this relation gives 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 $\chi_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \eta_k$ 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 $\frac{1}{n} \sum_{k=1}^{n} \zeta_k \sim -I^2$.

Thus relation (2.2) suggests the approximation $\tilde{\vartheta}_n = -\dfrac{\sum_{k=1}^{n} \eta_k}{\sum_{k=1}^{n} \zeta_k}$ of the maximum-likelihood estimate $\hat{\vartheta}_n$, and $\sqrt{n}(\tilde{\vartheta}_n - \vartheta_0)$ is asymptotically normal with expectation zero and variance $\frac{1}{I^2}$. 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 $\hat{\vartheta}_n - \tilde{\vartheta}_n = O\left(\frac{1}{n}\right)$. Hence $\sqrt{n}(\hat{\vartheta}_n - \vartheta_0)$ 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 committed by a replacement of the original expression by a simpler linearized version is negligible, and to show this the solution of the problems mentioned in the introduction is needed. The discussion of the following problem, called the Kaplan–Meyer method for the estimation of the empirical distribution function with the help of censored data shows such an example.

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 some unknown 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 $S(u) = 1 - F(u)$:

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

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. The complementary event has only an exponentially small probability. 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 $n^{-1/2}$ times 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

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

It is not difficult to 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. It 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 $N(Y_i) = \sum_{j=1}^{n} I(Y_j > Y_i)$ 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

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

and express the terms $\frac{1}{N(Y_i)}$ in the sum defining $\tilde{\Lambda}_n$ by means of the relation $\frac{1}{1+z} = \sum_{k=0}^{\infty}(-1)^k z^k = 1 - z + \varepsilon(z)$ 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

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

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 relations (2.9) and (2.10) yield that

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

This means that to solve our problem the asymptotic behaviour of the random variables $A(u)$ and $B(u)$ has to be given. We can get a better insight to this problem by rewriting the sum $A(u)$ as an integral and the double sum $B(u)$ as a two-fold integral with respect to empirical measures. Then these integrals can be rewritten as sums of random integrals with respect to normalized empirical measures and deterministic measures. Such an approach yields 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

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

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 normalized empirical processes and B_4 is a nonlinear functional of normalized 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 relations (2.11), (2.12) and (2.13) it follows that

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

Integration of B_2 and B_3 with respect to the variable x and then integration by parts in the expression B_2 yields that

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

With the help of the above expressions for B_2 and B_3, (2.14) can be rewritten as

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

Formula (2.15) almost agrees with the statement we wanted to prove. Here the normalized error $\sqrt{n}\,(\Lambda_n(u) - \Lambda(u))$ is expressed as a sum of linear functionals of normalized empirical measures plus some negligible error terms plus 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 $B_4(u)$ integration is taken with respect to 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 $\sup_{u \le T} |\sqrt{n}(\Lambda_n(u) - \Lambda(u)) - B_2(u) + B_3(u)|$ 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 $\sup_u |B_4(u)|$, 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 the limit behaviour of the Kaplan–Meyer estimate can be described by means of an appropriate expansion. The steps of the calculation leading to such an expansion are fairly standard, the only hard part is the solution of the problems mentioned in the introduction. It can be expected that such a method also works in a much more general situation.

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 they can be generalized without any effort to sums of not necessarily identically distributed random variables the condition about identical distribution of the summands will be dropped. 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 otherwise.

More explicitly, the following problem will be considered: 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 $S_n = \sum_{j=1}^n X_j$ and its variance $\operatorname{Var} S_n = V_n^2 = \sum_{j=1}^n \sigma_j^2$. We want to get a good bound on the probability $P(S_n > uV_n)$. The central limit theorem suggests that under general conditions an upper bound of the order $1 - \Phi(u)$ should hold for this probability, where $\Phi(u)$ denotes the standard normal distribution function. Since the standard normal distribution function satisfies the inequality $\left(\frac{1}{u} - \frac{1}{u^3}\right) \frac{e^{-u^2/2}}{\sqrt{2\pi}} < 1 - \Phi(u) < \frac{1}{u} \frac{e^{-u^2/2}}{\sqrt{2\pi}}$ for all $u > 0$ it is natural to ask when the probability $P(S_n > uV_n)$ is comparable with the value $e^{-u^2/2}$. More generally, we shall call an upper bound of the form $P(S_n > uV_n) \le e^{-Cu^2}$ with some constant $C > 0$ a Gaussian type estimate.

First I formulate Bernstein's inequality which tells for which values u the probability $P(S_n > uV_n)$ has a Gaussian type estimate. It supplies such an estimate if $u \le \text{const.}\, V_n$.

On the other hand, for $u \geq \text{const.}\, V_n$ it yields a much weaker estimate. I also present an example which shows that in this case only a very weak improvement of Bernstein's inequality is possible. I also discuss another result, called Bennett's inequality, which shows that such an improvement is possible. The main difficulties we meet in this work are closely related to the weakness of the estimates we have for the probability of the event $P(S_n > uV_n)$ if $u \gg \text{const.}\, V_n$.

In the usual formulation of Bernstein's inequality 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 $M = 1$ I shall consider only this special case.

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

$$P\left(S_n > uV_n\right) \leq \exp\left\{-\frac{u^2}{2\left(1 + \frac{1}{3}\frac{u}{V_n}\right)}\right\} \quad \text{for all } u > 0. \tag{3.1}$$

Proof of Theorem 3.1. Let us give a good bound on the exponential moments Ee^{tS_n} for appropriate parameters $t > 0$. Since $EX_j = 0$ and $E|X_j^{k+2}| \leq \sigma^2$ for $k \geq 0$ we can write $Ee^{tX_j} = \sum_{k=0}^{\infty} \frac{t^k}{k!} EX_j^k \leq 1 + \frac{t^2\sigma_j^2}{2}\left(1 + \sum_{k=1}^{\infty} \frac{2t^k}{(k+2)!}\right) \leq 1 + \frac{t^2\sigma_j^2}{2}\left(1 + \sum_{k=1}^{\infty} 3^{-k}t^k\right) = 1 + \frac{t^2\sigma_j^2}{2}\frac{1}{1-\frac{t}{3}} \leq \exp\left\{\frac{t^2\sigma_j^2}{2}\frac{1}{1-\frac{t}{3}}\right\}$ if $0 \leq t < 3$. Hence $Ee^{tS_n} = \prod_{j=1}^n Ee^{tX_j} \leq \exp\left\{\frac{t^2V_n^2}{2}\frac{1}{1-\frac{t}{3}}\right\}$ for $0 \leq t < 3$.

The above relation implies that

$$P\left(S_n > uV_n\right) = P(e^{tS_n} > e^{tuV_n}) \leq Ee^{tS_n}e^{-tuV_n} \leq \exp\left\{\frac{t^2V_n^2}{2}\frac{1}{1-\frac{t}{3}} - tuV_n\right\}$$

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

If the random variables $X_1, \dots, X_n$ satisfy the conditions of Bernstein's inequality, then also the random variables $-X_1, \dots, -X_n$ satisfy them. By applying the above result in both cases we get that $P(|S_n| > uV_n) \leq 2\exp\left\{-\frac{u^2}{2\left(1+\frac{1}{3}\frac{u}{V_n}\right)}\right\}$ under the conditions of Bernstein's inequality.

By Bernstein's inequality there is some sufficiently small number $\alpha(\varepsilon) > 0$ for all $\varepsilon > 0$ such that in the case $\frac{u}{V_n} < \alpha(\varepsilon)$ $P(S_n > uV_n) \leq e^{-(1-\varepsilon)u^2/2}$. Beside this, for all

fixed numbers $A > 0$ there is some constant $C = C(A) > 0$ such that in the case $\frac{u}{V_n} < A$ the inequality $P(S_n > uV_n) \le e^{-Cu^2}$ holds. This can be interpreted as a Gaussian type estimate for the probability $P(S_n > uV_n)$ if $u \le const.V_n$.

On the other hand, if $\frac{u}{V_n}$ is very large, then Bernstein's inequality yields a much worse estimate. The next example explains its cause. There are sequences of independent and identically distributed random variables $X_1, \dots, X_n$ bounded by one and with expectation zero such that with the notations $S_n = \sum_{j=1}^{n} X_j$, $\sigma^2 = EX_j^2$, $V_n^2 = \sum_{j=1} EX_j^2 = n\sigma^2$ the probability $P(S_n > uV_n)$ is relatively large if $\frac{u}{V_n}$ is large; it is much larger than the value suggested by the normal approximation.

Example 3.2. (Sums of independent random variables with bad tail distribution for large values). *Let us fix some positive integer n, real numbers u and σ^2 such that $0 < \sigma^2 \le \frac{1}{8}$, $n > 3u \ge 6$ and $u > 4n\sigma^2$. Take a sequence of independent and 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 $S_n = \sum_{j=1}^{n} X_j$ and $V_n^2 = n\sigma^2$. Then $ES_n = 0$,* $\operatorname{Var} S_n = V_n^2$, *and*

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

with some appropriate constant $B > 0$ (not depending on n, σ and u.

Proof of Example 3.2. Let us fix an integer u such that $n > 3u$ and $u > 4n\sigma^2$. Let $B = B(u)$ denote the event that among the random variables X_j, $1 \le j \le n$, there are exactly $3u$ terms with values $+1$ or -1, and the other random variables X_j equal zero. Let us also define the event $A = A(u) \subset B(u)$ which holds if $2u$ random variables X_j are equal to 1, u random variables X_j are equal to -1, and all remaining random variables X_j, $1 \le j \le n$, are equal to zero. Clearly, $P(S_n \ge u) \ge P(A) = P(B)P(A|B)$. On the other hand, $P(B) = \binom{n}{3u} \left(\sigma^2\right)^{3u} \left(1-\sigma^2\right)^{n-3u} \ge \left(\frac{n}{3u}\right)^{3u} \left(\sigma^2\right)^{3u} e^{-4n\sigma^2} = e^{-3u\log(3u/n\sigma^2)-4n\sigma^2}$. Here we exploited that because of the condition $\sigma^2 \le \frac{1}{8}$ we have $1 - \sigma^2 \ge e^{-4\sigma^2}$. Beside this, $u \ge 4n\sigma^2$, and $P(B) \ge e^{-3u\log(3u/n\sigma^2)-u} \ge e^{-B_1 u \log(u/n\sigma^2)}$ with some appropriate $B_1 > 0$ under our assumptions.

Let us consider a set of $3u$ elements, and choose a random subset of it by taking all elements of this set with probability $1/2$ to this random subset independently of each other. I claim that the conditional probability $P(A|B)$ equals the probability that this random subset has $2u$ elements. Indeed, even the conditional probability of the event A under the condition that for a prescribed set of indices $J \subset \{1, \dots, n\}$ with exactly $3u$ elements we have $X_j = \pm 1$ if $j \in J$ and $X_j = 0$ if $j \notin J$ equals the probability of the event that the above defined random subset has $2u$ elements. This is so, because under this condition the random variables X_j take the value $+1$ with probability $1/2$ for all $j \in J$ independently of each other. Hence $P(A|B) = \binom{3u}{2u} 2^{-3u} \ge$

$e^{-Cu} \ge e^{-B_2 u \log(u/n\sigma^2)}$ with some appropriate constants $C > 0$ and $B_2 > 0$ under our conditions, since $\frac{u}{n\sigma^2} \ge 4$ in this case. The estimates given for $P(B)$ and $P(A|B)$ imply the statement of Example 3.2.

In the case $u > V_n^2$ Bernstein's inequality yields the estimate $P(S_n > u) \le e^{-\alpha u}$ with some universal constant $\alpha > 0$ in the case $u > V_n^2$, and the above example shows that at most an additional logarithmic factor can be expected 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 $\sigma_j^2 = EX_j^2$, $1 \le j \le n$, $S_n = \sum_{j=1}^n X_j$ and $V_n^2 = \operatorname{Var} S_n = \sum_{j=1}^n \sigma_j^2$. Then*

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

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

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

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

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

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 $P(S_n > u) \le e^{-tu} Ee^{tS_n} \le e^{-tu + V_n^2(e^t - 1 - t)}$ for all $t \ge 0$. We get relation (3.2) from this inequality with the choice $t = \log\left(1 + \frac{u}{V_n^2}\right)$. (This is the place of minimum of the function $-tu + V_n^2(e^t - 1 - t)$ for fixed u in the parameter t.)

Relation (3.2) and the observation $\lim_{v \to \infty} \frac{(v+1)\log(v+1) - v}{v \log v} = 1$ with the choice $v = \frac{u}{V_n^2}$ imply formula (3.3). Because of relation (3.3) to prove formula (3.4) it is enough to check it for $2 \le \frac{u}{V_n^2} \le B$ 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 $\frac{V_n^2\left[\left(1+\frac{u}{V_n^2}\right)\log\left(1+\frac{u}{V_n^2}\right) - \frac{u}{V_n^2}\right]}{u \log \frac{u}{V_n^2}}$ is a continuous and positive function

of the variable $\frac{u}{V_n^2}$ in the interval $2 \le \frac{u}{V_n^2} \le B$, hence its minimum in this interval is strictly positive.

Let me make a short comparison between Bernstein's and Bennett's inequality. Both results yield an estimate on the probability $P(S_n > u)$, and their proofs are very similar. They are based on an estimate of the moment generating functions $R_j(t) = Ee^{tX_j}$ of the summands X_j, but Bennett's inequality yields a better estimate. It may be worth mentioning that the estimate given for $R_j(t) = Ee^{tX_j}$ in the proof of Bennett's inequality agrees with the moment generating function $Ee^{t(Y_j - EY_j)}$ of the normalization $Y_j - EY_j$ of a Poissonian random variable Y_j with parameter $\operatorname{Var} X_j$. As a consequence, we get, by using the standard method of estimating tail-distributions by means of the moment generating functions such an estimate for the probability $P(S_n > u)$ which is comparable with the probability $P(T_n - ET_n > u)$, where T_n is a Poissonian random variable with parameter $V_n = \operatorname{Var} S_n$. It can be told that Bernstein's inequality yields a Gaussian and Bennett's inequality a Poissonian type estimate for sums of independent random variables. As Example 3.2 shows the latter estimate is sharp also in the case $u \gg V_n^2$ when Bernstein's inequality yields only a weak bound.

Actually Bernstein's inequality can be derived from Bennett's inequality. On the other hand, it gives a good, 'visible' bound for the probability $P(S_n > u)$ for not too large numbers u, while the estimate of Bennett's inequality is less attractive. Beside this, the improvement supplied by Bennett's inequality for large numbers u has a limited importance.

I finish this section with another estimate due to Hoeffding which will be later useful in certain symmetrization arguments.

Theorem 3.4. (Hoeffding's inequality). *Let $\varepsilon_1, \dots, \varepsilon_n$ be independent random variables, $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = \frac{1}{2}$, $1 \le j \le n$, and let $a_1, \dots, a_n$ be arbitrary real numbers. Put $V = \sum_{j=1}^{n} a_j \varepsilon_j$. Then*

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

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

Remark 2: The Rademacher functions $r_k(x)$, $k = 1, 2, \dots$, defined by the formulas $r_k(x) = 1$ if $(2j-1)2^{-k} \le x < 2j2^{-k}$ and $r_k(x) = -1$ if $2(j-1)2^{-k} \le x < (2j-1)2^{-k}$, $1 \le j \le 2^{k-1}$, for all $k = 1, 2, \dots$, can be considered 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 $[0, 1]$. The Rademacher functions as random variables on the above probability space are independent with the same distribution

as the random variables $\varepsilon_1, \dots, \varepsilon_n$ considered in Theorem 3.4. Therefore results about such sequences of random variables whose distributions agree with those in Theorem 3.4 are also called results about Rademacher functions in the literature. At some points we will also use this terminology.

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

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

This section contains an estimate about the supremum of an appropriate class of random one-fold integrals with respect to a normalized empirical measure. This result can be considered as the solution of the one-variate version of the general problem about the behaviour of multiple integrals with respect to a normalized empirical measure mentioned in the introduction. An equivalent version of this estimate about the supremum of a nice class of sums of independent and identically distributed random variables will be also presented. Some natural questions related to these results will be also discussed. It will be examined how restrictive the conditions of these results are. In particular, we are interested in the question how the condition about the countable cardinality of the class of random variables can be weakened. A natural Gaussian counterpart of the supremum problems about random one-fold integrals will be also considered. Most proofs will be postponed to later sections.

To formulate these results first a notion will be introduced that plays a most important role in the sequel.

Definition of L_p-dense classes of functions. *Let a measurable space $(Y, \mathcal{Y})$ be given together with a set $\mathcal{G}$ of $\mathcal{Y}$ measurable real valued functions on this space. The class of functions $\mathcal{G}$ is called an L_p-dense class of functions, $1 \le p < \infty$, with parameter D and exponent L if for all numbers $0 < \varepsilon \le 1$ 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 $L_p(Y, \mathcal{Y}, \nu)$ with $m \le D\varepsilon^{-L}$ elements, i.e. there exists such a set $\mathcal{G}_{\varepsilon,\nu} \subset \mathcal{G}$ with $m \le D\varepsilon^{-L}$ elements for which $\inf_{g_j \in \mathcal{G}_{\varepsilon,\nu}} \int |g - g_j|^p \, d\nu < \varepsilon^p$ 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.)*

In most results of this work where L_p-dense classes will be considered the above definition will be applied for $p = 2$, i.e. L_2-dense classes of functions will be considered. But in some special considerations it will be useful to work also with L_p-dense classes

with a different parameter p. Hence to avoid some repetitions I introduced the above definition for a general parameter p.

The following estimate will be proved.

Theorem 4.1. (Estimate on the supremum of a class of partial sums). *Let us consider a sequence of independent and identically distributed random variables $\xi_1, \dots, \xi_n$, $n \ge 2$, with values in a measurable space $(X, \mathcal{X})$ and with some distribution μ. Beside this, let a countable and L_2-dense class of functions $\mathcal{F}$ with some parameter $D > 0$ and exponent $L \ge 1$ be given 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 $S_n(f) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} f(\xi_k)$ 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 universal constants $C > 0$, $\alpha > 0$ and $M > 0$ such that the supremum of the normalized random sums $S_n(f)$, $f \in \mathcal{F}$, satisfies the inequality

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

with $\beta = \max\left(\frac{\log D}{\log n}, 0\right)$, 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 $\sqrt{n}\sigma^2 \ge u \ge \sqrt{M}(L+\beta)^{3/4} \sigma \log^{1/2} \frac{2}{\sigma}$ about the number u in formula (4.4) is natural. I discuss it after the formulation of Theorem 4.2 which can be considered as the Gaussian counterpart of Theorem 4.1.

The condition about the countable cardinality of $\mathcal{F}$ can be weakened with the help of the notion of countable approximability introduced below. For the sake of later applications it will be defined in a more general form than it is needed in this section.

Definition of countably approximable classes of random variables. *Let us have a class of random variables $U(f)$, $f \in \mathcal{F}$, indexed by a class of functions $f \in$*

$\mathcal{F}$ on a measurable space $(Y, \mathcal{Y})$. This class of random variables is called countably approximable if there is a countable subset $\mathcal{F}' \subset \mathcal{F}$ such that for all numbers $u > 0$ the sets $A(u) = \{\omega\colon \sup_{f \in \mathcal{F}} |U(f)(\omega)| \ge u\}$ and $B(u) = \{\omega\colon \sup_{f \in \mathcal{F}'} |U(f)(\omega)| \ge u\}$ satisfy the identity $P(A(u) \setminus B(u)) = 0$.

Clearly, $B(u) \subset A(u)$. In the above definition it was 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, only Theorem 4.1 has to be applied for the class $\mathcal{F}'$. To do this it has to be checked 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 $0 < \varepsilon \le 1$ 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 $\bigcup_{j=1}^{m} \mathcal{D}_j = Y$. 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.

Next I formulate in Theorem 4.1′, a result about the supremum of the integral of a class of functions with respect to a normalized empirical distribution. It can be considered as a simple version of Theorem 4.1. I formulated this result, because Theorems 4.1 and 4.1′ are special cases of their multivariate counterparts about the supremum of so-called U-statistics and multiple integrals with respect to a normalized empirical distribution functions discussed in Section 8. These results are also closely related, but the explanation of their relation is not self-evident.

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

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

and define for all measurable (and μ 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 $J_n(f) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (f(\xi_j) - Ef(\xi_j)) = S_n(\bar{f})$ with $\bar{f}(x) = f(x) - \int f(x)\mu(\,dx)$. It is not difficult to see that $\sup_{x \in X} |\bar{f}(x)| \le 2$ if $\sup_{x \in X} |f(x)| \le 1$, $\int \bar{f}(x)\mu(\,dx) = 0$, $\int \bar{f}^2(x)\mu(\,dx) \le \int f^2(x)\mu(\,dx)$, 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.

Theorem 4.1′. (Estimate on the supremum of random integrals with respect to a normalized empirical measure). *Let us have a sequence of independent and identically distributed 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 satisfies the conditions of Theorem 4.1 with the possible exception of condition (4.3). The estimate (4.4) remains valid if the random sums $S_n(f)$ are replaced in it by the random integrals $J_n(f)$ defined in (4.6). Moreover, similarly to the corollary of Theorem 4.1, the condition about 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 as $n \to \infty$ to those of a Gaussian field $Z(f)$, $f \in \mathcal{F}$, 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. We can expect that the random variables of a Gaussian field with such properties satisfy an estimate similar to Proposition 4.1. The following Theorem 4.2, which can be considered as the Gaussian counterpart of Theorem 4.1, contains such a result. Let me also remark that in Section 10 so-called multiple Wiener–Itô integrals of functions of k variables with respect to a white noise will be defined for all $k \ge 1$. In the special case $k = 1$ the Wiener–Itô integrals $Z(f)$, $f \in \mathcal{F}$, of functions of one variable constitute a Gaussian field with the above properties.

Theorem 4.2. (Estimate on the supremum of a class of Gaussian random variables). *Let a probability measure μ be given 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 $0 < \varepsilon \le 1$ 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 $f \in \mathcal{F}$ $\inf_{1 \le j \le m} \int (f_j(x) - f(x))^2 \mu(\,dx) < \varepsilon^2$. Let us also assume that the class of functions $\mathcal{F}$ satisfies condition (4.2) with some $0 < \sigma \le 1$. 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.

The exponent at the right-hand side of inequality (4.7) does not contain the best possible universal constant. One could choose 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 (4.7) if the universal constants $C > 0$ and $M > 0$ are chosen sufficiently large in this inequality. Actually such an estimate will be proved in Theorem 8.6 which can be considered as the multivariate generalization of Theorem 4.2.

The condition about the countable cardinality of the set $\mathcal{F}$ in Theorem 4.2 could be weakened similarly to Theorem 4.1. But 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 the class of functions $\mathcal{F}$ considered in Theorem 4.1 had to be L_2-dense, while in Theorem 4.2 only a weaker version of this property was needed. In that result it was only demanded that there exists a subset of $\mathcal{F}$ of relatively small cardinality which is dense in the $L_2(\mu)$ norm. In the L_2-density property imposed in Theorem 4.1 a similar property was demanded for all probability measures ν. It may seem strange why such a property was demanded for such probability measures ν 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 a lot of new conditioned measures appear, and the L_2-density property is needed to work with all of them. One would also like to know some results that enable us to check when this condition holds. In the next section a popular notion, the notion of Vapnik–Červonenkis classes will be introduced, and it will be shown 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 $\sqrt{n}\sigma^2 > u$, and no such condition was imposed in formula (4.7). The appearance of this condition in Theorem 4.1 can be explained by comparing this result with the results discussed in Section 3. As we have seen, we do not loose much information if we restrict our attention to the case $u \le \text{const.}\, V_n^2 = \text{const.}\, n\sigma^2$ in Bernstein's inequality (if sums of independent and identically distributed random variables are considered). Theorem 4.1 gives an almost as good estimate for the supremum of normalized partial sums, under appropriate conditions for the class $\mathcal{F}$ of functions we consider in this theorem, as Berstein's inequality for the normalized partial sums of independent and identically distributed random variables with a variance bounded by σ^2. But to get such a result it was enough to consider only the case $\sqrt{n}\sigma^2 > u$. It has also a natural 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 $\mathcal{F}_\sigma = \{f_{s,t}\colon \int f_{s,t}^2(u)\,du = t - s \le$

σ^2, s and t 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 $P\left(\sup_{f\in\mathcal{F}_\sigma} Z(f) > u\right) \le e^{-\text{const.}\,(u/\sigma)^2}$. 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 $P\left(\sup_{f\in\mathcal{F}_\sigma} Z(f) > u\right) \to 1$, 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 the integrals $J_n(f_{s,t})$ of the functions from the same the class $\mathcal{F}_\sigma$ are considered 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. However the details of such an argument will be omitted.

At a heuristic level it is clear that if Theorem 4.1 is considered with such a class of functions $\mathcal{F}$ which is L_2-class with a large exponent L then for the validity of relation (4.4) such a lower bound has to be imposed for u where the expression $\sqrt{n}\sigma \log^{1/2} \frac{2}{\sigma}$ is multiplied with a large coefficient. A similar statement can be told about 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 one 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. The next simple Lemma 4.3 yields 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 be given on a space $(Y, \mathcal{Y})$. 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 $P^*(A(U) \setminus B(u)) \le \lim\limits_{\varepsilon\to 0} P(B(u-\varepsilon)\setminus B(u)) = 0$, where $P^*(X)$ denotes the outer measure of a not necessarily measurable set $X \subset \Omega$, since $\bigcap\limits_{\varepsilon\to 0} B(u-\varepsilon) = B(u)$, 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 with respect to a normalized empirical measure. 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 $(X, \mathcal{X})$ we define the integral $J_{n,k}(f)$ of the function f with respect to the k-fold product of the normalized version of the 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)

In the case $k \ge 2$ it will be assumed that the probability measure μ has no atoms.

Lemma 4.3 enables us to prove that certain classes of random integrals $J_{n,k}(f)$, $f \in \mathcal{F}$, defined with the help of some set of functions $f \in \mathcal{F}$ of k variables are countably approximable. I present an example of a class of such random integrals which is important in certain applications.

Let us consider the case when $X = R^s$, the s-dimensional Euclidean space with some $s \ge 1$. Given some $u = (u^{(1)}, \dots, u^{(s)}) \in R^s$, $v = (v^{(1)}, \dots, v^{(s)}) \in R^s$ such that $u < v$, i.e. $u^{(j)} < v^{(j)}$ for all $1 \le j \le s$, let $B(u, v)$ denote the s-dimensional rectangle $B(u, v) = \{z\colon u < z < v\}$. Let us fix some function $f(x_1, \dots, x_k)$ of k variables such that $\sup |f(x_1, \dots, x_k)| \le 1$, 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 $(u_1, v_1) \times \cdots \times (u_k, v_k)$, and it is zero outside of this rectangle. Let us call a class of functions $\mathcal{F}$ consisting of functions of the form $f_{u_1,\dots,u_k,v_1,\dots,v_k}$ closed if it has the following property. If $f_{u_1,\dots,u_k,v_1,\dots,v_k} \in \mathcal{F}$ for some vectors $(u_1, \dots, u_k)$ and $(v_1, \dots, v_k)$, and $u_j \le \bar{u}_j < \bar{v}_j \le v_j$, $1 \le j \le k$, then $f_{\bar{u}_1,\dots,\bar{u}_k,\bar{v}_1,\dots,\bar{v}_k} \in \mathcal{F}$. In Lemma 4.4 it will be proved that the random integrals introduced in formula (4.8) of functions from a closed class $\mathcal{F}$ constitute a countably approximable class.

Lemma 4.4. *Let us have a function f on the Euclidean space R^{ks} such that the $|f| \le 1$ in all points, and consider a closed class $\mathcal{F}$ of functions of the form $f_{u_1,\dots,u_k,v_1,\dots,v_k} \in (R^{sk}, \mathcal{B}^{sk})$, $u_j, v_j \in R^s$, $u_j \le v_j$, $1 \le j \le k$, introduced in the previous paragraph with the help of this function f. Let us take n independent and identically distributed random variables $\xi_1, \dots, \xi_n$ with some distribution μ and values in the space $(R^s, \mathcal{B}^s)$. Let μ_n denote the empirical distribution of this sequence. Then the class of random integrals $J_{n,k}(f_{u_1,\dots,u_k,v_1,\dots,v_k})$ defined in formula (4.8) with functions $f_{u_1,\dots,u_k,v_1,\dots,v_k} \in \mathcal{F}$ 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 $0 < \varepsilon < 1$ 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 $u_j \le \bar{u}_j < \bar{v}_j \le v_j$ such that the sets $K_j = B(u_j, v_j) \setminus B(\bar{u}_j, \bar{v}_j)$ 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

$$|J_{n,k}(f_{\bar{u}_1,\dots,\bar{u}_k,\bar{v}_1,\dots,\bar{v}_k})(\omega) - J_{n,k}(f_{u_1,\dots,u_k,v_1,\dots,v_k})(\omega)| \le \varepsilon. \tag{4.9}$$

Then 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 difference of integrals at its left-hand side 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 $x_j = x_{\bar{j}}$, $1 \le j, \bar{j} \le k$, $j \ne \bar{j}$, 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 $F_n(x) = F_n(x, \omega) = \frac{1}{n}\#\{j\colon 1 \le$

$j \le n$, $\xi_j(\omega) < x\}$ is introduced, and the difference $F_n(x) - F(x)$ is considered. This result states that $\sup_x |F_n(x) - F(x)|$ tends to zero with probability one.

Observe that $\sup_x |F_n(x) - F(x)| = n^{-1/2} \sup_{f \in \mathcal{F}} |J_n(f)|$, 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 $P\left(\sup_{f \in \mathcal{F}} |J_n(f)| > u\right)$. 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 $P\left(n^{-1/2} \sup_{f \in \mathcal{F}} |J_n(f)| > u\right) \le e^{-Cnu^2}$ if $1 \ge u \ge \bar{C} n^{-1/2}$ 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. The classes of L_2-dense classes played an important role in the study of the previous section. The results of this section may help to find interesting classes of functions with this property. Some of the results formulated in this section will be proved in Appendix A.

First I recall the following notions.

Definition of Vapnik-Červonenkis classes of sets and functions. *Let a set X be given, and let us select a class $\mathcal{D}$ of subsets of this set X. 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 $S(n) = \{x_1, \dots, x_n\} \subset X$ of cardinality n of the set X the collection of sets of the form $S(n) \cap D$, $D \in \mathcal{D}$, contains no more than Bn^K subsets of $S(n)$. 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 $X = Y \times R^1$.

The following result which was first proved by Sauer is of fundamental importance in the theory of Vapnik–Červonenkis classes. This result provides a relatively simple condition for a class $\mathcal{D}$ of subsets of a set X to be a Vapnik–Červonenkis class. Its proof is given in Appendix A. Before its formulation I introduce some terminology which seems to be wide spread and generally accepted in the literature.

Definition of shattering of a set. *Let a set S and a class $\mathcal{E}$ of subsets of S be given. A finite set $F \subset S$ is called shattered by the class $\mathcal{E}$ if all its subsets $H \subset F$ can be written in the form $H = E \cap F$ with some element $E \in \mathcal{E}$ of the class of sets of $\mathcal{E}$.*

Theorem 5.1. (Sauer's lemma). *Let a finite set $S = S(n)$ consisting of n elements be given together with a class $\mathcal{E}$ of subsets of S. If $\mathcal{E}$ shatters no subset of S of cardinality k, then $\mathcal{E}$ contains at most $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k-1}$ subsets of S.*

The estimate of Sauer's lemma is sharp. Indeed, if $\mathcal{E}$ contains all subsets of S of cardinality less than or equal to $k-1$, then it shatters no subset of a set F of cardinality k (a set F of cardinality k cannot be written in the form $E \cap F$, $E \in \mathcal{E}$), and $\mathcal{E}$ contains $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k-1}$ subsets of S.

Let us have a set X and a class of subsets $\mathcal{D}$ of it. One may be interested in when $\mathcal{D}$ is a Vapnik–Červonenkis class. Sauer's lemma gives a most useful condition for it. Namely, it implies that if there exists a positive integer k such that the class $\mathcal{D}$ shatters no subset of X of cardinality k, then $\mathcal{D}$ is a Vapnik–Červonenkis class. Indeed, let us take some number $n \geq k$, fix an arbitrary set $S(n) = \{x_1, \ldots, x_n\} \subset X$ of cardinality n, and introduce the class of subsets $\mathcal{E} = \mathcal{E}(S(n)) = \{S(n) \cap D\colon\ D \subset \mathcal{D}\}$. If the above condition is satisfied, then $\mathcal{E}$ shatters no subset of $S(n)$ of cardinality k, hence by Sauer's lemma the class $\mathcal{E}$ contains at most $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k-1}$ elements. Let me remark that it is also proved that $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k-1} \leq 1.5\frac{n^{k-1}}{(k-1)!}$ if $n \geq k+1$. This estimate gives a bound on the parameter and exponent of a Vapnik–Červonenkis class which satisfies the above condition.

Moreover, Theorem 5.1 also has the following consequence. Take an (infinite) set X and a class of its subsets $\mathcal{D}$. There are two possibilities. Either there is some set $S(n) \subset X$ of cardinality n for all integers n such that $\mathcal{E}(S(n))$ contains all subsets of $S(n)$, i.e. $\mathcal{D}$ shatters this set, or $\sup\limits_{S\colon\, S \subset X,\, |S|=n} |\mathcal{E}(S)|$ tends to infinity at most in a polynomial order as $n \to \infty$, where $|S|$ and $|\mathcal{E}(S)|$ denote the cardinality of S and $\mathcal{E}(S)$.

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. (A relation between the L_1-dense class and Vapnik–Červonenkis class property). *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 $\sup\limits_{y \in Y} |f(y)| \leq 1$ 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 $B \geq 1$ 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 $0 < \varepsilon \le 1$. We are going to show that 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 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 $\bar{\mathcal{D}}(\varepsilon, \nu) = \{f_1, \dots, f_M\} \subset \mathcal{F}$ 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 $f_{M+1} \in \mathcal{F}$ can be attached to $\bar{\mathcal{D}}(\varepsilon, \nu)$ without violating this condition. Thus the inequality $M \le D\varepsilon^{-L}$ means that $\bar{\mathcal{D}}(\varepsilon, \nu)$ is an ε-dense subset of $\mathcal{F}$ in the space $L_1(Y, \mathcal{Y}, \nu)$ with no more than $D\varepsilon^{-L}$ elements.

In the estimation of the cardinality M of a (finite) set $\mathcal{D}(\varepsilon, \nu) = \{f_1, \dots, f_M\}$ with the property $\int |f_j - f_k| \, d\nu \ge \varepsilon$ if $j \ne k$ the Vapnik–Červonenkis class property of $\mathcal{F}$ is exploited 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 $A(f_j) = \{(y, t)\colon\ y \in Y,\ \min(0, f_j(y)) \le t \le \max(0, f_j(y))\}$, $f_j \in \mathcal{D}(\varepsilon, \nu) \subset \mathcal{F}$ 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 there exists a set $S_0(p)$ with the above property and with a relatively small number p, then this yields 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 the estimate $P((y_l, t_l) \in A(f_j)\Delta A(f_k)) = \frac{1}{2}(\nu \times \lambda)(A(f_j)\Delta A(f_k)) = \frac{1}{2}\int |f_j - f_k| \, d\nu \ge \frac{\varepsilon}{2}$ hold, where λ denotes the Lebesgue measure. This implies that the probability that the (random) sets $A(f_j) \cap S_0(p)$ 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 $\frac{1}{8}$, 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 $M \ge 1$ 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 $\sup_{y \in Y} |f(y)| \le 1$ 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 probability measure ν on $(Y, \mathcal{Y})$ together with a number $0 < \varepsilon \le 1$, and $\mathcal{D}(\varepsilon, \nu) = \{f_1, \dots, f_M\}$ is an $\frac{\varepsilon^2}{2}$-dense set of $\mathcal{F}$ in the space $L_1(Y, \mathcal{Y}, \nu)$, $M \le 2^L D\varepsilon^{-2L}$, then for all function $f \in \mathcal{F}$ some function $f_j \in \mathcal{D}(\varepsilon, \nu)$ can be chosen in such a way that $\int (f - f_j)^2 \, d\nu \le 2 \int |f - f_j| \, d\nu \le \varepsilon^2$. This implies that $\mathcal{F}$ is 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 X 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 given in Appendix A.

Theorem 5.3. (A way to construct Vapnik–Červonenkis classes). *Let us consider a k-dimensional subspace $\mathcal{G}_k$ of the linear space of real valued functions defined on a set X, and define the level-set $A(g) = \{x\colon\ x \in X,\, g(x) \ge 0\}$ 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 X consisting of the above introduced level sets. No subset $S = S(k+1) \subset X$ of cardinality $k+1$ is shattered by $\mathcal{D}$. Hence by Theorem 5.1 $\mathcal{D}$ is a Vapnik–Červonenkis class of subsets of X.*

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 complementary sets $\mathcal{C}^c = \{S \setminus C\colon\ C \in \mathcal{C}\}$ 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.

Lemma 5.4. (Some useful properties of L_2-dense classes). *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 $\mathcal{G}_1 + \mathcal{G}_2 = \{g_1 + g_2\colon\ g_1 \in \mathcal{G}_1,\, g_2 \in \mathcal{G}_2\}$, $\mathcal{G}_1 \cdot \mathcal{G}_2 = \{g_1 g_2\colon\ g_1 \in \mathcal{G}_1,\, g_2 \in \mathcal{G}_2\}$, $\min(\mathcal{G}_1, \mathcal{G}_2) = \{\min(g_1, g_2)\colon\ g_1 \in \mathcal{G}_1,\, g_2 \in \mathcal{G}_2\}$, $\max(\mathcal{G}_1, \mathcal{G}_2) = \{\max(g_1, g_2)\colon\ g_1 \in \mathcal{G}_1,\, g_2 \in \mathcal{G}_2\}$ 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 was 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 a Vapnik-Červonenkis class of sets is an L_2-dense class of functions, and 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.

In this section some results will be proved by means of a simple but useful method, called the chaining argument. This enables us to prove Theorem 4.2, but if the supremum of sums of independent and identically distributed random variables is considered, then it yields a much weaker estimate than Theorem 4.1, which will be presented in Proposition 6.1. But even this result turned out to be useful, because it enables to reduce the proof of Theorem 4.1 to the proof of a weaker version of it formulated in Proposition 6.2. It will be shown that Propositions 6.1 and 6.2 together imply Theorem 4.1. The proof of Proposition 6.2 which is based on a symmetrization argument is postponed to the next section.

The method of proof of Theorem 4.2 does not suffice to prove Theorem 4.1, because we have relatively weak estimates about the tail 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 an earlier point. Proposition 6.1 contains the result that can be obtained in such a way. Before the study of this problem we shall 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 let $f_p \in \mathcal{F}_p$. 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

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

and

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

The above estimates together with the relation $\bigcup_{p=0}^{\infty} \mathcal{F}_p = \mathcal{F}$ which implies that $\{|Z(f)| \ge u\} \subset \bigcup_{p=1}^{\infty} \bigcup_{j=1}^{m_p} A(p,j) \cup \bigcup_{s=1}^{m_0} B(s)$ 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 $M \ge 16$ (and $L \ge 1$), then

$$2^{2pL}\sigma^{-L}\exp\left\{-\frac{2^{2p}u^2}{256\sigma^2}\right\} \le \left(\frac{1}{2}\right)^{-2pL}\sigma^{-L}\left(\frac{\sigma}{2}\right)^{2^{2p}M^2L/256} \le 2^{-pL} \le 2^{-p}$$

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 an 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 (universal) constants in this estimate are chosen sufficiently large.

The proof of Theorem 4.2 was based on a sufficiently good estimate on the probabilities $P(|Z(f) - Z(g)| > u)$ for pairs of functions $f, g \in \mathcal{F}$ and numbers $u > 0$. In the case of Theorem 4.1 only a weaker bound can be given for the corresponding probabilities. There is no good estimate on the tail 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, where the tail distribution of 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 will be given in Proposition 6.1. Another result will

be formulated in Proposition 6.2 whose proof is postponed to the next section. It will be shown that Theorem 4.1 follows from Propositions 6.1 and 6.2.

Before the formulation of Proposition 6.1 I recall an estimate which is a simple consequence of Bernstein's inequality: If $S_n(f) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(\xi_j)$ 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 a version of (6.1) 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 relations (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 $\xi_1, \dots, \xi_n$, $n \ge 2$, and define the normalized random sums $S_n(f) = \frac{1}{\sqrt{n}} \sum_{l=1}^{n} f(\xi_l)$, 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 $u > 0$ such that $n\sigma^2 \ge \left(\frac{u}{\sigma}\right)^2 \ge ML \log \frac{2}{\sigma}$ 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 $\bigcup_{j=1}^{m} \mathcal{D}_j = \mathcal{F}$, and the normalized random 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 $\mathcal{F}$. Beside this, 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 $n\bar{\sigma}^2 \ge \frac{M^{2/3}(L+\beta)\log n}{1000\bar{A}^{4/3}}$ 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 $\sup_{f \in D_j} |S_n(f - f_j)|$ 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. Its formulation contains some parameters. In the proof of Theorem 4.1 with the help of Propositions 6.1 and 6.2 the parameters appearing in these Propositions must be fitted to each other. The parameter $\bar{A} \ge 2$ in Proposition 6.1 was introduced to make this fitting simpler, and this was also the reason to formulate inequality (6.2) under the condition $M \ge M_0(\bar{A})$ with a bound $M_0(\bar{A})$ depending on $\bar{A}$. We wanted to guarantee the validity of Proposition 6.1 with such a number $\bar{\sigma} = \bar{\sigma}(u)$ which also satisfies the inequality $n\bar{\sigma}^2 \ge K \log n$ holds with a previously fixed number $K > 0$. The last relation in Proposition 6.1 shows that this is possible if first the number $\bar{A}$ and then the number $M_0 = M_0(\bar{A})$ is chosen sufficiently large.

Now I 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 $K > 0$ 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 normalized random sums $S_n(f)$, $f \in \mathcal{F}$, introduced in Theorem 4.1 satisfy the inequality*

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

(A possible choice of the parameters is $K = 4$, $A_0 = 2^{10} \cdot 10^{16}$ and $\gamma = \frac{1}{2}$.)

I did not try to find optimal parameters in formula (6.3). Even the exponent $\frac{1}{2}$ 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 form $P\left(\sup_{f \in \mathcal{F}} |S_n(f)| \ge u\right)$ with some class of functions $\mathcal{F}$. 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 $u = An^{1/2}\sigma^2$ with $A \ge A_0$ is taken where A_0 is a sufficiently large positive number. Let us observe that in this case 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 $P(S_n(f) > An^{1/2}\sigma^2) = P\left(\sum_{l=1}^{n} f(\xi_l) > uV_n\right) < e^{-\text{const.}\, An\sigma^2}$ with $u = A\sqrt{n}\sigma$ 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, since 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 such a 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 chaining argument applied in the proof of Theorem 4.1 does not give a good estimate on the expression at the left-hand side of (6.3). 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 Proposition 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 $\mathcal{F} = \left\{ \frac{g - f_j}{2} \colon\ g \in \mathcal{D}_j \right\}$ introduced in Proposition 6.1, where f_j is the function appearing in the definition of the class of functions $\mathcal{D}_j$. Clearly,

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

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 first term at the right-hand side can be well bounded by means of Proposition 6.1, the second term can be well bounded by means of Proposition 6.2, and Theorem 4.1 follows from these estimates.

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 the number $M \ge M_0(\bar{A})$ in Proposition 6.1 with such a treshold $M_0 = M_0(\bar{A})$ (this number M can be chosen also in formula (4.4) of Theorem 4.1) which also satisfies the inequality $\frac{M_0^{2/3}}{1000\bar{A}^{4/3}} \ge K$ 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 $\bar{A}_0 \ge A_0$ Proposition 6.2 yields the estimation

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

(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 estimate in 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 $\gamma\bar{A}_0^{1/2}n\bar{\sigma}^2\ge\frac{1}{K}n\bar{\sigma}^2\ge L\log n\ge 2L\log\frac{1}{\bar{\sigma}}$, i.e. $\bar{\sigma}^{-L}\le e^{\gamma\bar{A}_0^{1/2}n\bar{\sigma}^2/2}$. By the estimates of Proposition 6.1 $n\bar{\sigma}^2\ge\frac{1}{64}\left(\frac{u}{\bar{A}\sigma}\right)^2$. 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 $\inf_{1\le j\le m_p}\int(f-f_{a(p,j)})^2\,d\mu\le 2^{-4p}\sigma^2$ 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 $\int\left(\frac{f_{a(p,j)}-f_{a(p-1,j')}}{2}\right)^2 d\mu\le 4\sigma^2 2^{-4p}$ and $\sup_{x_j\in X,\,1\le j\le k}\left|\frac{f_{a(p,j)}(x_1,\dots,x_k)-f_{a(p-1,j')}(x_1,\dots,x_k)}{2}\right|\le 1$. Relation (6.1) yields that

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

and

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

Choose an integer number R, $R\ge 0$, such that $\frac{2^{6(R+1)}}{256}\left(\frac{u}{\bar{A}\sigma}\right)^2>n\sigma^2\ge\frac{2^{6R}}{256}\left(\frac{u}{\bar{A}\sigma}\right)^2$, 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 $\bar{A}\ge 2$ 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 $m_R \le D2^{2RL}\sigma^{-L} = D\bar\sigma^{-L}$, 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 $\bigcup_{j=1}^{m} \mathcal{D}_j = \mathcal{F}$. Beside this, with our choice of the number R 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

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

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.

The inequality

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

holds, and this implies (together with the relation $R \ge 1$) that

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

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 $\sigma \ge n^{-1/3}$, we can exploit that the inequality $2^{6R}\left(\frac{u}{\bar{A}\sigma}\right)^2 \le 256 n\sigma^2$ holds because of the definition of the number R. It can be rewritten as $2^{-4R} \ge 2^{-16/3}\left[\frac{\left(\frac{u}{\bar{A}\sigma}\right)^2}{n\sigma^2}\right]^{2/3}$. Hence $n\bar{\sigma}^2 = 2^{-4R}n\sigma^2 \ge \frac{2^{-16/3}}{\bar{A}^{4/3}}(n\sigma^2)^{1/3}\left(\frac{u}{\sigma}\right)^{4/3}$. Since $n\sigma^2 \ge n^{1/3}$ and $(\frac{u}{\sigma})^2 \ge M(L+\beta)^{3/2}\log\frac{2}{\sigma} \ge \frac{M}{3}(L+\beta)^{3/2}$, 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.

This section contains the proof of Proposition 6.2 with the help of a symmetrization argument which completes the proof of Theorem 4.1. By symmetrization argument I mean the reduction of the investigation of sums of the form $\sum f(\xi_j)$ to sums of the form $\sum \varepsilon_j f(x_j)$, where ε_j are independent random variables, independent also of the random variables ξ_j, and $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = \frac{1}{2}$. 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*

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

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

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

Proof of Lemma 7.1. Put $\tau = \min\{n\colon |Z_n| > u+\alpha\}$ 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

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

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 and identically distributed random variables $\xi_1, \dots, \xi_n$ with values in the space $(X, \mathcal{X})$ 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 $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = \frac{1}{2}$, $1 \le j \le n$, independent also of the random sequence $\xi_1, \dots, \xi_n$. Then*

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

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 $S_n(f) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} f(\xi_j)$ and $\bar{S}_n(f) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} f(\bar{\xi}_j)$ for all $f \in \mathcal{F}$. The inequality

$$P\left(\sup_{f\in\mathcal{F}}|S_n(f)| > A\sqrt{n}\sigma^2\right) \le 2P\left(\sup_{f\in\mathcal{F}}|S_n(f) - \bar{S}_n(f)| > \frac{2}{3}A\sqrt{n}\sigma^2\right). \tag{7.3}$$

follows from Lemma 7.1 if it is applied for the countable set of random variables $Z_n(f) = S_n(f)$ and $\bar{Z}_n(f) = \bar{S}_n(f)$, $f \in \mathcal{F}$, and the numbers $u = \frac{2}{3}A\sqrt{n}\sigma^2$ and $\alpha = \frac{1}{3}A\sqrt{n}\sigma^2$, since the random fields $S_n(f)$ and $\bar{S}_n(f)$ are independent, and $P(|\bar{S}_n(f)| \le \alpha) > \frac{1}{2}$ for all $f \in \mathcal{F}$. Indeed, $\alpha = \frac{1}{3}A\sqrt{n}\sigma^2 \ge \sqrt{2}\sigma$, $E\bar{S}_n(f)^2 \le \sigma^2$, thus Chebishev's inequality implies that $P(|\bar{S}_n(f)| \le \alpha) \ge P(|\bar{S}_n(f)| \le \sqrt{2}\sigma) \ge \frac{1}{2}$ for all $f \in \mathcal{F}$.

Let us observe that the random field

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

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 those indices j for which $\varepsilon_j = -1$ and do not change them for those indices j for which $\varepsilon_j = 1$. On the other hand, the distribution of the random field obtained in such a way agrees with the conditional distribution of the random field defined in (7.4′) under the condition that the values of the random variables ε_j are prescribed.

The above relation together with formula (7.3) imply that

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

Lemma 7.2 is proved.

Let me briefly explain the method of proof of Proposition 6.2. A probability of the form $P\left(n^{-1/2} \sup_{f \in \mathcal{F}} \left| \sum_{j=1}^{n} f(\xi_j) \right| > u \right)$ has to be estimated. Lemma 7.2 enables us to replace this problem by the estimation of the probability $P\left(n^{-1/2} \sup_{f \in \mathcal{F}} \left| \sum_{j=1}^{n} \varepsilon_j f(\xi_j) \right| > \frac{u}{3} \right)$ with some independent random variables ε_j, $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = \frac{1}{2}$, $j = 1, \dots, n$, 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. We shall show that such an approach leads to the estimation of the probability $P\left(n^{-1/2}\sup_{f\in\mathcal{F}'}\left|\sum_{j=1}^{n}f(\xi_j)\right|>u^{1+\alpha}\right)$ 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. Nevertheless, it turned out to be useful to turn to the investigation of this new inequality. We shall exploit that the number u is replaced by a larger number $u^{1+\alpha}$ when turning to this new inequality. Let us also observe that if the sum of bounded random variables is considered, then for very large values u the probability we investigate equals zero, since the sum of bounded random variables is bounded with probability 1.On the basis of these observations such a backward induction procedure can be worked out, in which the lower bound of the numbers u for which we can give a good upper bound on the probability $P\left(n^{-1/2}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^{n}f(\xi_j)\right|>u\right)$ is diminished at each step.

Proposition 6.2 contains the estimate we can get with the help of the above argument. To work out the details we introduce the following notion.

Definition of good tail behaviour for a class of normalized random sums. *Let us have some measurable space $(X,\mathcal{X})$ and a probability measure μ on it together with some integer $n\geq 2$ and real number $\sigma>0$. Consider some class $\mathcal{F}$ of functions $f(x)$ on the space $(X,\mathcal{X})$, and take a sequence of independent and μ distributed random variables $\xi_1,\dots,\xi_n$, with values in the space $(X,\mathcal{X})$. Define the normalized random sums $S_n(f)=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}f(\xi_j)$, $f\in\mathcal{F}$. Given some real number $T>0$ we say that the set of normalized random sums $S_n(f)$, $f\in\mathcal{F}$, has a good tail behaviour at level T (with parameters n and σ^2 which will be fixed in the sequel) if the inequality*

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

holds for all numbers $A>T$.

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

Proposition 7.3. *Let us fix a positive integer $n\geq 2$, 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 $L\geq 1$ and parameter D. Let us also assume that all functions $f\in\mathcal{F}$ satisfy the conditions $\sup_{x\in X}|f(x)|\leq\frac{1}{4}$, $\int f^2(x)\mu(\,dx)\leq\sigma^2$, and $n\sigma^2>K(L+\beta)\log n$ with a sufficiently large fixed number K and $\beta=\max\left(\frac{\log D}{\log n},0\right)$.*

Let a number $T>1$ be such that for all classes of functions $\mathcal{F}$ which satisfy the above conditions the class of normalized random sums $S_n(f)=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}f(\xi_j)$, $f\in\mathcal{F}$, 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 $T^{4/3}$. There is a universal constant $\bar{A}_0$ such that if $T \ge \bar{A}_0$ and the number T has the above property, then the number $\bar{T} = T^{3/4}$ also has it. The above result holds for instance with the choice $\bar{A}_0 = 64 \cdot 10^{12}$ and $K = 1$.

Proposition 6.2 simply follows from Proposition 7.3. To show this let us first observe that a 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 this class of functions $\mathcal{F}$ satisfies the conditions of Proposition 7.3. Indeed, in this case $P\left(\sup_{f\in\mathcal{F}} |S_n(f)| \ge A\sqrt{n}\sigma^2\right) \le P\left(\sup_{f\in\mathcal{F}} |S_n(f)| > \frac{\sqrt{n}}{4}\right) = 0$ for all $A > T_0$. Then the repetitive application of Proposition 7.3 yields that a class of random sums $S_n(f)$, $f \in \mathcal{F}$, has a good tail behaviour at all levels $T \ge T_0^{(3/4)^j}$ with an index j such that $T_0^{(3/4)^j} \ge \bar{A}_0$ if the class of functions $\mathcal{F}$ satisfies the conditions of Proposition 7.3. Hence it has a good tail behaviour for $T = \bar{A}_0^{4/3}$. If a 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 7.3, (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 $\mathcal{F}$ satisfies formula (6.3) in Proposition 6.2 with $4K$, $A_0 = 4\bar{A}_0^{4/3}$ and $\gamma = \frac{1}{2}$, and this is what we had to show.

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, $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = \frac{1}{2}$, $1 \le j \le n$, and investigate the conditional probability

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

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

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

with

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

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

$$P((\xi_1,\dots,\xi_n)\in H)\le e^{-A^{2/3}n\sigma^2}\quad \text{if } A\ge T. \tag{7.8}$$

(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 (7.8) 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 $\bar{\mathcal{F}}=\{\bar f(f)\colon\ f\in\mathcal{F}\}$. Let us show that the class of functions $\bar{\mathcal{F}}$ satisfies the conditions of Proposition 7.3, hence the estimate (7.5) holds for the class of functions $\bar{\mathcal{F}}$ 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 $\int\bar f^2(x)\mu(\,dx)\le\int f^4(x)\mu(\,dx)\le\frac{1}{16}\int f^2(x)\,\mu(\,dx)\le\frac{\sigma^2}{16}<\sigma^2$ if $f\in\mathcal{F}$. It remained to show that $\bar{\mathcal{F}}$ is an L_2-dense class with exponent L and parameter D. For this goal we need a good estimate on $\int(\bar f(x)-\bar g(x))^2\rho(\,dx)$, where $\bar f,\,\bar g\in\bar{\mathcal{F}}$, and ρ is an arbitrary probability measure.

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 $\bar{\mathcal{F}}$ in the space $L_2(X,\mathcal{X},\rho)$, and not only $\mathcal{F}$, but also $\bar{\mathcal{F}}$ is an L_2-dense class with exponent L and parameter D.

An application of the conditions of Proposition 7.3 for the number $A^{4/3}\ge T^{4/3}$ and the class of functions $\bar{\mathcal{F}}$ yields that

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

i.e. relation (7.8) holds.

By formula (7.6) and the definition of the set H given in (7.7) the estimate

$$P(f,A|\xi_1,\dots,\xi_n)\le 2e^{-A^{2/3}n\sigma^2/144}\quad\text{if } (\xi_1,\dots,\xi_n)\notin H \tag{7.9}$$

holds 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

$$P(\mathcal{F},A|\xi_1,\dots,\xi_n)=P\left(\sup_{f\in\mathcal{F}}\frac{1}{\sqrt n}\left|\sum_{j=1}^n\varepsilon_jf(\xi_j)\right|\ge\frac{A}{3}\sqrt n\sigma^2\,\middle|\,\xi_1,\dots,\xi_n\right)$$

holds for all $(\xi_1,\dots,\xi_n)$ and $A\ge T$. We shall estimate this conditional probability with the help of relation (7.9) if $(\xi_1,\dots,\xi_n)\notin H$. Given some set consisting 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 $(X, \mathcal{X})$ which is concentrated in the points $x_1, \dots, x_n$, and $\nu(\{x_j\}) = \frac{1}{n}$ for all points x_j, $j = 1, \dots, n$. If $\int f^2(x)\nu(\,dx) \le \delta^2$ for a function f, then $\left| \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j f(x_j) \right| \le n^{1/2} \int |f(x)| \nu(\,dx) \le n^{1/2}\delta$. Since the condition $n\sigma^2 \ge K(L+\beta)\log n$ in Proposition 7.3 also implies that $n\sigma^2 \ge 1$ (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 $\left| \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j f(x_j) - \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j g(x_j) \right| \le \frac{A}{6\sqrt{n}} \le \frac{A}{6}\sqrt{n}\sigma^2$.

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 $\frac{1}{\sqrt{n}} \left| \sum_{j=1}^n \varepsilon_j f(\xi_j) \right| \ge \frac{A}{3}\sqrt{n}\sigma^2$ for some function $f \in \mathcal{F}$, then $\frac{1}{\sqrt{n}} \left| \sum_{j=1}^n \varepsilon_j f_l(\xi_j) \right| \ge \frac{A}{6}\sqrt{n}\sigma^2$ 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 $P(\mathcal{F}, A|\xi_1, \dots, \xi_n) \le \sum_{l=1}^m P(f_l, A|\xi_1, \dots, \xi_n)$ with these functions $f_1, \dots, f_m$, and relation (7.9) yields that

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

This inequality together with Lemma 7.2 (under the restriction that $A \ge \bar{A}_0 \ge \frac{3\sqrt{2}}{\sqrt{n}\sigma} \ge 3\sqrt{2}$) and estimate (7.8) imply that

$$\begin{aligned} P\left(\frac{1}{\sqrt{n}} \sup_{f\in\mathcal{F}} \left| \sum_{j=1}^n f(\xi_j) \right| \ge An^{1/2}\sigma^2 \right) &\le 4P\left(\frac{1}{\sqrt{n}} \sup_{f\in\mathcal{F}} \left| \sum_{j=1}^n \varepsilon_j f(\xi_j) \right| \ge \frac{A}{3} n^{1/2}\sigma^2 \right) \\ &\le 8D(6n)^L e^{-A^{2/3}n\sigma^2/144} + 4e^{-A^{2/3}n\sigma^2} \quad \text{if } A \ge T. \end{aligned} \tag{7.10}$$

By the conditions of Proposition 7.3 the inequality $n\sigma^2 \ge K(L+\beta)\log n = KL\log n + K\log(\max(D,1))$ holds. Hence the first term at the right-hand side of (7.10) can be bounded by

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

if $A \ge T \ge \bar{A}_0 \ge 64 \cdot 10^{12}$ and $K \ge 1$. (With such parameters $\frac{A^{2/3}}{144} - A^{1/2} \ge A^{1/2}$, and $e^{-A^{1/2}n\sigma^2} \le n^{-LA^{1/2}} \max(D,1)^{-A^{1/2}}$.) With such a choice of the parameters the

inequality $\frac{3\sqrt{2}}{\sqrt{n}\sigma} \le \frac{3\sqrt{2}}{\sqrt{K \log 2}} \le \bar{A}_0 \le A$, needed for the validity of relation (7.2), also holds. The second term at the right-hand side of (7.10) be bounded as $4e^{-A^{2/3}n\sigma^2} \le \frac{1}{2}e^{-A^{1/2}n\sigma^2}$ with the above choice of the numbers $\bar{A}_0$ and K.

By the above calculation formula (7.10) 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 $\bar{A}_0 = 64 \cdot 10^{12}$ 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 the the tail distribution of multiple stochastic integrals and of their supremum. The supremum results were formulated in Section 4 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. Estimates about the tail distribution of multiple stochastic integrals with respect to a normalized empirical measure have a natural version for U-statistics. Such estimates can be considered as the multivariate counterpart of estimates on sums of independent random variables. It will turn out natural to concentrate on tail distribution estimation for a special class of U-statistics, called the degenerate U-statistics. The definition of U-statistics and degenerate U-statistics will be introduced in this section. U-statistics are natural multivariate versions of sums of independent (and identically distributed) random variables. Degenerate U-statistics have some additional special properties. They can be considered as the multivariate analogs of sums of independent and identically distributed random variables with zero expectation. The proof of the results presented in this section requires a more detailed study of the properties of U-statistics, a problem which is of special interest in itself. This will be the subject of the next section.

A Gaussian counterpart of the above results will also be presented. Some estimates will be formulated about multiple Wiener–Itô integrals and the supremum of such integrals. They are the natural Gaussian analogs of the results about degenerate U-statistics and multiple integrals with respect to normalized empirical measures. Wiener–Itô integrals will be introduced, and their properties will be discussed only in Section 10. Their study yields an invaluable help in understanding the properties of U-statistics and multiple integrals with respect to normalized empirical measures. This section is closed with the presentation of a two-dimensional version of Example 3.2 which shows that certain conditions in the estimates of this Section cannot be omitted.

Let us consider a sequence of independent and identically distributed random variables $\xi_1, \dots, \xi_n$ with values in a measurable space $(X, \mathcal{X})$. Let μ denote the distribution of the random variables ξ_j, and introduce the empirical distribution of the sequence $\xi_1, \dots, \xi_n$ defined in (4.5). Given a measurable function $f(x_1, \dots, x_k)$ on the k-fold

product space $(X^k, \mathcal{X}^k)$ consider 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). In the definition of this integral the diagonals $x_j = x_l$, $1 \le j < l \le k$, were omitted from the domain of integration. 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. (Estimate on the tail distribution of a multiple random integral with respect to a normalized empirical distribution). *Let us take a measurable function $f(x_1, \dots, x_k)$ on the k-fold product $(X^k, \mathcal{X}^k)$ of a measurable 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 independent and identically distributed 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

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

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*

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

with a number σ, $0 \le \sigma \le 1$, satisfying relation 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 $u > n^{k/2}\sigma^{k+1}$. Indeed, in this case Theorem 8.1′ can be applied with $\bar{\sigma} = \left(un^{-k/2}\right)^{1/(k+1)} \ge \sigma$ if $u \le n^{k/2}$, hence also condition $0 < \bar{\sigma} \le 1$ is satisfied. 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 $n^{k/2} \ge u > n^{k/2}\sigma^{k+1}$, and relation (8.3) holds in this case. If $u > n^{k/2}$, then $P(|J_{n,k}(f)| > u) = 0$, and relation (8.3) holds again.

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 the number $0 \le \sigma \le 1$ satisfies relation (8.2), provided that the level u we consider is less than $n^{k/2}\sigma^{k+1}$. As we shall see later (see Corollary 1 of Theorem 9.4), 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 the condition $u \le n^{k/2}\sigma^{k+1}$ is really needed in Theorem 8.1′.

The next result, Theorem 8.2, is the generalization of Theorem 4.1 for multiple random integrals. In its formulation the notions of L_2-dense classes and countably approximability introduced in Section 4 are applied.

Theorem 8.2. (Estimate on the supremum of multiple random integrals with respect to an empirical distribution). *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

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

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 such a version of Theorems 8.1 and 8.2 which corresponds to the results about sums of independent random variables in the case $k = 1$ the following notions will be introduced.

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 in this space* $(X, \mathcal{X})$. *The expression*

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

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

Remark. In later calculations sometimes we shall work with U-statistics with kernel functions of the form $f(x_{u_1}, \dots, x_{u_k})$ instead of $f(x_1, \dots, x_k)$, where $\{u_1, \dots, u_k\}$ is an arbitrary set with different elements. The U-statistic with such a kernel function will also be defined, and it equals the U-statistic with the original kernel function f defined in (8.7), i.e.

$$I_{n,k}(f(x_{u_1}, \dots, x_{u_k})) = I_{n,k}(f(x_1, \dots, x_k)). \tag{8.7'}$$

(Observe that if we define the function $f_\pi(x_1, \dots, x_k) = f(x_{\pi(1)}, \dots, x_{\pi(k)})$ for all permutations π of the set $\{1, \dots, k\}$, then $I_{n,k}(f_\pi) = I_{n,k}(f)$, hence the above definition is legitimate.) Such a definition is natural, and it simplifies the notation in some calculations. A similar convention will be introduced about Wiener–Itô integrals in Section 10.

Some special U-statistics, called degenerate U-statistics, will be also introduced. They can be considered as the natural multivariate version of sums of identically distributed random variables with expectation zero. This notion will be defined together with canonical kernel functions, because they are closely related to each other.

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

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

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

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

For the sake of more convenient notations in the future we shall speak also of U-statistics of order zero. We shall write $I_{n,0}(c) = c$ for any constant c, and $I_{n,0}(c)$ will

be called a degenerate U-statistic of order zero. A constant will be considered also as a canonical function with zero arguments.

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. Let us also observe that

$$I_{n,k}(f) = I_{n,k}(\operatorname{Sym} f) \tag{8.9}$$

for all functions of k variables.

The next two results, Theorems 8.3 and 8.4, deal with degenerate U-statistics. Theorem 8.3 is the U-statistic version of Theorem 8.1 and Theorem 8.4 is the U-statistic version of Theorem 8.2. Actually Theorem 8.3 yields a sharper estimate than Theorems 8.1, because it contains more explicit and better universal constants. I shall return to this point later.

Theorem 8.3. (Estimate on the tail distribution of a degenerate U-statistic). *Let us have a measurable function $f(x_1, \dots, x_k)$ on the k-fold product $(X^k, \mathcal{X}^k)$, $k \geq 1$, of a measurable space $(X, \mathcal{X})$ together with a probability measure μ on $(X, \mathcal{X})$ and a sequence of independent and identically distributed random variables $\xi_1, \dots, \xi_n$ with distribution μ on $(X, \mathcal{X})$. Let us consider the U-statistic $I_{k,n}(f)$ of order k with this sequence of random variables $\xi_1, \dots, \xi_n$. Assume that this U-statistic is degenerate, i.e. its kernel function $f(x_1, \dots, x_k)$ 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 $0 < \sigma \leq 1$. Then 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*

$$P(k!n^{-k/2}|I_{n,k}(f)| > u) \leq A \exp\left\{ -\frac{u^{2/k}}{2\sigma^{2/k}\left(1 + B\left(un^{-k/2}\sigma^{-(k+1)}\right)^{1/k}\right)} \right\} \tag{8.10}$$

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

Let us also formulate the following simple corollary of Theorem 8.3.

Corollary of Theorem 8.3 *Under the conditions of Theorem 8.3 there exist some universal constants $C = C(k) > 0$ and $\alpha = \alpha(k) > 0$ that*

$$P(k!n^{-k/2}|I_{n,k}(f)| > u) \leq C \exp\left\{ -\alpha \left(\frac{u}{\sigma}\right)^{2/k} \right\} \quad \text{for all } 0 \leq u \leq n^{k/2}\sigma^{k+1}. \tag{8.10'}$$

The following estimate holds about the supremum of degenerate U-statistics.

Theorem 8.4. (Estimate on the supremum of degenerate U-statistics). *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 $0 < \sigma \le 1$. Let us take a sequence of independent μ distributed random variables $\xi_1, \dots, \xi_n$, and consider the U statistics $I_{n,k}(f)$ with these random variables and kernel functions $f \in \mathcal{F}$. Let us assume that all these U-statistics $I_{n,k}(f)$, $f \in \mathcal{F}$, are degenerate, 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

$$P\left(\sup_{f\in\mathcal{F}} n^{-k/2}|I_{n,k}(f)| \ge u\right) \le CD\exp\left\{-\alpha\left(\frac{u}{\sigma}\right)^{2/k}\right\} \quad \text{if} \quad n\sigma^2 \ge \left(\frac{u}{\sigma}\right)^{2/k} \ge M(L+\beta)^{3/2}\log\frac{2}{\sigma}, \tag{8.11}$$

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.

I also formulate a Gaussian counterpart of the above results. To do this I need some notions that will be introduced in Section 10. In that section the white noise with a reference measure μ will be defined. It is an appropriate set of jointly Gaussian random variables indexed by the measurable sets $A \in \mathcal{X}$ of a measurable space $(X, \mathcal{X})$, and it also depends on a σ-finite measure μ on $(X, \mathcal{X})$ called the reference measure of the white noise.

In Section 10 it will be also shown that given a white noise μ_W with a non-atomic σ-additive reference measure μ on a measurable space $(X, \mathcal{X})$ and a measurable function $f(x_1, \dots, x_k)$ of k variables on the product space $(X^k, \mathcal{X}^k)$ such that

$$\int f^2(x_1, \dots, x_k)\mu(\,dx_1)\dots\mu(\,dx_k) \le \sigma^2 < \infty \tag{8.12}$$

a k-fold Wiener-Itô integral of the function f with respect to the white noise μ_W

$$Z_{\mu,k}(f) = \frac{1}{k!}\int f(x_1, \dots, x_k)\mu_W(\,dx_1)\dots\mu_W(\,dx_k) \tag{8.13}$$

can be defined, and the main properties of this integral will be proved. It will be seen that Wiener-Itô integrals have a similar relation to degenerate U-statistics and multiple integrals with respect to normalized empirical measures as normally distributed random variables to partial sums of independent random variables. Hence it is useful to find the analogs of the previous estimates in this section to the distribution of Wiener-Itô integrals. The subsequent Theorems 8.5 and 8.6 contain such results.

Theorem 8.5. (Estimate on the tail distribution of a multiple Wiener–Itô integral). *Let us fix a measurable space $(X, \mathcal{X})$ together with a σ-finite non-atomic measure μ on it, and let μ_W be a white noise with reference measure μ on $(X, \mathcal{X})$. If $f(x_1, \dots, x_k)$ is a measurable function on $(X^k, \mathcal{X}^k)$ which satisfies relation (8.12) with some $0 < \sigma < \infty$, then*

$$P(k!|Z_{\mu,k}(f)| > u) \le C \exp\left\{ -\frac{1}{2} \left(\frac{u}{\sigma} \right)^{2/k} \right\} \tag{8.14}$$

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

Theorem 8.6. (Estimate on the supremum of Wiener–Itô integrals). *Let $\mathcal{F}$ be a countable class of functions of k variables on a measurable space $(X, \mathcal{X})$ such that*

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

with some non-atomic σ-additive measure μ on $(X, \mathcal{X})$. Let us assume that there exist some constants $D > 1$ and $L > 0$ such that for all $0 < \varepsilon \le 1$ a subset $\{f_1, \dots, f_m\} \subset \mathcal{F}$ can be chosen with $m = m(\varepsilon) \le D\varepsilon^{-L}$ elements for which

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

Take a white noise μ_W on $(X, \mathcal{X})$ with reference measure μ, and define the Wiener–Itô integrals $Z_{\mu,k}(f)$ for all $f \in \mathcal{F}$. Fix some $0 < \varepsilon \le 1$. The inequality

$$\begin{gathered} P\left(\sup_{f \in \mathcal{F}} k!|Z_{\mu,k}(f)| > u \right) \le CD \exp\left\{ -\frac{1}{2} \left(\frac{(1-\varepsilon)u}{\sigma} \right)^{2/k} \right\} \\ \text{if } u \ge ML^{k/2} \varepsilon^{-k/2} \sigma \left(\log^{k/2} \frac{2}{\sigma} + \log^{k/2} \frac{1}{\varepsilon} \right) \end{gathered} \tag{8.15}$$

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

Formula (8.15) yields an almost as good estimate for the supremum of Wiener–Itô integrals with the choice of a small $\varepsilon > 0$ as formula (8.14) for a single Wiener–Itô integral. But the lower bound imposed on the number u in the estimate (8.15) depends on ε, and for a small number $\varepsilon > 0$ it is large.

The subsequent result presented in Example 8.7 may help to understand why Theorems 8.3 and 8.5 are sharp. Its proof and the discussion of the question about the sharpness of Theorems 8.3 and 8.5 will be postponed to Section 13.

Example 8.7. (A converse estimate to Theorem 8.5). *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 $\sigma > 0$ together with the Wiener–Itô integral $Z_{\mu,k}(f)$ introduced in formula (8.13).

Then the relation $\int f(x_1, \dots, x_k)^2 \, \mu(\, dx_1) \dots \mu(\, dx_k) = \sigma^2$ holds, and the Wiener–Itô integral $Z_{\mu,k}(f)$ satisfies the inequality

$$P(k!|Z_{\mu,k}(f)| > u) \ge \frac{\bar{C}}{\left(\frac{u}{\sigma}\right)^{1/k} + 1} \exp\left\{ -\frac{1}{2} \left(\frac{u}{\sigma}\right)^{2/k} \right\} \quad \textit{for all } u > 0 \tag{8.16}$$

with some constant $\bar{C} > 0$.

The above results show that multiple integrals with respect to a normalized empirical measure or degenerate U-statistics satisfy some estimates similar to those about multiple Wiener–Itô integrals, but they hold under more restrictive conditions. The difference between the estimates in these problems is similar to the difference between the corresponding results in Section 4 whose reason was explained there. Hence they will be only briefly discussed here. The estimates of Theorem 8.1 and 8.3 are similar to that of Theorem 8.5. Moreover, for $0 \le u \le \varepsilon n^{k/2} \sigma^{k+1}$ with a small number $\varepsilon > 0$ Theorem 8.3 yields an almost as good estimate about degenerate U-statistics as Theorem 8.5 yields for a Wiener–Itô integral with the same kernel function f and underlying measure μ. Example 8.7 shows that the constant in the exponent of formula (8.14) cannot be improved, at least there is no possibility of an improvement if only the L_2-norm of the kernel function f is known. Some results discussed later indicate that the estimate of Theorem 8.3 can neither be improved.

The main difference between Theorem 8.5 and the results of Theorem 8.1 or 8.3 is that in the latter case not only an L_2 but also an L_∞ norm is imposed on the kernel function f, and the validity of the estimate is stated only under the condition $u \le n^{k/2} \sigma^{k+1}$. It can be shown that the condition about the L_∞ norm of the kernel function cannot be dropped from the conditions of these theorems, and as a version of Example 3.2 presented at the end of this section shows, in the case $u \gg n^{k/2} \sigma^{k+1}$ the left-hand side of (8.10) may satisfy only a much weaker estimate. This estimate will be given only for $k = 2$, but with some work it can be generalized for general indices k.

Theorems 8.2, 8.4 and 8.6 show that for the tail distribution of the supremum of a not too large class of degenerate U-statistics or multiple integrals a similar upper bound can be given as for the tail distribution of a single degenerate U-statistic or multiple integral, only the universal constants may be worse in the new estimates. However, they hold only under the additional condition that the level at which the tail distribution of the supremum is estimated is not too low. A similar phenomenon appeared already in the results of Section 4. Moreover, such a restriction had to be imposed in the formulation of the results here and in Section 4 for the same reason.

In Theorem 8.2 and 8.4 an L_2-dense class of kernel functions was considered, and this meant that the class of random integrals or U-statistics we consider in this result is not too large. In Theorem 8.6 a similar, but weaker condition was imposed on the class of kernel functions. They had to satisfy a similar condition, but only for the reference measure μ of the white noise appearing in the Wiener–Itô integral. A similar difference

appears in the comparison of Theorems 4.1 or 4.1′ with Theorem 4.2, and this difference has the same reason in the two cases.

I finish this section with the proof of the following Example 8.8 which is a multivariate version of Example 3.2. For the sake of simplicity I restrict my attention to the case $k = 2$.

Example 8.8. (A converse estimate to Theorem 8.3). *Let us take a sequence of independent and identically distributed random variables $\xi_1, \dots, \xi_n$ taking values in the plane $X = R^2$ such that $\xi_j = (\eta_{j,1}, \eta_{j,2})$, $\eta_{j,1}$ and $\eta_{j,2}$ are independent random variables, $P(\eta_{j,1} = 1) = P(\eta_{j,1} = -1) = \frac{\sigma^2}{4}$, $P(\eta_{j,1} = 0) = 1 - \frac{\sigma^2}{2}$, and $P(\eta_{j,2} = 1) = P(\eta_{j,2} = -1) = \frac{1}{2}$ 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 sequence of independent random variables $\xi_1, \dots, \xi_n$. Then $I_{n,2}(f)$ is a degenerate U-statistic. If $u \ge B_1 n\sigma^3$ with some appropriate constant $B_1 > 1$, $\bar{B}_2^{-1} n \ge u \ge \bar{B}_2 n^{-2}$ with a sufficiently large fixed number $\bar{B}_2 > 0$ and $\frac{1}{4} \ge \sigma^2 \ge \frac{1}{n}$, $n \ge 100$, then the estimate

$$P(n^{-1} I_{n,2}(f) > u) \ge \exp\left\{-Bn^{1/3}u^{2/3} \log\left(\frac{u}{n\sigma^3}\right)\right\} \tag{8.17}$$

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 if we disregard the value of the universal constants in our estimates, then in the case $u = n\sigma^3$ the right-hand side of (8.17) has the same order as Theorem 8.3 suggests. (In this example $\int f^2(x, y)\mu(\,dx)\mu(\,dy) = 2E(\eta_{j,1}\eta_{j,2})^2 = \sigma^2$, where μ is the distribution of $(\eta_{j,1}, \eta_{j,2})$.) If the probability in (8.17) at the same level u is considered for such a modified version of Example 8.8 where the same construction is taken, but with a much smaller parameter σ^2, then the probability at the right-hand side of (8.17) 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 (8.17) can be slightly improved, the term log can be replaced by $\log^{2/3}$ in the exponent of the right-hand side of (8.17). To get such an improvement some additional calculation is needed where the numbers u_1 and u_2 have to be replaced by $v_1 = 8n^{1/3}u^{2/3} \log^{-1/3}\left(\frac{u}{n\sigma^3}\right)$ and $v_2 = \frac{1}{4}n^{2/3}u^{1/3} \log^{1/3}\left(\frac{u}{n\sigma^3}\right)$.

It is simple to check that the U-statistic introduced 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 (8.17) the results of Section 3, in particular Example 3.2 can be applied for the sequence $\eta_{j,1}$, $j = 1, 2, \dots, n$. Beside this, the following result holds from the theory of large deviations: If $X_1, \dots, X_n$ are independent and identically distributed random variables, $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$, 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 $P\left(\sum_{j=1}^{n} X_j > u\right) \ge C_1 e^{-C_2 u^2/n}$ for all $0 \le u \le \alpha n$.

Proof of Example 8.8. The inequality

$$P(n^{-1}I_{n,2}(f) > u) \ge P\left(2\left(\sum_{j=1}^{n}\eta_{j,1}\right)\left(\sum_{j=1}^{n}\eta_{j,2}\right) > 2nu\right) - P\left(2\sum_{j=1}^{n}\eta_{j,1}\eta_{j,2} > nu\right) \tag{8.18}$$

holds. Because of the independence of the random variables $\eta_{j,1}$ and $\eta_{j,2}$ the first probability at the right-hand side of (8.18) can be bounded from below with the choice $v_1 = 4n^{1/3}u^{2/3}$ and $v_2 = \frac{1}{4}n^{2/3}u^{1/3}$ by means of Example 3.2. (The estimate of Example 3.2 can be applied with the choice $y = v_1$, since the relation $v_1 \ge 4n\sigma^2$ holds if $u \ge B_1 n\sigma^3$ with $B_1 > 1$, and the remaining conditions $0 \le \sigma^2 \le \frac{1}{8}$ and $n \ge 3v_1 \ge 6$ also hold under the conditions of Example 8.8.) This estimate together with the large-deviation result mentioned after the remark imply that

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

with appropriate constants $B_1 > 1$, $B_2 > 0$ 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 $X_j = 2\eta_{j,1}\eta_{j,2}$ at level nu instead of level u we get the following upper bound for the second term at the right-hand side of (8.18).

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

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. In this case also the constant B_4 can be chosen sufficiently large in the last inequality. The above estimates imply the statement of the example.

9. Some results about U-statistics.

This section contains the proof of the Hoeffding decomposition theorem, an important result about U-statistics. It states that all U-statistics can be represented as a sum of degenerate U-statistics of different order. This representation can be considered as the natural multivariate version of the decomposition of a random variable as the sum of a random variable with expectation zero plus a constant (which can be interpreted as a random variable of zero variable). Some important properties of the Hoeffding decomposition will also be proved. The properties of the kernel function of a U-statistic will be compared to those of the kernel functions of the U-statistics in its Hoeffding decomposition.

If the Hoeffding decomposition of a U-statistic is taken, then the L_2 and L_∞-norms of the kernel functions appearing in the U-statistics of the Hoeffding decomposition will be bounded by means of the corresponding norm of the kernel function of the original U-statistic. It will be also shown that if a class of U-statistics defined with the help of an L_2-dense class of kernel functions (and the same sequence of independent and identically distributed random variables) is considered, and the Hoeffding decomposition of all of these U-statistics is taken, then the kernel functions of the degenerate U-statistics appearing in these Hoeffding decompositions also constitute an L_2-dense class. Another important result of this section is Theorem 9.4. It gives a decomposition of a k-fold random integral with respect to a normalized empirical measure to the linear combination of degenerate U-statistics. This result makes possible to derive Theorem 8.1 from Theorem 8.3 and Theorem 8.2 from Theorem 8.4, and it is also useful in the proof of Theorems 8.3 and 8.4.

Let us first consider the Hoeffding's decomposition. In the special case $k = 1$ it states that the sum $S_n = \sum_{j=1}^{n} \xi_j$ of independent and identically distributed random variables can be rewritten as $S_n = \sum_{j=1}^{n} (\xi_j - E\xi_j) + \left(\sum_{j=1}^{n} E\xi_j \right)$, i.e. as the sum of independent random variables with zero expectation plus a constant. For the sake of a simpler terminology in the sequel a constant will be considered as a degenerate U-statistic of order zero, and the notation $I_{n,0}(c) = c$ will be applied for a constant c. I wrote down the above trivial formula, because Hoeffding's decomposition is actually its adaptation to a more general situation. To understand this let us first see how to adapt the above construction to the case $k = 2$.

In this case a sum of the form $I_{n,2}(f) = \sum_{1 \le j,k \le n, j \ne k} f(\xi_j, \xi_k)$ has to be considered. Write $f(\xi_j, \xi_k) = [f(\xi_j, \xi_k) - E(f(\xi_j, \xi_k)|\xi_k)] + E(f(\xi_j, \xi_k)|\xi_k) = f_1(\xi_j, \xi_k) + \bar{f}_1(\xi_k)$ with $f_1(\xi_j, \xi_k) = f(\xi_j, \xi_k) - E(f(\xi_j, \xi_k)|\xi_k)$, and $\bar{f}_1(\xi_k) = E(f(\xi_j, \xi_k)|\xi_k)$ to make the conditional expectation of $f_1(\xi_j, \xi_k)$ with respect to ξ_k equal zero. Repeating this procedure for the first coordinate we define $f_2(\xi_j, \xi_k) = f_1(\xi_j, \xi_k) - E(f_1(\xi_j, \xi_k)|\xi_j)$ and $\bar{f}_2(\xi_j) = E(f_1(\xi_j, \xi_k)|\xi_j)$. Let us also write $\bar{f}_1(\xi_k) = [\bar{f}_1(\xi_k) - E\bar{f}_1(\xi_k)] + E\bar{f}_1(\xi_k)$ and $\bar{f}_2(\xi_j) = [\bar{f}_2(\xi_j) - E\bar{f}_2(\xi_j)] + E\bar{f}_2(\xi_j)$. 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 $I_{n,2}(f)$ into a sum of degenerate U-statistics of different orders.

Hoeffding's decomposition can be obtained 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 the following notations will be introduced.

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 $Q_j = I - P_j$ i.e. $Q_j f = f - P_j f$ for all integrable functions on f on the space $(X^k, \mathcal{X}^k, \mu^k)$, $1 \le j \le k$. 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. The following result holds.

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 such a decomposition*

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

for which all functions f_V, $V \subset \{1, \dots, k\}$, in (9.2) are canonical with respect to the probability measure μ, and the function f_V depends on 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

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

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 $I_{n,k}(f)$.

Proof of Theorem 9.1. Write $f = \prod_{j=1}^{k} (P_j + Q_j) f$. 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 $P_j f_V \equiv 0$ (in all coordinates x_s, $s \in V \setminus \{j\}$ if $j \in V$). Since $P_j = P_j^2$, and the identity $P_j Q_j = P_j - P_j^2 = 0$ 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 $P_j f_V = \left(\prod_{s \in \{1,\dots,k\} \setminus V} P_s \prod_{s \in V \setminus \{j\}} Q_s \right) P_j Q_j f = 0$. By applying 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).

In the Hoeffding decomposition we rewrote a general U-statistic in the form of a linear combination of degenerate U-statistics. In many applications of this result we still we have to know how the properties of the kernel function f of the original U-statistic are reflected in the properties of the kernel functions f_V of the degenerate U-statistics taking part in the Hoeffding composition. In particular, we need a good estimate on the L_2 and L_∞ norm of the functions f_V by means of the corresponding norm of the function f. Moreover, if we want to prove estimates on the tail distribution of the supremum of U-statistics $I_{n,k}(f)$ for a nice class of kernel functions $f \in \mathcal{F}$ which is an L_2-dense class of functions with some exponent L and parameter D, then we may need a similar estimate on the class of kernel functions f_V, $f \in \mathcal{F}$, with some $V \in \{1, \dots, k\}$ appearing in the Hoeffding decomposition of these functions. We have to show that this class of functions is also L_2-dense, and we also need a good bound on the exponent and parameter of this L_2-dense class. The next result formulates such a statement.

Theorem 9.2. (Some properties of the Hoeffding decomposition). *Let us consider a square integrable function $f(x_1, \dots, x_k)$ 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 $V \subset \{1, \dots, k\}$. (In particular, $f_\emptyset^2 \le \int f^2(x_1, \dots, x_k) \mu(\, dx_1) \dots \mu(\, dx_k)$ for $V = \emptyset$.)

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 functions f_V taking part in this decomposition. These 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 will be proved as a consequence of Proposition 9.3 presented below. To formulate it first some notations will be introduced:

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

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

defined for those $y \in Y$ for which the above 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 $Q = Q_\mu = I - P = I - P_\mu$

$$Q_\mu f(y,z) = (I - P_\mu)f(y,z) = f(y,z) - P_\mu f(y,z) = f(y,z) - \int f(y,z)\mu(\,dz), \quad (9.6)$$

defined for those points $(y,z) \in Y \times Z$ whose first coordinate y is such that the expression $P_\mu f(y)$ is meaningful. (Here, and in the sequel a function $g(y)$ defined on the space $(Y, \mathcal{Y})$ will be sometimes identified 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 holds:

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.*

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

and

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

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 $\mathcal{G}_\mu = \{\frac{1}{2}Q_\mu f = \frac{1}{2}(f - P_\mu f),\ f \in \mathcal{F}\}$ 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 $(Y_1 \times Z \times Y_2, \mathcal{Y}_1 \times \mathcal{Z} \times \mathcal{Y}_2)$ of three measurable spaces $(Y_1, \mathcal{Y}_1)$, $(Z, \mathcal{Z})$ and $(Y_2, \mathcal{Y}_2)$ with a probability measure μ on the space $(Z, \mathcal{Z})$ and a probability measure ρ on $Y_1 \times Y_2, \mathcal{Y}_1 \times \mathcal{Y}_2)$, and define the transformations*

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

and

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

for the measurable functions f on the space $Y_1 \times Z \times Y_2$ integrable with respect the measure $\mu \times \rho$. 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 $\mathcal{G}_\mu = \{\frac{1}{2} Q_\mu f = \frac{1}{2}(f - P_\mu f),\ f \in \mathcal{F}\}$ 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 $(Y_1 \times Z \times Y_2, \mathcal{Y}_1 \times \mathcal{Z} \times \mathcal{Y}_2)$. Beside this, we apply that measure on $(Y_1 \times Z \times Y_2, \mathcal{Y}_1 \times \mathcal{Z} \times \mathcal{Y}_2)$ which is the image of the product measure $\rho \times \mu$ with respect to the map induced by the above transformation on the space of measures.

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 $\mathcal{F}_{P_s} = \{P_s f\colon\ f \in \mathcal{F}\}$ and $\mathcal{F}_{Q_s} = \{\frac{1}{2} Q_s f\colon\ f \in \mathcal{F}\}$, $1 \le s \le k$. These relations together with the identity $f_V = \left(\prod_{s \in V} P_s \prod_{s \in \{1, \dots, k\} \setminus V} Q_s \right) f$ 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). This 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 if the same relation holds for $\mathcal{F}$ we have to find for all $0 < \varepsilon \le 1$ a set $\{f_1, \dots, f_m\} \subset \mathcal{F}_\mu$, $1 \le j \le m$ with $m \le D\varepsilon^{-L}$ elements, 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 for $\mathcal{F}$ 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 under the same condition we have to find for all numbers $0 < \varepsilon \le 1$ 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 $g \in \mathcal{G}_\mu$.

To show this let us consider the probability measure $\tilde{\rho} = \frac{1}{2}(\rho + \bar{\rho} \times \mu)$ 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 $\mathcal{F}_0(\varepsilon, \tilde{\rho}) = \{f_1, \dots, f_m\} \subset \mathcal{F}$ with $m \le D\varepsilon^{-L}$ elements 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 $\{g_1, \dots, g_m\} = \{\frac{1}{2}Q_\mu f_1, \dots, \frac{1}{2}Q_\mu f_m\}$. All functions $g \in \mathcal{G}_\mu$ can be written in the form $g = \frac{1}{2}Q_\mu f$ 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 $\int \frac{1}{4}(Q_\mu f - Q_\mu \bar{f})^2 \, d\rho \le \int (f - \bar{f})^2 \, d\tilde{\rho}$ for all pairs $f, \bar{f} \in \mathcal{F}$. This inequality holds, since $\int \frac{1}{4}(Q_\mu f - Q_\mu \bar{f})^2 \, d\rho \le \int \frac{1}{2}(f - \bar{f})^2 \, d\rho + \int \frac{1}{2}(P_\mu f - P_\mu \bar{f})^2 \, d\rho$, 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 $\int \frac{1}{4}(Q_\mu f - Q\mu \bar{f})^2 \, d\rho \le \int (f - \bar{f})^2 \frac{1}{2} d\,(\rho + \bar{\rho} \times \mu) = \int (f - \bar{f})^2 d\,\tilde{\rho}$ as we have claimed.

Now we shall discuss the relation between Theorem 8.1′ and Theorem 8.3 and between Theorem 8.2 and Theorem 8.4. First we shall show that Theorem 8.1 (or Theorem 8.1′) is equivalent to the corollary of Theorem 8.3 which contains the estimate (8.10′) instead of the slightly stronger estimate (8.10) formulated in Theorem 8.3. We also claim that Theorems 8.2 and 8.4 are equivalent. Both in Theorem 8.2 and in Theorem 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 integration with respect to the measure $\mu_n - \mu$ in the definition (4.8) of the integral $J_{n,k}(f)$ yields some kind of normalization which is missing 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 the corollary of Theorem 8.3 from Theorems 8.1′ or of Theorem 8.4 from Theorem 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 identity we need the identity $\int f(x_1, \dots, x_k)\mu(\,dx_j) \equiv 0$ in the case when the domain of integration is not the whole space X but the set $X \setminus \{x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k\}$.

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 in-

troduce 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 $\bar{f}((x_1, u_1), \dots, (x_k, u_k)) = f(x_1, \dots, x_k)$ in this enlarged space. Then $I_{n,k}(f) = I_{n,k}(\bar{f})$, the measure $\bar{\mu} = \mu \times \lambda$ is non-atomic, and $\bar{f}$ is canonical with respect to $\bar{\mu}$ if f is canonical with respect toμ. Hence the corollary of Theorem 8.3 and Theorem 8.4 can be derived from Theorems 8.1′ and 8.2 respectively by proving them first for their counterpart in the above constructed enlarged space with the above defined functions.

Also Theorems 8.1′ and 8.2 can be derived from Theorems 8.3 and 8.4 respectively, but this demands some additional work. For this goal let us observe that a random integral $J_{n,k}(f)$ can be written as a sum of U-statistics of different order, and it can also be expressed as a sum of degenerate U-statistics if Hoeffding's decomposition is applied for each U-statistic in this sum. We show that the coefficients of the degenerate U-statistics in the above representation have relatively small coefficients. This result is formulated in the following Theorem 9.4. To make its content more understandable I describe it in the special case of two-fold random integrals in a more explicit form in Corollary 2 of Theorem 9.4.

Theorem 9.4. (Decomposition of a multiple random integral with respect to a normalized empirical measure to a linear combination of degenerate U-statistics). *Let a non-atomic measure μ be given on a measurable space $(X, \mathcal{X})$ together with a sequence of independent, μ-distributed random variables $\xi_1, \dots, \xi_n$. Take a function $f(x_1, \dots, x_k)$ of k variables on the space $(X^k, \mathcal{X}^k)$ such that*

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

and 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 $\lim_{n\to\infty} C(n,k,V) = C(k,V)$ hold with some appropriate constant such that $0 \le |C(k,V)| < \infty$, and $C(n,k,\{1,\dots,k\}) = 1$ for $V = \{1, \dots, k\}$.

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. $C(n,k,V) = C(n,k,|V|)$ can be written. This fact will be not applied in this work.

Theorems 8.1′ and 8.2 can be simply derived from Theorems 8.3 and 8.4 respectively with the help of Theorem 9.4. Indeed, to get Theorem 8.1′ observe that formula (9.9) implies the inequality

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

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). Hence Theorem 8.1′ follows 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 is 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. It is enough to estimate each term at the right-hand side of (9.10) by means of Theorem 8.3. It can be assumed that $2^k C(k) > 1$. Let us first assume that also the inequality $\frac{u}{2^k C(k)\sigma} \ge 1$ holds. In this case formula (8.3′) in Theorem 8.1′ can be obtained by means of the estimation of each term at the right-hand side of (9.10). Observe that $\exp\left\{ -\alpha \left(\frac{u}{2^k C(k)\sigma} \right)^{2/s} \right\} \le \exp\left\{ -\alpha \left(\frac{u}{2^k C(k)\sigma} \right)^{2/k} \right\}$ for all $s \le k$ if $\frac{u}{2^k C(k)\sigma} \ge 1$. In the other case, if $\frac{u}{2^k C(k)\sigma} \le 1$, formula (8.3′) holds again with a sufficiently large $C > 0$, because in this case its right-hand side is greater than 1.

Theorem 8.2 can be similarly derived from Theorem 8.4 by observing that relation (9.10) remains valid if $|J_{n,k}(f)|$ is replaced by $\sup_{f \in \mathcal{F}} |J_{n,k}(f)|$ and $|I_{n,|V|}(f_V)|$ by $\sup_{f_V \in \mathcal{F}_V} |I_{n,|V|}(f_V)|$ in it, and we have the right to choose the constant M in formula (8.6) of Theorem 8.2 sufficiently large. The only difference in the argument is that beside formulas (9.4) and (9.4′) the last statement of Theorem 9.2 also has to be applied in this case. It 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.

Next I make some comments about the content of Theorem 9.4. The expression $J_{n,k}(f)$ was defined as a k-fold random integral, where the diagonals were omitted from the domain of integration. We have integrated with respect to the signed measure $\mu_n - \mu$, which means some kind of normalization. Thus it is not surprising that $J_{n,k}(f)$ has a tail distribution behaviour similar to that of degenerate U-statistics. Theorem 9.4 has such a consequence. Formula (9.9) expresses the random integral $J_{n,k}(f)$ as a linear combination of degenerate U-statistics of different order. This is similar to the Hoeffding decomposition of the U-statistic $I_{n,k}(f)$ where the same functions f_V appear. But the coefficients $C(n,k,|V|)n^{-|V|/2}$ of the terms $I_{n,|V|}(f_V)$ in the expansion (9.9) are small. On the other hand, they do not have to disappear. In particular, the expansion (9.9) may contain a non-zero constant term, in which case the expected value $EJ_{n,k}(f)$ is not equal to zero. But even in this case it can be bounded by a number not depending on the sample size n. Next I show an example for such a random integral $J_{n,2}(f)$ where $EJ_{n,2}(f) \ne 0$.

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.)

Now I turn to the proof of Theorem 9.4.

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

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

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 $P = \{R_1, \dots, R_u\}$ $\bigcup_{j=1}^{u} R_j = \{1, \dots, k\}$, $R_j \cap R_l = \emptyset$, $1 \le j < l \le u$, 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 the set $\{1, \dots, k\}$ in which two points s and t, $1 \le s, t \le k$, belong to the same component of this partition if and only if $l_s = l_t$. For a partition $P \in \mathcal{P}_k$ let us define the set $\mathcal{H}(P) = \mathcal{H}_n(P)$ consisting of sequences $(l_1, \dots, l_k)$ with $1 \le l_j \le n$ for all $1 \le j \le k$ 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

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

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 if $l_s = l_t$ for some $s \ne t$, then 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 because of this fact. 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. Beside this, it will be seen from the calculations that the same degenerate U-statistics $I_{n,|V|}(f_V)$

appear in the representation of $J_{n,k}(f)$ as in formula (9.2). This is 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 a 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 $s \geq 2$. 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

$$\sum_{j=1}^{s} (-1)^{s-1} \mu(\, dx_1) \dots \mu(\, dx_{j-1}) \mu^{(l)}(\, dx_j) \mu(\, dx_{j+1}) \dots \mu(\, dx_s) + (-1)^s \mu(dx_1) \dots \mu(dx_s)$$

$$= \sum_{j=1}^{s} (-1)^{s-1} \mu(\, dx_1) \dots \mu(\, dx_{j-1}) (\mu^{(l)}(\, dx_j) - \mu(\, dx_l)) \mu(\, dx_{j+1}) \dots \mu(\, dx_s)$$

$$+ (-1)^{s-1} (s-1) \mu(dx_1) \dots \mu(dx_s). \tag{9.13}$$

Here we have exploited 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 $1 \leq l \leq n$. We have assumed that $s \geq 2$, 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 calculation leading to relation (9.13), only the notations will be more complicated. To write down what we get first we define a set $\mathcal{T}(P)$ whose elements are certain subsets of $\{1, \dots, k\}$ depending on the partition $P = \{R_1, \dots, R_u\}$ together with a subset $\bar{\mathcal{T}}(P)$ of it. The elements of the set $\mathcal{T}(P)$ are all those sets $\{j_1, \dots, j_{u'}\} \subset \{1, \dots, k\}$, $u' \leq u$, for which the numbers $j_1, \dots, j_{u'}$ belong to different components 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 satisfy the following additional condition: If some components $R_t = \{b_t\}$, $1 \leq t \leq u$, of the partition P consists of only one point, then all 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), that

$$\int{}' f(x_1, \dots, x_k) \Big(\mu^{(l_1)}(\, dx_1) - \mu(\, dx_1)\Big) \dots \Big(\mu^{(l_k)}(\, dx_k) - \mu(\, dx_k)\Big) \tag{9.14}$$

$$= \sum_{V \in \bar{\mathcal{T}}(P)} \alpha(V, P) \int f(x_1, \dots, x_k) \prod_{j \in V} \Big(\mu^{(l_j)}(\, dx_j) - \mu(\, dx_j)\Big) \prod_{j' \in \{1, \dots, k\} \setminus V} \mu(\, dx_{j'})$$

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 $V \in \bar{\mathcal{T}}(P)$. On the other hand, it is important to observe that a term with non-zero coefficient $\alpha(V,P)$ appears at the right-hand side of (9.14) only for $V \in \bar{\mathcal{T}}(P)$. The class of functions $\bar{\mathcal{T}}(P)$ was introduced, because they have this property. This relation 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), and a component of the form $\mu(\,dx_{b_t})$ is missing. Hence the product $\prod_{j' \in \{1,\dots,k\} \setminus V} \mu(\,dx_{j'})$ cannot appear at the right-hand side of (9.14) if $V \notin \bar{\mathcal{T}}(P)$.

Let me remark that at the right-hand side of (9.14) $\int$ was written and not $\int'$, i.e. the diagonal was not omitted 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

$$\begin{aligned} &\int' f(x_1,\dots,x_k) \Big(\mu^{(l_1)}(\,dx_1) - \mu(\,dx_1)\Big) \dots \Big(\mu^{(l_k)}(\,dx_k) - \mu(\,dx_k)\Big) \\ &\qquad = \sum_{V \in \bar{\mathcal{T}}(P)} \alpha(V,P) \left(\prod_{j' \in \{1,\dots,k\} \setminus V} P_{\mu,j'} \prod_{j \in V} Q_{\mu,j} \right) f(\xi_{l_j},\, j \in V) \end{aligned} \tag{9.15}$$

if $(l_1,\dots,l_k) \in \mathcal{H}(P)$. 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 product 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 $\big(\mu^{(l_j)}(\,dx_j) - \mu(\,dx_j)\big)$ 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 $P_{\mu,j'}$. Beside this, the operators $Q_{\mu,j}$ and $P_{\mu,j'}$ are exchangeable.

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 $I_{n,|V|}(f_v)$ is the U-statistic of order $|V|$ with the kernel function

$$f_V(x_j, j \in V) = \left(\prod_{j' \in \{1,\dots,k\} \setminus V} P_{\mu,j'} \prod_{j \in V} Q_{\mu,j} \right) f \tag{9.17}$$

with that function on $f(x_1, \dots, x_k)$ which is considered in Theorem 9.4 and some appropriate coefficients $\bar{\alpha}(V, P, k, n)$ at the right-hand side of (9.16). These coefficients could be explicitly calculated. We do not need an explicit formula for $\bar{\alpha}(V, P, k, n)$, but we shall need 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 $R_j \cap V = \emptyset$, (here u denotes the cardinality of the partition P), and the constant $D(k) < \infty$ depends only on the multiplicity k of the integral $J_{n,k}(f)$.

To show that $|\bar{\alpha}(V, P, k, n)| \le D(k) n^{\beta(P,V)}$ let us observe that if we fix the coordinates l_j, $j \in V$, in an arbitrary way and sum up the expression at left-hand side of (9.16) for the remaining indices $l_{j'}$, $j' \notin V$, 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 $\bar{\alpha}(V, P, k, n)$. Hence to get a good estimate on $\bar{\alpha}(V, P, k, n)$ the number of the vectors $(l_j, j \notin V)$ taking part in the summation at the left-hand side of (9.16) has to be well bounded. 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. Beside this, 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

$$C(k, V, P) = \lim_{n \to \infty} n^{-\beta(P,V)} \bar{\alpha}(V, P, k, n), \qquad |C(k, V, P)| < \infty, \tag{9.18}$$

also exists.

We get by applying relations (9.12) and (9.15) 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

$$J_{n,k}(f) = \sum_{V \subset \{1,2,\dots,k\}} n^{-|V|/2} C(n, k, V) \frac{1}{k!} \sum_{\substack{1 \le l_j \le n, \\ l_j \neq l_{j'} \text{ if } j \neq j' \text{ for } j \in V}} f_V(\xi_{l_j},\, j \in V) \tag{9.19}$$

holds with the functions $f_V(x_j,\, j \in V)$ defined in (9.17) for all $V \subset \{1, \dots, k\}$ and some appropriate coefficients $C(n, k, V)$. We shall show that these coefficients satisfy the inequality $|C(n, k, V)| \le C(k)$ with some constant $C(k) > 0$. Beside this, it is not difficult to see that the identity $C(n, k, \{1, \dots, k\}) = 1$ holds. To see the estimate $|C(n, k, V)| \le C(k)$ observe that $n^{-|V|/2} C(n, k, |V|)$ can be written as a sum of finitely many terms, (their number is less than a number depending only on k) such that all of them can be bounded by a number of the form $\frac{|\bar{\alpha}(V,P,k,n)|}{n^{k/2} k!} \le D(k) n^{-k/2 + \beta(P,V)}$ 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 such a partition of a set with less than or equal to $k - |V|$ elements whose components have at least 2 elements.

Relation (9.19) can be rewritten as $J_{n,k}(f) = \sum_{V \subset \{1,2,\dots,k\}} C(n,k,V) n^{-|V|/2} I_{n,|V|}(f_V)$, 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 (9.17). This kernel function 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 (9.18) together with the bound on the terms $\beta(P,V)$ also imply that the finite limits $\lim_{n\to\infty} C(n,k,V) = C(k,V)$ also exist. Theorem 9.4 is proved.

Two corollaries of Theorem 9.4 will be formulated. 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*

$$E\left(n^{-k/2} I_{n,k}(f)\right)^2 = \frac{n(n-1)\cdots(n-k+1)}{k! n^k} \int \operatorname{Sym} f^2(x_1,\dots,x_k)\mu(dx_1)\dots\mu(dx_k)$$
$$\le \frac{1}{k!} \int f^2(x_1,\dots,x_k)\mu(dx_1)\dots\mu(dx_k), \tag{9.20}$$

where μ is the distribution of the random variables taking part in the definition of the U-statistic $I_{n,k}(f)$, and $\operatorname{Sym} f$ *is the symmetrization of the function f. The k-fold multiple random integral $J_{k,n}(f)$ with an arbitrary square integrable kernel function f satisfies the inequality*

$$EJ_{n,k}(f)^2 \le \bar{C}(k) \int f^2(x_1,\dots,x_k)\mu(dx_1)\dots\mu(dx_k)$$

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. The identity

$$E(n^{-k/2} I_{n,k}(f))^2 = \frac{1}{(k!)^2 n^k} {\sum}' Ef(\xi_{l_1},\dots,\xi_{l_k}) f(\xi_{l'_1},\dots,\xi_{l'_k}) \tag{9.21}$$

holds, 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 $Ef(\xi_{l_1},\dots,\xi_{l_k}) f(\xi_{l'_1},\dots,\xi_{l'_k}) = 0$ if the two sets $\{l_1,\dots,l_k\}$ and $\{l'_1,\dots,l'_k\}$ 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. On the other hand,

$$Ef(\xi_{l_1},\dots,\xi_{l_k}) f(\xi_{l'_1},\dots,\xi_{l'_k}) = \int f(x_1,\dots,x_k) f(x_{\pi(1)},\dots,x_{\pi(k)})\mu(dx_1)\dots\mu(dx_k)$$

if $(l'_1,\dots,l'_k) = (\pi(l_1),\dots,\pi(l_k))$ with some $(\pi(1),\dots,\pi(k)) \in \Pi_k$, where Π_k denotes the set of all permutations of the set $\{1,\dots,k\}$. By summing up the above identities for

all pairs $(l_1, \dots, l_k)$ and $(l'_1, \dots, l'_k)$ and by applying formula (9.21) we get the identity at the left-hand side of formula (9.20). The second relation in (9.20) is obvious.

The bound for $J_{n,k}(f)$ follows from Theorem 9.4, formula (9.4) in Theorem 9.2 by which the L_2-norm of the functions f_V is not greater than the L_2-norm of the function f and the bound that formula (9.20) yields for the second moment of the 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. Multiple Wiener–Itô integrals and their properties.

In this section I present the definition of multiple Wiener–Itô integrals and some of its most important properties needed in the proof of the results in Section 8. First the notion of the white noise with some reference measure will be introduced, then multiple Wiener–Itô integrals with respect to a white noise with some non-atomic reference measure will be defined. A most important result in the theory of multiple Wiener–Itô integrals is the so-called diagram formula presented in Theorem 10.2A which enables us to write the product of two Wiener–Itô integrals in the form of a sum of Wiener–Itô integrals. Its proof is given in Appendix B.

Another interesting result about Wiener-Itô integrals, formulated at the end of this section in Theorem 10.5 states that the class of random variables which can be written in the form of a sum of Wiener–Itô integrals of different order is sufficiently rich. All random variables with finite second moment which are measurable with respect to the σ-algebra generated by the (Gaussian) random variables appearing in the underlying white noise in the construction of multiple Wiener–Itô integrals can be written in such a form. This result explains why it is natural to expect a result like the diagram formula. The product of two Wiener–Itô integrals is also measurable with respect to the σ-algebra generated by the random variables in the underlying white noise, hence if we know that such a product is square integrable, then Theorem 10.5 implies that it can be written as the sum of multiple Wiener–Itô integrals of different order. What makes the diagram formula especially useful is the fact that it yields an explicit representation of a product of Wiener–Itô integrals in the form of a sum of Wiener–Itô integrals. I shall also give a heuristic explanation of the diagram formula which may explain why it has the form appearing in Theorem 10.2A. It also helps to find the analogs of the diagram formula for (random) integrals with respect to the product of normalized empirical measures. Such a result will be useful later.

Once the diagram formula is proved, it is not difficult to generalize it to the product of finitely many Wiener–Itô integrals. This generalization, formulated in Theorem 10.2, will be also called the diagram formula. It has an important corollary about the calculation of the moments of Wiener–Itô integrals. Theorem 8.5 can be proved relatively simply by means of this corollary.

I shall give the proof of two other results about Wiener–Itô integrals in Appendix C. The first one, Theorem 10.3, is called Itô's formula for Wiener–Itô integrals, and it explains the relation between multiple Wiener-Itô integrals and Hermite polynomials of Gaussian random variables. This result is a relatively simple consequence of the diagram formula and some basic recursive relations about Hermite polynomials.

The other result proved in Appendix C, Theorem 10.4, is a limit theorem about a sequences of appropriately normalized degenerate U-statistics, where the limit is a multiple Wiener–Itô integral. This result is interesting for us, because it helps to compare Theorems 8.3 and 8.1 with their one-variate counterpart, Bernstein's inequality. In the one-variate case Bernstein's inequality provides a comparison of the distribution of sums of independent random variables and normal distribution functions, i.e. the limit distribution in the central limit theorem. Theorem 8.3 yields a similar result about de-

generate U-statistics. Its comparison with Theorem 8.5 and the limit theorem proved in Appendix C about the limit distribution of degenerate U-statistics show that degenerate U-statistics satisfy an estimate similar to Bernstein's inequality. The upper bound in it is similar to the estimate on the tail-distribution of the limit distribution of normalized degenerate U-statistics, which equals to the distribution of an appropriate multiple Wiener–Itô integral. The estimate of Theorem 8.1 about multiple integrals with respect to normalized empirical distribution functions also have a similar interpretation.

My Lecture Note [28] contains a rather detailed description of Wiener–Itô integrals. But in that work the emphasis was put on the study of a slightly different version of it. The original version introduced here was also only briefly discussed there, not all details were worked out. In particular, the diagram formula needed in this work was formulated and proved only for modified Wiener–Itô integrals. I shall discuss the difference between these random integrals together with the question why a different version of Wiener–Itô integrals was introduced in [28] at the end of the section.

To define multiple Wiener–Itô integrals first the notion of a white noise has to be introduced. This is done in the following definition.

Definition of a white noise with some reference measure. *Let us have a σ-finite measure μ on a measurable space $(X, \mathcal{X})$. A white noise with reference measure μ is a Gaussian random field $\mu_W = \{\mu_W(A)\colon\ A \in \mathcal{X}, \mu(A) < \infty\}$, i.e. a set of jointly Gaussian random variables indexed by the above sets A, which satisfies the relations $E\mu_W(A) = 0$ and $E\mu_W(A)\mu_W(B) = \mu(A \cap B)$.*

It is worth making some comments about this definition.

Remark: In the definition of a white noise sometimes also the property $\mu_W(A \cup B) = \mu_W(A) + \mu_W(B)$ with probability 1 if $A \cap B = \emptyset$, and $\mu(A) < \infty$, $\mu(B) < \infty$ is mentioned. But this condition can be omitted, because it follows from the remaining properties of the white noise. Indeed, simple calculation shows that $E(\mu_W(A \cup B) - \mu_W(A) - \mu_W(B))^2 = 0$ if $A \cap B = \emptyset$, hence $\mu_W(A \cup B) - \mu_W(A) - \mu_W(B) = 0$ with probability 1 in this case. It also can be observed that if some sets $A_1, \dots, A_k \in \mathcal{X}$, $\mu(A_j) < \infty$, $1 \le j \le k$, are disjoint, then the random variables $\mu_W(A_j)$, $1 \le j \le k$, are independent because of the uncorrelatedness of these jointly Gaussian random variables.

It is not difficult to see that for an arbitrary reference measure μ on a space $(X, \mathcal{X})$ a white noise μ_W with this reference measure really exists. This follows simply from Kolmogorov's fundamental theorem, by which if the finite dimensional distributions of a random field are prescribed in a consistent way, then there exists a random field with these finite dimensional distributions.

Now I turn to the definition of multiple Wiener–Itô integrals with respect to a white noise with some reference measure. First I introduce the class of functions whose Wiener–Itô integrals with respect to a white noise μ_W with a non-atomic reference measure μ will be defined.

Let us consider a measurable space $(X, \mathcal{X})$, a non-atomic σ-finite measure μ on it and a white noise μ_W on $(X, \mathcal{X})$ with reference measure μ. Let us define the classes

of functions $\mathcal{H}_{\mu,k}$, $k = 1, 2, \dots$, consisting of functions of k variables on $(X, \mathcal{X})$ by the formula

$$\mathcal{H}_{\mu,k} = \Big\{ f(x_1, \dots, x_k) \colon\ f(x_1, \dots, x_k) \text{ is an } \mathcal{X}^k \text{ measurable, real valued} \\ \text{function on } X^k, \text{ and } \int f^2(x_1, \dots, x_k) \mu(\, dx_1) \dots, \mu(\, dx_k) < \infty \Big\}. \tag{10.1}$$

The k-fold Wiener-Itô integrals of the functions $f \in \mathcal{H}_{\mu,k}$ with respect to the white noise μ_W will be defined in a rather standard way. First they will be defined for some simple functions, called elementary functions, then it will be shown that the integral for this elementary functions have an L_2 contraction property which makes possible to extend it to the class of functions in $\mathcal{H}_{\mu,k}$.

Let us first introduce the following class of elementary functions $\bar{\mathcal{H}}_{\mu,k}$ of k variables. A function $f(x_1, \dots, x_k)$ on $(X^k, \mathcal{X}^k)$ belongs to $\bar{\mathcal{H}}_{\mu,k}$ if there exist finitely many disjoint measurable subsets $A_1, \dots, A_M$, $1 \le M < \infty$, of the set X (i.e. $A_j \cap A_{j'} = \emptyset$ if $j \neq j'$) such that $\mu(A_j) < \infty$ for all $1 \le j \le M$, and the function f has the form

$$f(x_1, \dots, x_k) = \begin{cases} c(j_1, \dots, j_k) \quad \text{if } (x_1, \dots, x_k) \in A_{j_1} \times \dots \times A_{j_k} \\ \qquad \text{with some indices } (j_1, \dots, j_k), \quad 1 \le j_s \le M,\ 1 \le s \le k, \\ \qquad \text{such that all numbers } j_1, \dots, j_k \text{ are different} \\ 0 \quad \text{if } (x_1, \dots, x_k) \notin \bigcup_{\substack{(j_1,\dots,j_k)\colon 1 \le j_s \le M,\ 1 \le s \le k, \\ \text{and all } j_1,\dots,j_k \text{ are different.}}} A_{j_1} \times \dots \times A_{j_k} \end{cases} \tag{10.2}$$

with some real numbers $c(j_1, \dots, j_k)$, $1 \le j_s \le M$, $1 \le s \le k$, if all $j_1, \dots, j_k$ are different numbers. This means that the function f is constant on all k-dimensional rectangles $A_{j_1} \times \dots \times A_{j_k}$ with different, non-intersecting edges, and it equals zero on the complementary set of the union of these rectangles. The property that the support of the function f is on the union of rectangles with non-intersecting edges is sometimes interpreted so that the diagonals are omitted from the domain of integration of Wiener–Itô integrals.

The Wiener-Itô integral of an elementary function $f(x_1, \dots, x_k)$ of the form (10.2) with respect to a white noise μ_W with the (non-atomic) reference measure μ is defined by the formula

$$\int f(x_1, \dots, x_k) \mu_W(\, dx_1) \dots \mu_W(\, dx_k) \\ = \sum_{\substack{1 \le j_s \le M,\ 1 \le s \le k \\ \text{all } j_1,\dots,j_k \text{ are different}}} c(j_1, \dots, j_k) \mu_W(A_{j_1}) \cdots \mu_W(A_{j_k}). \tag{10.3}$$

(The representation of the function f in (10.2) is not unique, the sets A_j can be divided to smaller disjoint sets, but its Wiener–Itô integral defined in (10.3) does not depend

on its representation.) The notation

$$Z_{\mu,k}(f) = \frac{1}{k!} \int f(x_1, \dots, x_k) \mu_W(\, dx_1) \dots \mu_W(\, dx_k), \tag{10.4}$$

will be used in the sequel, and the expression $Z_{\mu,k}(f)$ will be called the normalized Wiener–Itô integral of the function f. Such a terminology will be applied also for the Wiener–Itô integrals of all functions $f \in \mathcal{H}_{\mu,k}$ to be defined later.

If f is an elementary function in $\bar{\mathcal{H}}_{\mu,k}$ defined in (10.2), then its normalized Wiener–Itô integral defined in (10.3) and (10.4) satisfies the relations

$$\begin{aligned} E k! Z_{\mu,k}(f) &= 0, \\ E(k! Z_{\mu,k}(f))^2 &= \sum_{\substack{(j_1,\dots,j_k)\colon 1 \le j_s \le M,\ 1 \le s \le k, \\ \text{and all } j_1,\dots,j_k \text{ are different.}}} \sum_{\pi \in \Pi_k} c(j_1, \dots, j_k) c(j_{\pi(1)}, \dots, j_{\pi(k)}) \\ &\qquad E\mu_W(A_{j_1}) \cdots \mu_W(A_{j_k}) \mu_W(A_{j_{\pi(1)}}) \cdots \mu_W(A_{j_{\pi(k)}}) \qquad (10.5) \\ &= k! \int \operatorname{Sym} f^2(x_1, \dots, x_k) \mu(\, dx_1) \dots \mu(\, dx_k) \\ &\le k! \int f^2(x_1, \dots, x_k) \mu(\, dx_1) \dots \mu(\, dx_k), \end{aligned}$$

where Π_k denotes the set of all permutations $\pi = \{\pi(1), \dots, \pi(k)\}$ of the set $\{1, \dots, k\}$, and $\operatorname{Sym} f(x_1, \dots, x_k) = \frac{1}{k!} \sum_{\pi \in \Pi_k} f(x_{\pi(1)}, \dots, x_{\pi(k)})$.

The identities written down in (10.5) can be simply checked. The first relation follows from the identity $E\mu_W(A_{j_1}) \cdots \mu_W(A_{j_k}) = 0$ for disjoint sets $A_{j_1}, \dots, A_{j_k}$, which holds, since the expectation of the product of independent random variables with zero expectation is taken. The second identity follows similarly from the identity

$$\begin{aligned} &E\mu_W(A_{j_1}) \cdots \mu_W(A_{j_k}) \mu_W(A_{j'_1}) \cdots \mu_W(A_{j'_k}) = 0 \\ &\qquad \text{if the sets of indices } \{j_1, \dots, j_k\} \text{ and } \{j'_1, \dots, j'_k\} \text{ are different,} \\ &E\mu_W(A_{j_1}) \cdots \mu_W(A_{j_k}) \mu_W(A_{j'_1}) \cdots \mu_W(A_{j'_k}) = \mu(A_{j_1}) \cdots \mu(A_{j_k}) \\ &\qquad \text{if } \{j_1, \dots, j_k\} = \{j'_1, \dots, j'_k\} \text{ i.e. if } j'_1 = j_{\pi(1)}, \dots, j'_k = j_{\pi(k)} \\ &\qquad \text{with some permutatation } \pi \in \Pi_k, \end{aligned}$$

which holds because of the facts that the μ_W measure of disjoint sets are independent with expectation zero, and $E\mu_W(A)^2 = \mu(A)$. The remaining relations in (10.5) can be simply checked.

It is not difficult to check that

$$E Z_{\mu,k}(f) Z_{\mu,k'}(g) = 0 \tag{10.6}$$

for all functions $f \in \bar{\mathcal{H}}_{\mu,k}$ and $g \in \bar{\mathcal{H}}_{\mu,k'}$ if $k \neq k'$, and

$$Z_{\mu,k}(f) = Z_{\mu,k}(\operatorname{Sym} f) \tag{10.7}$$

for all functions $f \in \bar{\mathcal{H}}_{\mu,k}$.

The definition of Wiener–Itô integrals can be extended to general functions $f \in \mathcal{H}_{\mu,k}$ with the help of the estimate (10.5). But to carry out this extension we still have to know that the class of functions $\bar{\mathcal{H}}_{\mu,k}$ is a dense subset of the class $\mathcal{H}_{\mu,k}$ in the Hilbert space $L_2(X^k, \mathcal{X}^k, \mu^k)$, where μ^k is the k-th power of the reference measure μ of the white noise μ_W. I briefly explain how this property of $\bar{\mathcal{H}}_{\mu,k}$ can be proved. The non-atomic property of the measure μ is exploited at this point.

To prove this statement it is enough to show that the indicator function of any product set $A_1 \times \cdots \times A_k$ such that $\mu(A_j) < \infty$, $1 \le j \le k$, but the sets $A_1, \dots, A_k$ may be non-disjoint is in the $L_2(\mu^k)$ closure of $\bar{\mathcal{H}}_{\mu,k}$. In the proof of this statement it will be exploited that if μ is a non-atomic measure, then for all $\varepsilon > 0$ and $1 \le j \le k$ the set A_j can be represented as a finite union $A_j = \bigcup_s B_{j,s}$ of disjoint sets $B_{j,s}$ such that $\mu(B_{j,s}) < \varepsilon$.[(1)] By means of these relations the product $A_1 \times \cdots \times A_k$ can be written in the form

$$A_1 \times \cdots \times A_k = \bigcup_{s_1,\dots,s_k} B_{1,s_1} \times \cdots \times B_{k,s_k} \tag{10.8}$$

with some sets B_{j,s_j} such that $\mu(B_{j,s_j}) < \varepsilon$ for all sets in this union. Moreover, we may assume, by refining the partitions of the sets A_j if this is necessary that any two sets B_{j,s_j} and $B_{j',s'_{j'}}$ in this representation are either disjoint, or they agree. Take such a representation of $A_1 \times \cdots \times A_k$, and consider the set we obtain by omitting those products $B_{1,s_1} \times \cdots \times B_{k,s_k}$ from the union at the right-hand side of (10.8) for which $B_{i,s_i} = B_{j,s_j}$ for some $1 \le i < j \le k$. The indicator function of the remaining set is in the class $\bar{\mathcal{H}}_{\mu,k}$. Hence it is enough to show that the distance between this indicator function and the indicator function of the set $A_1 \times \cdots \times A_k$ is less than const.ε in the $L_2(\mu^k)$ norm with some const. which may depend on the sets $A_1, \dots, A_k$, but not on ε. Indeed, by letting ε tend to zero we get from this relation that the indicator function of the set $A_1 \times A_2 \times \cdots \times A_k$ is in the closure of $\bar{\mathcal{H}}_{\mu,k}$ in the $L_2(\mu^k)$ norm.

Hence to prove the desired property of $\bar{\mathcal{H}}_{\mu,k}$ it is enough to prove the following statement. Take the representation (10.8) of $A_1 \times \cdots \times A_k$ (which depends on ε) and an arbitrary pair of integers i and j such that $1 \le i < j \le k$. Then the sum of the measures $\mu^k(B_{1,s_1} \times \cdots \times B_{k,s_k})$ of those sets $B_{1,s_1} \times \cdots \times B_{k,s_k}$ at the right-hand side of (10.8) for which $B_{i,s_i} = B_{j,s_j}$ is less than const.ε. To prove such an estimate observe

---

(1) For the sake of simplicity let us call a σ-finite measure μ on a measurable space $(X, \mathcal{X})$ non-atomic if for all sets $A \in \mathcal{X}$ such that $\mu(A) < \infty$ and numbers $\varepsilon > 0$ there is a finite partition $A = \bigcup_{s=1}^{N} B_s$ of the set A with the property $\mu(B_s) < \varepsilon$ for all $1 \le s \le N$. The original definition of a non-atomic measure is a formally weaker statement. It calls a measure μ non-atomic if for all A such that $0 < \mu(A) < \infty$ there exists a $B \subset A$ with the property $0 < \mu(B) < \mu(A)$. But the original definition of the non-atomic measure implies the property suggested in this footnote. (See e.g. Example 49 at the end of Chapter 2 in [*].) I omit the proof of this non-trivial statement, because it is a little bit outside from the direction of the present work.

that the μ^k measure of such a set can be bounded by the μ^{k-1} measure of the set we obtain by omitting the i-th term from the product defining it in the following way:

$$\mu^k(B_{1,s_1} \times \cdots \times B_{k,s_k}) \le \varepsilon \mu^{k-1}(B_{1,s_1} \times \cdots \times B_{i-1,s_{i-1}} \times B_{i+1,s_{i+1}} \times \cdots \times B_{k,s_k}).$$

Let us sum up this inequality for all such sets $B_{1,s_1} \times \cdots \times B_{k,s_k}$ at the right-hand side of (10.8) for which $B_{i,s_i} = B_{j,s_j}$. The left-hand side of the inequality we get in such a way equals the quantity we want to estimate. The expression at its right-hand side is less than $\varepsilon \prod_{1 \le s \le k,\, s \ne i} \mu(A_s)$, since ε-times the μ^{k-1} measure of such disjoint sets are summed up in it which are contained in the set $A_1 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_k$. In such a way we get the estimate we wanted to prove.

Knowing that $\bar{\mathcal{H}}_{\mu,k}$ is a dense subset of $\mathcal{H}_{\mu,k}$ in $L_2(\mu^k)$ norm we can finish the definition of k-fold Wiener–itô integrals in the standard way. Given any function $f \in \mathcal{H}_{\mu,k}$, a sequence of functions $f_n \in \bar{\mathcal{H}}_{\mu,k}$, $n = 1, 2, \dots$, can be defined in such a way that $\int |f(x_1, \dots, x_k) - f_n(x_1, \dots, x_k)|^2 \mu(\, dx_1) \dots \mu(\, dx_k) \to 0$ as $n \to \infty$. By relation (10.5) the normalizations $Z_{\mu,k}(f_n)$ of the already defined Wiener–Itô integrals of the functions f_n, $n = 1, 2, \dots$, constitute a Cauchy sequence in the space of square integrable random variables on the probability space, where the white noise is given. (Observe that the difference of two functions from the class $\bar{\mathcal{H}}_{\mu,k}$ also belongs to this class.) Hence the limit $\lim_{n\to\infty} Z_{\mu,k}(f_n)$ exists in L_2 norm, and this limit can be defined as the normalized Wiener–Itô integral $Z_{\mu,k}(f)$ of the function f. The definition of this limit does not depend on the choice of the approximating functions f_n, hence it is meaningful. It can be seen that relations (10.5) and (10.6) remain valid for all functions $f \in \mathcal{H}_{\mu,k}$. The following Theorem 10.1 describes the properties of multiple Wiener–Itô integrals. It contains already proved results. The only still non-discussed part of this Theorem is Property f) of Wiener–Itô integrals. But it is easy to check this property by observing that one-fold Wiener–Itô integrals are (jointly) Gaussian, they are measurable with respect to the σ-algebra generated by the white noise μ_W. Beside this, the random variable $\mu_W(A)$ for a set $A \in \mathcal{X}$, $\mu(A) < \infty$, equals the (one-fold) Wiener–Itô integral of the indicator function of the set A.

Theorem 10.1. (Characterization of multiple Wiener–Itô integrals). *Let a white noise μ_W be given with some non-atomic σ-additive reference measure on a measurable space $(X, \mathcal{X})$. Then the k-fold Wiener–Itô integral of all functions in the class $\mathcal{H}_{\mu,k}$ introduced in formula (10.1) can be defined, and its normalized version $Z_{\mu,k}(f) = \frac{1}{k!} \int f(x_1, \dots, x_k) \mu_W(\, dx_1) \dots \mu_W(dx_k)$ satisfies the following relations:*

a) $Z_{\mu,k}(\alpha f + \beta g) = \alpha Z_{\mu,k}(f) + \beta Z_{\mu,k}(g)$ for all $f, g \in \mathcal{H}_{\mu,k}$ and real numbers α and β.

b) If $A_1, \dots, A_k$ are disjoint sets, $\mu(A_j) < \infty$, then the function $f_{A_1,\dots,A_k}$ defined by the relation $f_{A_1,\dots,A_k}(x_1, \dots, x_k) = 1$ if $x_1 \in A_1$, $\dots$, $x_k \in A_k$, $f_{A_1,\dots,A_k}(x_1, \dots, x_k) = 0$ otherwise, satisfies the identity

$$Z_{\mu,k}(f_{A_1,\dots,A_k}(x_1, \dots, x_k)) = \frac{1}{k!} \mu_W(A_1) \cdots \mu_W(A_k).$$

c)

$$EZ_{\mu,k}(f) = 0, \quad \textit{and} \quad EZ^2_{\mu,k}(f) = \frac{1}{k!}\|\mathrm{Sym}\, f\|_2^2 \le \frac{1}{k!}\|f\|_2^2$$

for all $f \in \mathcal{H}_{\mu,k}$, where $\|f\|_2^2 = \int f^2(x_1, \dots, x_k)\mu(\,dx_1)\dots\mu(\,dx_k)$ is the square of the L_2 norm of a function $f \in \mathcal{H}_{\mu,k}$.

d) Relation (10.6) holds for all functions $f \in \mathcal{H}_{\mu,k}$ and $g \in \mathcal{H}_{\mu,k'}$ if $k \neq k'$.

e) Relation (10.7) holds for all functions $f \in \mathcal{H}_{\mu,k}$.

f) The Wiener–Itô integrals $Z_{\mu,1}(f)$ of order $k = 1$ are jointly Gaussian. The smallest σ-algebra with respect to which they are all measurable agrees with the σ-algebra generated by the random variables $\mu_W(A)$, $A \in \mathcal{X}$, $\mu(A) < \infty$, of the white noise.

We have defined Wiener–Itô integrals of order k for all $k = 1, 2, \dots$. For the sake of completeness let us introduce the class $\mathcal{H}_{\mu,0}$ for $k = 0$ which consists of the real constants (functions of zero variables), and put $Z_{\mu,0}(c) = c$. Because of relation (10.7) we could have restricted our attention to Wiener–Itô integrals with symmetric kernel functions. But it turned out more convenient to work also with Wiener–Itô integrals of not necessarily symmetric functions.

Now I formulate the diagram formula for the product of two Wiener–Itô integrals. For this goal some notations have to be introduced. To present the product of the multiple Wiener–Itô integrals of two functions $f(x_1, \dots, x_k) \in \mathcal{H}_{\mu,k}$ and $g(x_1, \dots, x_l) \in \mathcal{H}_{\mu,l}$ in the form of sums of Wiener–Itô integrals a class of diagrams $\Gamma = \Gamma(k, l)$ will be defined. The diagrams $\gamma \in \Gamma(k, l)$ have vertices $(1, 1), \dots, (1, k)$ and $(2, 1), \dots, (2, l)$, and edges $((1, j_1), (2, j'_1)), \dots, ((1, j_s), (2, j'_s))$ with some $1 \le s \le \min(k, l)$. The indices $j_1, \dots, j_s$ in the definition of the edges are all different, and the same relation holds for the indices $j'_1, \dots, j'_s$. All such diagrams γ belongs to $\Gamma(k, l)$. The set of vertices of the form $(1, j)$, $1 \le j \le k$, will be called the first row, and the set of vertices of the form $(2, j')$, $1 \le j' \le l$, the second row of a diagram. We demanded that edges of a diagram can connect only vertices of different rows, and at most one edge may start from each vertex of the diagram.

Given a diagram $\gamma \in \Gamma(k, l)$ with the set of edges

$$E(\gamma) = \{(1, j_1), (2, j'_1)), \dots, ((1, j_s), (2, j'_s)\}$$

let $V_1(\gamma) = \{(1, 1), \dots, (1, k)\} \setminus \{(1, j_1), \dots, (1, j_s)\}$ and $V_2(\gamma) = \{(2, 1), \dots, (2, l)\} \setminus \{(2, j'_1), \dots, (2, j'_s)\}$ denote the set of vertices in the first and in the second row of the diagram γ respectively from which no edge starts. Put $\alpha_\gamma(1, j) = (2, j')$ if $((1, j), (2, j')) \in E(\gamma)$ and $\alpha_\gamma(1, j) = (1, j)$ if the diagram γ contains no edge of the form $((1, j), (2, j')) \in E(\gamma)$. In words, the function $\alpha_\gamma(\cdot)$ is defined on the vertices of the first row of the diagram γ, it replaces a vertex to the vertex it is connected to by an edge of the diagram if there is such a vertex, and it does not change those vertices from which no edge starts. Put $|\gamma| = k + l - 2s$, i.e. $|\gamma|$ equals the number of vertices in γ from which no edge starts.

Given two functions $f(x_1,\dots,x_k)\in\mathcal{H}_{\mu,k}$ and $g(x_1,\dots,x_l)\in\mathcal{H}_{\mu,l}$ let us introduce their product

$$\begin{aligned}&F(x_{(1,1)},\dots,x_{(1,k)},x_{(2,1)},\dots,x_{(2,l)})\\&\quad=F_{f,g}(x_{(1,1)},\dots,x_{(1,k)},x_{(2,1)},\dots,x_{(2,l)})\\&\quad=f(x_{(1,1)},\dots,x_{(1,k)})g(x_{(2,1)},\dots,x_{(2,l)})\end{aligned}\tag{10.9}$$

together with its modification

$$\begin{aligned}&\bar F_\gamma(x_{(1,j)},:\ (1,j)\in V_1(\gamma),\ x_{(2,1)}\dots,x_{(2,l)})\\&\quad=f(x_{\alpha_\gamma(1,1)},\dots,x_{\alpha_\gamma(1,k)})g(x_{(2,1)},\dots,x_{(2,l)}).\end{aligned}\tag{10.9a}$$

(Here the function $f(x_1,\dots,x_k)$ is replaced by $f(x_{(1,1)},\dots,x_{(1,k)})$ and the function $g(x_1,\dots,x_l)$ by $g(x_{(2,1)},\dots,x_{(2,l)})$.) With the help of the above introduced sets $V_1(\gamma)$, $V_2(\gamma)$ and function $\alpha_\gamma(\cdot)$ let us introduce the functions $F_\gamma=F_\gamma(f,g)$ as

$$\begin{aligned}&F_\gamma(x_{(1,j)},x_{(2,j')}\colon\ (1,j)\in V_1(\gamma),\ (2,j')\in V_2(\gamma))\\&\quad=\int\bar F_\gamma(x_{\alpha_\gamma(1,j)}\colon\ (1,j)\in V_1(\gamma),\ x_{(2,1)},\dots,x_{(2,l)})\\&\qquad\prod_{(2,j)\in\{(2,1),\dots,(2,l)\}\setminus V_2(\gamma)}\mu(\,dx_{(2,j)})\end{aligned}\tag{10.10}$$

for all diagrams $\gamma\in\Gamma(k,l)$. In words: We take the product defined in (10.9), then if the index $(1,j)$ of a variable $x_{(1,j)}$ is connected with the index $(2,j')$ of some variable $x_{(2,j')}$ by an edge of the diagram γ, then we replace the variable $x_{(1,j)}$ by $x_{(2,j')}$ in this product. Finally we integrate the function obtained in such a way with respect to the arguments with indices from the set $\{(2,1),\dots,(2,l)\}\setminus V_2(\gamma)$. It is clear that F_γ is a function of $|\gamma|$ variables. It depends on those coordinates whose indices are such vertices of γ from which no edge starts.

For the sake of simpler notations we shall also consider Wiener–Itô integrals with such kernel functions whose variables are more generally indexed. If the k-fold Wiener–Itô integral with a kernel function $f(x_1,\dots,x_k)$ is well-defined, then we shall say that the Wiener–Itô integral with kernel function $f(x_{u_1},\dots,x_{u_k})$, where $\{u_1,\dots,u_k\}$ is an arbitrary set with k different elements, is also well defined, and it equals the Wiener–Itô integral with the original kernel function $f(x_1,\dots,x_k)$. (We have right to make such a convention since the value of a Wiener–Itô integral does not change if we permute the indices of the variables of the kernel function in an arbitrary way.) In particular, we shall speak about the Wiener–Itô integral of the function F_γ defined in (10.10) without reindexing its variables $x_{(1,j)}$ and $x_{(2,j')}$ 'in the right way'. Now we can formulate the diagram formula for the product of two Wiener–Itô integrals.

Theorem 10.2A. (The diagram formula for the product of two Wiener–Itô integrals). *Let a non-atomic σ-finite measure μ be given on a measurable space $(X,\mathcal{X})$ together with a white noise μ_W with reference measure μ, and take two functions $f(x_1,\dots,x_k)\in\mathcal{H}_{\mu,k}$ and $g(x_1,\dots,x_l)\in\mathcal{H}_{\mu,l}$. (The classes of functions $\mathcal{H}_{\mu,k}$ and*

$\mathcal{H}_{\mu,l}$ were introduced in (10.1).) Let us consider the class of diagrams $\Gamma(k,l)$ introduced above together with the functions F_γ, $\gamma \in \Gamma(k,l)$, defined by formulas (10.9), (10.9a) and (10.10) with its help. They satisfy the inequality

$$\|F_\gamma\|_2 \le \|f\|_2 \|g\|_2 \quad \textit{for all } \gamma \in \Gamma(k,l), \tag{10.11}$$

where the L_2 norm of a (generally indexed) function $h(x_{u_1}, \dots, x_{u_s})$ is defined as

$$\|h\|_2^2 = \int h^2(x_{u_1}, \dots, x_{u_s}) \mu(\, dx_{u_1}) \dots \mu\, dx_{u_s}).$$

Beside this, the product $Z_{\mu,k}(f) Z_{\mu,l}(g)$ of the normalized Wiener–Itô integrals of the functions f and g (the notation $Z_{\mu,k}$ was introduced in (10.4)) satisfies the identity

$$(k! Z_{\mu,k}(f))(l! Z_{\mu,l}(g)) = \sum_{\gamma \in \Gamma(k,l)} |\gamma|! Z_{\mu,|\gamma|}(F_\gamma) = \sum_{\gamma \in \Gamma(k,l)} |\gamma|! Z_{\mu,|\gamma|}(\mathrm{Sym}\, F_\gamma). \tag{10.12}$$

Theorem 10.2A will be proved in Appendix B. The following consideration yields a heuristic explanation for it. Actually, it can also be considered as a sketch of proof.

In the theory of general Itô integrals when stochastic processes are integrated with respect to a Wiener processes, one of the most basic results is Itô's formula about differentiation of functions of Itô integrals. It has a heuristic interpretation by means of the informal 'identity' $dW^2 = dt$. In the case of general white noises this 'identity' can be generalized as $\mu_W(\, dx)^2 = \mu(\, dx)$. Next we present a rather informal 'proof' of the diagram formula on the basis of this 'identity' and the fact that the diagonals are omitted from the domain of integration in the definition of Wiener–Itô integrals.

In this 'proof' we fix two numbers $k \ge 1$ and $l \ge 1$, and consider the product of the Wiener–Itô integrals of the functions f and g of order k and l. This product is a bilinear form of the functions f and g. Hence it is enough to check formula (10.12) for a sufficiently rich class of functions. It is enough to consider functions of the form $f(x_1, \dots, x_k) = I_{A_1}(x_1) \cdots I_{A_k}(x_k)$ and $g(x_1, \dots, x_l) = I_{B_1}(x_1) \cdots I_{B_l}(x_l)$ with disjoint sets $A_1, \dots, A_k$ and disjoint sets $B_1, \dots, B_l$, where $I_A(x)$ is the indicator function of a set A. (It is exploited at this point that the functions f and g disappear at the diagonals.) Let us divide the sets A_j into the union of small disjoint sets $D_j^{(m)}$, $1 \le j \le k$ with some fixed number $1 \le m \le M$ in such a way that $\mu(D_j^{(m)}) \le \varepsilon$ with some fixed $\varepsilon > 0$, and the sets B_j into the union of small disjoint sets $F_j^{(m)}$, $1 \le j \le l$, with some fixed number $1 \le m \le M$, in such a way that $\mu(F_j^{(m)}) \le \varepsilon$ with some fixed $\varepsilon > 0$. Beside this, we also require that two sets $D_j^{(m)}$ and $F_{j'}^{(m')}$ should be either disjoint or they should agree. (The sets $D_j^{(m)}$ are disjoint for different indices, and the same relation holds for the sets $F_{j'}^{(m')}$.)

Then the identity

$$k! Z_{\mu,k}(f) = \prod_{j=1}^{k} \left(\sum_{m=1}^{M} \mu_W(D_j^{(m)}) \right) \quad \text{and} \quad l! Z_{\mu,l}(g) = \prod_{j'=1}^{l} \left(\sum_{m'=1}^{M} \mu_W(F_{j'}^{(m')}) \right),$$

holds, and the product of these two Wiener–Itô integrals can be written in the form of a sum by means of a term by term multiplication. Let us divide the terms of the sum we get in such a way into subgroups indexed by the diagrams $\gamma \in \Gamma(k,l)$ in the following way: Each term in this sum is a product of the form $\prod_{j=1}^{k} \mu_W(D_j^{(m_j)}) \prod_{j'=1}^{l} \mu_W(F_{j'}^{(m'_j)})$. Let it belong to the subgroup indexed by the diagram γ with edges $((1,j_1),(2,j'_1)), \dots$, and $((1,j_s),(2,j'_s))$ if the elements in the pairs $(D_{j_1}^{m_{j_1}}, F_{j'_1}^{m_{j'_1}}), \dots , (D_{j_s}^{m_{j_s}}, F_{j'_s}^{m_{j'_s}})$ agree, and otherwise all terms are different. Then letting $\varepsilon \to 0$ (and taking partitions of the sets D_j and $F_{j'}$ corresponding to the parameter ε) the sums of the terms in each subgroup turn to integrals, and our calculation suggests the identity

$$(k!Z_{\mu,k}(f))(l!Z_{\mu,l}(g)) = \sum_{\gamma \in \Gamma(k,l)} \bar{Z}_\gamma \tag{10.13}$$

with

$$\begin{aligned} \bar{Z}_\gamma = \int & f(x_{\alpha_\gamma(1,1)}, \dots, x_{\alpha_\gamma(1,k)}) g(x_{(2,1)}, \dots, x_{(2,l)}) \\ & \mu_W(\, dx_{\alpha_\gamma(1,1)}) \dots \mu_W(\, dx_{\alpha_\gamma(1,k)}) \mu_W(\, dx_{(2,1)}) \dots \mu_W(\, dx_{(2,l)}) \end{aligned} \tag{10.13a}$$

with the function $\alpha_\gamma(\cdot)$ introduced before formula (10.9). The indices $\alpha(1,j)$ of the arguments in (10.13a) mean that in the case $\alpha_\gamma(1,j) = (2,j')$ the argument $x_{(1,j)}$ has to be replaced by $x_{(2,j')}$. In particular, $\mu_W(\, dx_{\alpha(1,j)}) \mu_W(\, dx_{(2,j')}) = \mu_W(\, dx_{(2,j')})^2 = \mu(\, dx_{(2,j')})$ in this case because of the 'identity' $\mu_W(\, dx)^2 = \mu(\, dx)$. Hence the above informal calculation yields the identity $\bar{Z}_\gamma = |\gamma|! Z_{\mu,|\gamma|}(F_\gamma)$. Thus relations (10.13) and (10.13a) imply formula (10.12).

Similar heuristic argument can be applied to get formulas for the product of integrals of normalized empirical distributions or Poisson processes, only the starting formula $\mu_W(\, dx)^2 = \mu(\, dx)$ changes in these cases, some additional terms appear which modify the final result. I return to this question in the next section.

It is not difficult to generalize Theorem 10.2A with the help of some additional notations to a diagram formula about the product of finitely many Wiener–Itô integrals. Let us consider $m \geq 2$ Wiener–Itô integrals $k_p! Z_{\mu,k_p}(f_p)$, of functions $f_p(x_1, \dots, x_{k_p}) \in \mathcal{H}_{\mu,k_p}$, of order $k_p \geq 1$, $1 \leq p \leq m$, and define a class of diagrams $\Gamma = \Gamma(k_1, \dots, k_m)$ in the following way.

The diagrams $\gamma \in \Gamma = \Gamma(k_1, \dots, k_m)$ have vertices of the form (p,r), $1 \leq p \leq m$, $1 \leq r \leq k_p$. The set of vertices $\{(p,r)\colon\ 1 \leq r \leq k_p\}$ with a fixed number p will be called the p-th row of the diagram γ. A diagram $\gamma \in \Gamma = \Gamma(k_1, \dots, k_m)$ may have some edges. All edges of a diagram connect vertices from different rows, and from each vertex there starts at most one edge. All diagrams satisfying these properties belong to $\Gamma(k_1, \dots, k_m)$. If a diagram γ contains an edge of the form $((p_1,r_1),(p_2,r_2))$ with $p_1 < p_2$, then (p_1,r_1) will be called the upper and (p_2,r_2) the lower end point of this edge. Let $E(\gamma) = \{((p_1^{(u)}, r_1^{(u)}), (p_2^{(u)}, r_2^{(u)})),\ p_1^{(u)} < p_2^{(u)},\ 1 \leq u \leq s\}$ denote the set of

all edges of a diagram γ (the number of edges in γ was denoted by s), and let us also introduce the sets $V^u(\gamma) = \{((p_1^{(u)}, r_1^{(u)}),\ 1 \le u \le s\}$, the set of all upper end points and $V^b(\gamma) = \{((p_2^{(u)}, r_2^{(u)}),\ 1 \le u \le s\}$, the set of all lower end points of edges in a diagram γ. Let $V = V(\gamma) = \{(p,r)\colon\ 1 \le p \le m, 1 \le r \le k_p\}$ denote the set of all vertices of γ, and let $|\gamma| = k_1 + \cdots + k_m - 2|E(\gamma)|$ be equal to the number of vertices in γ from which no edge starts. Let us also define the function $\alpha_\gamma(p,r)$ for a vertex (p,r) of the diagram γ in the following way: $\alpha_\gamma(p,r) = (\bar p, \bar r)$, if there is some pair of integers $(\bar p, \bar r)$ such that $((p,r),(\bar p,\bar r)) \in E(\gamma)$ and $p < \bar p$, i.e. $(p,r) \in V^u(\gamma)$ and $((p,r),(\bar p,\bar r)) \in E(\gamma)$, and put $\alpha_\gamma(p,r) = (p,r)$ for $(p,r) \in V(\gamma) \setminus V^u(\gamma)$. In words, the function $\alpha_\gamma(\cdot)$ was defined on the set of vertices $V(\gamma)$ in such a way that it replaces an upper end point of an edge with the lower end point of this edge, and it does not change the remaining vertices of the diagram.

With the help of the above quantities the appropriate multivariate version of the functions given in (10.9), (10.9a) and (10.10) can be defined. Put

$$\begin{aligned} F(x_{(p,r)},\colon\ 1 \le p \le m, 1 \le r \le k_p) &= F_{f_1,\dots,f_m}(x_{(p,r)},\colon\ 1 \le p \le m, 1 \le r \le k_p) \\ &= \prod_{p=1}^m f_p(x_{(p,1)},\dots,x_{(p,k_p)}), \end{aligned} \tag{10.14}$$

$$\bar F_\gamma(x_{(p,r)},\colon\ (p,r) \in V(\gamma) \setminus V^u(\gamma)) = \prod_{p=1}^m f_p(x_{\alpha_\gamma(p,1)},\dots,x_{\alpha_\gamma(p,k_p)}), \tag{10.14a}$$

and

$$\begin{aligned} &F_\gamma(x_{(p,r)}\colon\ (p,r) \in V(\gamma) \setminus (V^b(\gamma) \cup V^u(\gamma)) \\ &\quad = \int \bar F_\gamma(x_{(p,r)},\colon\ (p,r) \in V(\gamma) \setminus V^b(\gamma)) \prod_{(p,r)\in V^b(\gamma)} \mu(\,dx_{(p,r)}). \end{aligned} \tag{10.15}$$

With the help of the above notations the diagram formula for the product of arbitrarily many Wiener–Itô integrals can be formulated.

Theorem 10.2. (The diagram formula for the product of finitely many Wiener–Itô integrals). *Let a non-atomic σ-finite measure μ be given on a measurable space $(X, \mathcal{X})$ together with a white noise μ_W with reference measure μ. Take $m \ge 2$ functions $f_p(x_1,\dots,x_{k_p}) \in \mathcal{H}_{\mu,k_p}$ with some order $k_p \ge 1$, $1 \le p \le m$. Let us consider the class of diagrams $\Gamma(k_1,\dots,k_m)$ introduced above together with the functions F_γ, $\gamma \in \Gamma(k_1,\dots,k_m)$, defined by formulas (10.14), (10.14a) and (10.15) with its help. The L_2-norm of these functions satisfies the inequality*

$$\|F_\gamma\|_2 \le \prod_{p=1}^m \|f_p\|_2 \quad \textit{for all } \gamma \in \Gamma(k_1,\dots,k_m). \tag{10.16}$$

Beside this, the product $\prod_{p=1}^{m} Z_{\mu,k_p}(f_p)$ *of the normalized Wiener–Itô integrals of the functions* f_p, $1 \le p \le m$, *satisfies the identity*

$$\prod_{p=1}^{m} k_p! Z_{\mu,k_p}(f_p) = \sum_{\gamma \in \Gamma(k_1,\dots,k_m)} |\gamma|! Z_{\mu,|\gamma|}(F_\gamma) = \sum_{\gamma \in \Gamma(k_1,\dots,k_m)} |\gamma|! Z_{\mu,|\gamma|}(\mathrm{Sym}\, F_\gamma). \tag{10.17}$$

Theorem 10.2 can be relatively simply derived from Theorem 10.2A by means of induction with respect the number of terms whose product we consider. We still have to check that with the introduction of an appropriate notation Theorem 10.2A remains valid also in the case when the function f is a constant.

Let us also consider the case when $f = c$ and $g \in \mathcal{H}_{\mu,l}$. In this case we apply the convention $Z_{\mu,0}(c) = c$, define the class of diagrams $\Gamma(0, l)$ that consists only of one diagram γ whose first row is empty, its second row contains the vertices $(2,1), \dots, (2,l)$, and it has no edges. Beside this, we define $F_\gamma(x_{(2,1)}, \dots, x_{(2,l)}) = cg(x_{(2,1)}, \dots, x_{(2,l)})$ in this case. With such a convention Theorem 10.2A can be extended to the case of the product of two Wiener–Itô integrals of order $k \ge 0$ and $l \ge 1$. Theorem 10.2 can be derived from this slightly generalized result by induction.

By statement c) of Theorem 10.1 all Wiener–Itô integrals of order $k \ge 1$ have expectation zero. This fact together with Theorem 10.2 enable us to compute the expectation of a product of Wiener–Itô integrals. Theorem 10.2 makes possible to rewrite a product of Wiener–Itô integrals as a sum of Wiener–Itô integrals. Then its expectation can be calculated by taking the expected value of each term and summing them up. Only constant terms yield a non-zero contribution to this expectation. These constant terms agree with the functions F_γ corresponding to diagrams with no free vertices. The next corollary writes down the result we get in such a way.

Corollary of Theorem 10.2 about the expectation of a product of Wiener–Itô integrals. *Let a non-atomic* σ*-finite measure* μ *be given on a measurable space* $(X, \mathcal{X})$ *together with a white noise* μ_W *with reference measure* μ. *Take* $m \ge 2$ *functions* $f_p(x_1, \dots, x_{k_p}) \in \mathcal{H}_{\mu,k_p}$, *and consider their Wiener–Itô integrals* $Z_{\mu,k_p}(f_p)$, $1 \le p \le m$. *The expectation of the product of these random variables satisfies the identity*

$$E\left(\prod_{p=1}^{m} k_p! Z_{\mu,k_p}(f_p)\right) = \sum_{\gamma \in \bar{\Gamma}(k_1,\dots,k_m)} F_\gamma, \tag{10.18}$$

where $\bar{\Gamma}(k_1, \dots, k_m)$ *denotes the set of all such diagrams* $\gamma \in \Gamma(k_1, \dots, k_m)$ *which have no free vertices, i.e.* $|\gamma| = 0$. *Such diagrams will be called closed in the sequel. (If* $\bar{\Gamma}(k_1, \dots, k_m)$ *is empty, then the sum at the right-hand side of (10.17) equals zero.) The functions* F_γ *for* $\gamma \in \bar{\Gamma}(k_1, \dots, k_m)$ *are constants, and they satisfy the inequality*

$$|F_\gamma| \le \prod_{p=1}^{m} \|f_p\|_2 \quad \textit{for all } \gamma \in \bar{\Gamma}(k_1, \dots, k_m). \tag{10.19}$$

Proof of the Corollary. Relation (10.18) is a straight consequence of formula (10.17), part c) of Theorem 10.1 and the identity $Z_{\mu,0}(F_\gamma) = F_\gamma$, if $|\gamma| = 0$. Relation (10.19) follows from (10.16).

The next result I formulate, Itô's formula for multiple Wiener–Itô integrals, can also be considered as a consequence of the diagram formula. It will be proved in Appendix C.

Theorem 10.3. (Itô's formula for multiple Wiener–Itô integrals). *Let a non-atomic σ-finite measure μ be given on a measurable space $(X, \mathcal{X})$ together with a white noise μ_W with reference measure μ. Let us take some real valued, orthonormal functions $\varphi_1(x), \dots, \varphi_m(x)$ on the measure space $(X, \mathcal{X}, \mu)$. Let $H_k(u)$ denote the k-th Hermite polynomial with leading coefficient 1. Take the one-fold Wiener–Itô integrals $\eta_p = Z_{\mu,1}(\varphi_p)$, $1 \le p \le m$, and introduce the random variables $H_{k_p}(\eta_p)$, $1 \le p \le m$, with some integers $k_p \ge 1$, $1 \le p \le m$. Put $K_p = \sum_{j=1}^{p} k_r$, $1 \le p \le m$, $K_0 = 0$. Then $\eta_1, \dots, \eta_m$ are independent, standard normal random variables, and the identity*

$$\begin{aligned} \prod_{p=1}^{m} H_{k_p}(\eta_p) &= K_m! Z_{\mu,K_m}\left(\prod_{p=1}^{m}\left(\prod_{j=K_{p-1}+1}^{K_p} \varphi_p(x_j)\right)\right) \\ &= K_m! Z_{\mu,K_m}\left(\mathrm{Sym}\left(\prod_{p=1}^{m}\left(\prod_{j=K_{p-1}+1}^{K_p} \varphi_p(x_j)\right)\right)\right) \end{aligned} \tag{10.20}$$

holds. In particular, for a single real valued function $\varphi(x)$ such that $\int \varphi^2(x)\mu(\,dx) = 1$

$$H_k\left(\int \varphi(x)\mu_W(\,dx)\right) = \int \varphi(x_1)\cdots\varphi(x_k)\mu_W(\,dx_1)\dots\mu_W(\,dx_k). \tag{10.21}$$

I also formulate a limit theorem about the distribution of normalized degenerate U-statistics. The limit distribution in this result can be described by means of multiple Wiener–Itô integrals. It will be proved in Appendix C.

Theorem 10.4. (Limit theorem about normalized degenerate U-statistics). *Let us consider a sequence of degenerate U-statistics $I_{n,k}(f)$ of order k, $n = k, k+1, \dots$, defined in (8.7) with the help of a sequence of independent and identically distributed random variables $\xi_1, \xi_2, \dots$ taking values in a measurable space $(X, \mathcal{X})$ with a non-atomic distribution μ and a kernel function $f(x_1, \dots, x_k)$, canonical with respect to the measure μ, defined on the k-fold product $(X^k, \mathcal{X}^k)$ of the space $(X, \mathcal{X})$ for which $\int f^2(x_1, \dots, x_k)\mu(\,dx_1)\dots\mu(\,dx_k) < \infty$. Then the sequence of normalized U-statistics $n^{-k/2} I_{n,k}(f)$ converges in distribution, as $n \to \infty$, to the k-fold Wiener–Itô integral*

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

with kernel function $f(x_1, \dots, x_k)$ and a white noise μ_W with reference measure μ.

Remark. The limit behaviour of degenerate U-statistics $I_{n,k}(f)$ with an atomic measure μ which satisfy the remaining conditions of Theorem 10.4 can be described in the following way. Take the probability space $(U, \mathcal{U}, \lambda)$, where $U = [0,1]$, $\mathcal{U}$ is the Borel σ-algebra and λ is the Lebesgue measure on it. Introduce a sequence of indepedent random variables $\eta_1, \eta_2, \dots$ with uniform distribution on the interval $[0,1]$, which is independent also of the sequence $\xi_1, \xi_2, \dots$. Define the product space $(\tilde{X}, \tilde{\mathcal{X}}, \tilde{\mu}) = (X \times U, \mathcal{X} \times \mathcal{U}, \mu \times \lambda)$ together with the function $\tilde{f}(\tilde{x}_1, \dots, \tilde{x}_k) = \tilde{f}((x_1, u_1), \dots, (x_k, u_k)) = f(x_1, \dots, x_k)$ with the notation $\tilde{x} = (x, u) \in X \times U$, and $\tilde{\xi}_j = (\xi_j, \eta_j)$, $j = 1, 2, \dots$. Then $I_{n,k}(f) = I_{n,k}(\tilde{f})$ (with the above defined function $\tilde{f}$ and $\tilde{\mu}$ distributed random variables $\tilde{\xi}_j$). Beside this, Theorem 10.4 can be applied for the degenerate U-statistics $I_{n,k}(\tilde{f})$, $n = 1, 2, \dots$.

In the next result I give an interesting representation of the Hilbert space consisting of the square integrables functions measurable with respect to a white noise μ_W. An isomorphism will be given with the help of Wiener–Itô integrals between this Hilbert space and the so-called Fock space to be defined below. To formulate this result first some notations will be introduced.

Let $\mathcal{H}^0_{\mu,k} \subset \mathcal{H}_{\mu,k}$ denote the class of symmetric functions in the space $\mathcal{H}_{\mu,k}$, $k = 0, 1, 2, \dots$, i.e. $f \in \mathcal{H}_{\mu,k}$ is in its subspace $\mathcal{H}^0_{\mu,k}$ if and only if $f(x_1, \dots, x_k) = \operatorname{Sym} f(x_1, \dots, x_k)$. Let us introduce for all $k = 0, 1, 2, \dots$ the Hilbert space $\mathcal{G}_k$ consisting of those random variables η (on the probability space where the white noise μ_W is defined) which can be written in the form

$$\eta = Z_{\mu,k}(f) = \frac{1}{k!} \int f(x_1, \dots, x_k) \mu_W(\,dx_1) \dots \mu_W(\,dx_k) \quad \text{with some f } \in \mathcal{H}^0_{k,\mu}.$$

It follows from part a) and c) of Theorem 10.1 that the map $f \to Z_{\mu,k}(f)$ is a linear transformation of $\mathcal{H}^0_{\mu,k}$ to $\mathcal{G}_k$, and $\frac{1}{k!}\|f\|_2^2 = EZ^2_{\mu,k}(f)$ for all $f \in \mathcal{H}^0_{\mu,k}$, where $\|f\|_2$ denotes the usual L_2-norm of the function f with respect to the k-fold power of the measure μ. By the definition of Wiener–Itô integrals the set $\mathcal{G}_1$ consists of jointly Gaussian random variables with expectation zero. The spaces $\mathcal{H}_{\mu,0}$ and $\mathcal{G}_0$ consist of the real constants. Let us define the space $\operatorname{Exp}(\mathcal{H}_\mu)$ of infinite sequences $f = (f_0, f_1, \dots)$, $f_k \in \mathcal{H}^0_{\mu,k}$, $k = 0, 1, 2, \dots$, such that $\|f\|_2^2 = \sum_{k=0}^{\infty} \frac{1}{k!}\|f_k\|_2^2 < \infty$. The space $\operatorname{Exp}(\mathcal{H}_\mu)$ with the natural addition and multiplication by a constant and the above introduced norm $\|f\|_2$ for $f \in \operatorname{Exp}(H_\mu)$ is a Hilbert space which is called the Fock space in the literature.

Let $\mathcal{G}$ denote the class of random variables of the form

$$Z(f) = \sum_{k=0}^{\infty} Z_{\mu,k}(f_k), \quad f = (f_0, f_1, f_2, \dots) \in \operatorname{Exp}(\mathcal{H}_\mu).$$

The next result describes the structure of the space of random variables $\mathcal{G}$. It is useful for a better understanding of Wiener–Itô integrals, but it will be not used in the sequel. In its proof I shall refer to some basic measure theoretical results.

Theorem 10.5. (Isomorphism of the space of square integrable random variables measurable with respect to a white noise with a Fock space). *Let a non-atomic σ-finite measure μ be given on a measurable space $(X, \mathcal{X})$ together with a white noise μ_W with reference measure μ. Let us consider the class of functions $\mathcal{H}^0_{\mu,k}$, $k = 0, 1, 2, \dots$, and $\mathrm{Exp}\,(\mathcal{H}_\mu)$ together with the spaces of random variables $\mathcal{G}_k$, $k = 0, 1, 2, \dots$, and $\mathcal{G}$ defined above. The transformation Z: $Z(f) = \sum_{k=0}^{\infty} Z_{\mu,k}(f_k)$, $f = (f_0, f_1, f_2, \dots) \in \mathrm{Exp}\,(\mathcal{H}_\mu)$, is a unitary transformation from the Hilbert spaces $\mathrm{Exp}\,(\mathcal{H}_\mu)$ to $\mathcal{G}$. The Hilbert space $\mathcal{G}$ consists of all random variables with finite second moment, measurable with respect to the σ-algebra generated by the random variables $\mu_W(A)$, $A \in \mathcal{X}$, $\mu(A) < \infty$. This σ-algebra agrees with the σ-algebra generated by the random variables $Z_{\mu,1}(f_1)$, $f_1 \in \mathcal{H}^0_{\mu,1}$.*

Proof of Theorem 10.5. Properties a) and c) in Theorem 10.1 imply that the transformation $f_k \to Z_{\mu,k}(f_k)$ is a linear transformation of $\mathcal{H}^0_{\mu,k}$ to $\mathcal{G}_k$, and $\frac{1}{k!}\|f_k\|_2^2 = EZ_{\mu,k}(f)^2$. Beside this, $EZ_{\mu,k}(f)Z_{\mu,k'}(f'_{k'}) = 0$ if $f_k \in \mathcal{H}^0_{\mu,k}$, and $f'_{k'} \in \mathcal{H}^0_{\mu,k'}$ with $k \neq k'$ by properties d) and c). (The latter property is needed to guarantee this relation also holds if $k = 0$ or $k' = 0$.) From these relations follows that the map Z: $Z(f) = \sum_{k=0}^{\infty} Z_{\mu,k}(f_k)$, $f = (f_0, f_1, f_2, \dots) \in \mathrm{Exp}\,(\mathcal{H}_\mu)$ is an isomorphism between the Hilbert spaces $\mathrm{Exp}\,(\mathcal{H}_\mu)$ and $\mathcal{G}$.

It remained to show that $\mathcal{G}$ contains all random variables with finite second moment, measurable with respect to the corresponding σ-algebra. Let $g_j(u)$, $j = 1, 2, \dots$, be an orthonormal basis in $\mathcal{H}^0_{\mu,1} = \mathcal{H}_{\mu,1}$, and introduce the random variables $\eta_j = Z_{\mu,1}(g_j)$, $j = 1, 2, \dots$. By Itô's formula for Wiener–Itô integrals (Theorem 10.3) these random variables are independent with standard normal distribution, and all expressions of the form $H_{r_1}(\eta_{j_1}) \dots H_{r_p}(\eta_{j_p})$ with $r_1 + \dots + r_p = k$ are in the space $\mathcal{G}_k$, where $H_r(\cdot)$ denotes the Hermite polynomial of order r with leading coefficient 1. To prove the desired statement by means of these relations we still need the following results from the classical analysis:

a) Hermite polynomials constitute a complete orthonormal system in the L_2-space on the real line with respect to the standard normal distribution. (This result will be proved in Section C in Proposition C2.)

b) If a random variable ζ is measurable with respect to the σ-algebra generated by some random variables $\eta_1, \eta_2, \dots$, then there exists a Borel measurable function $f(x_1, x_2, \dots)$ on the infinite product of the real line $(R^\infty, \mathcal{B}^\infty)$ in such a way that $\zeta = f(\eta_1, \eta_2, \dots)$.

This means in our case that any random variable ζ measurable with respect to the σ-algebra generated by the random variables $\eta_j = Z_{\mu,1}(g_j)$, $j = 1, 2, \dots$, can be written in the form $\zeta = f(\eta_1, \eta_2, \dots)$ with the above introduced independent, standard normal random variables $\eta_1, \eta_2, \dots$. If it has finite second moment, then the function f appearing in its representation has finite L_2-norm with respect to the infinite power of the standard normal distribution. Hence some results about orthogonal basis on

product spaces make possible to expand the function f with respect to product of Hermite polynomials, and this yields that

$$\zeta = \sum c(j_1, r_1, \dots, j_s, r_s) H_{r_1}(\eta_{j_1}) \cdots H_{r_s}(\eta_{j_s})$$

with some coefficients $c(j_1, r_1, \dots, j_s, r_s)$ such that

$$\sum c^2(j_1, r_1, \dots, j_s, r_s) \|H_{r_1}(u)\|^2 \cdots \|H_{r_s}(u)\|^2 < \infty.$$

(Actually it is known that $\|H_k(u)\|^2 = k!$, but here we do not need this knowledge.)

The above relations yield the desired representation of a random variable ζ with finite second moment, if it is measurable with respect to the σ-algebra generated by the random variables in $\mathcal{G}_1$. Indeed, the identity $\zeta = \sum_{k=0}^{\infty} \zeta_k$ holds with

$$\zeta_k = \sum_{r_1 + \dots + r_s = k} c(j_1, r_1, \dots, j_s, r_s) H_{r_1}(\eta_{j_1}) \cdots H_{r_s}(\eta_{j_s}),$$

and $\zeta_k \in \mathcal{G}_k$ by Itô's formula.

To complete the proof it is enough to remark that the σ-algebra generated by the random variables $\eta_1, \eta_2, \dots$ and $\mu_W(A)$, $A \in \mathcal{X}$, $\mu(A) < \infty$ agree, as it was stated in part f) of Theorem 10.1.

The results about Wiener–Itô integrals discussed in this Section are useful in the study of non-linear functionals of a set of jointly Gaussian random variables defined by means of a white noise. In my Lecture Note [28] similar problems were discussed, but in that work a slightly different version of Wiener–Itô integrals was introduced. The reason for it was that that the solution of the problems studied in [28] demanded different methods.

In work [28] stationary Gaussian random fields were considered, and the main problem studied there was the description of the limit distribution of certain sequences of non-linear functionals of such Gaussian random fields. In a stationary Gaussian random field a shift operator can be introduced. The shift of all random variables measurable with respect to the underlying stationary Gaussian random field can be defined. In [28] we needed a technique which helps in working with the shift operator.The Fourier analysis is a useful tool in the study of the shift operator. In paper [28] it was tried to unify the tools of multiple Wiener–Itô integrals and Fourier analysis. This led to the definition of a slightly different version of Wiener–Itô integrals.

The idea behind this definition was the observation that not only the correlation function of a stationary Gaussian field can be expressed by means of the Fourier transform of its spectral measure, but also a random spectral measure can be constructed whose Fourier transform expresses the stationary Gaussian process itself. After the introduction of this random spectral measure a version of the multiple Wiener–Itô integral can be defined with respect to it, and all square integrable random variables measurable

with respect to the σ-algebra generated by the underlying Gaussian stationary random field can be expressed with its help. Moreover, it enables us to apply the methods of multiple Wiener–Itô integrals and Fourier analysis simultaneously. In [28] such a method was worked out. The modified Wiener–Itô integral introduced there shows a behaviour similar to that of the original Wiener–Itô integral, only it has to be taken into account that the random spectral measure behaves not like a white noise, but as its 'Fourier transform'. I omit the details which can be found in [28].

The spaces $\mathcal{G}_k$ consisting of all k-fold Wiener–Itô integrals were introduced also in [28], and this had for a special reason. In that work the Hilbert space of square integrable functions, measurable with respect to the underlying stationary Gaussian field was studied together with the shift operator acting on it, which are unitary operators on this Hilbert space. It was useful to decompose this Hilbert space to the direct sum of orthogonal subspaces, invariant with respect to the shift operator. The spaces $\mathcal{G}_k$ were elements of such a decomposition.

In the present work no shift operator was defined, and no limit theorem was studied for non-linear functionals of a Gaussian field. Here the introduction of the spaces $\mathcal{G}_k$ was useful because of a different reason. In the study of our problems we shall need good estimates on the $2p$-th moment of random variables, measurable with respect to the underlying white noise for all positive integers p. As it will be shown, the high moments of the random variables in the spaces $\mathcal{G}_k$ with different indices k show an essentially different behaviour. For a large number p the p-th moment of a random variable in $\mathcal{G}_k$ behaves similarly to that of the k-th power ξ^k of a Gaussian random variable ξ with zero expectation. An estimate of this type will be formulated in Proposition 13.1 or in its consequence, in formula (13.2) and in a partial converse of this result, in Theorem 13.6.

11. The diagram formula for products of degenerate U-statistics.

There is a natural analog of the diagram formula for the products of Wiener–Itô integrals both for the products of multiple integrals with respect to normalized empirical measures and for the products of degenerate U-statistics. These two results are closely related. They express the products of multiple random integrals or of degenerate U-statistics as a sum of multiple random integrals or degenerate U-statistics respectively. In this work the diagram formula for multiple integrals with respect to a normalized empirical measure will be discussed only at an informal level, while a complete proof of the analogous result about degenerate U-statistics will be given. The reason for such an approach is that the diagram formula for the product of degenerate U-statistics is more useful in the study of the problems discussed in this work.

We want to get good estimates about the high moments both of multiple random integrals and of degenerate U-statistics. In the case of degenerate U-statistics the diagram formula yields an explicit formula for these moments. It expresses the product whose expected value has to be calculated as a sum of degenerate U-statistics of different order. Beside this the expected value of all degenerated U-statistics of order $k \geq 1$ equals zero. Hence the expected value we are interested in equals the sum of the zero order terms appearing in the diagram formula.

The analogous problem about the moments of multiple integrals with respect to a normalized empirical measure is more difficult. The diagram formula enables us to express these moments as the sum of the expectation of multiple random integrals of different order also in this case. But the expected value of random integrals of order $k \geq 1$ with respect to a normalized empirical distribution may be non-zero. It is shown in an example presented before the proof of Theorem 9.4 that this is really possible.

First I give an informal description of the diagram formula for the product of two random integrals with respect to a normalized empirical measure. Its analog, the diagram formula for the product of two Wiener–Itô integrals can be described in an informal way by means of formulas (10.13) and (10.13a) together with the 'identity' $\mu_W(\,dx)^2 = \mu(\,dx)$ in their interpretation. The diagram formula for the product of two multiple integrals with respect to a normalized empirical measure has a similar representation. (Observe that in the definition of the random integral $J_{n,k}(\cdot)$ given in formula (4.8) the diagonals are omitted, similarly to the case of Wiener–Itô integrals, from the domain of integration.) In this case such a version of formulas (10.13) and (10.13a) can be applied, where the random integrals $Z_{\mu,k}$ are replaced by $J_{n,k}$, and the white noise measures μ_W are replaced by the normalized empirical measures $\nu_n = \sqrt{n}(\mu_n - \mu)$. But the analog of the 'identity' $\mu_W(\,dx)^2 = \mu(\,dx)$ needed in the interpretation of these formulas has a different form. Namely, it states that $\nu_n(\,dx)^2 = \mu(\,dx) + \frac{1}{\sqrt{n}}\nu_n(\,dx)$. Let us 'prove' this new 'identity'.

Take a small set Δ, i.e. a set Δ such that $\mu(\Delta)$ is very small, write down the identity $\nu_n(\Delta)^2 = n\mu_n(\Delta)^2 + n\mu(\Delta)^2 - 2n\mu_n(\Delta)\mu(\Delta)$, and observe that only a second order error is committed if the terms $n\mu(\Delta)^2$ and $2n\mu_n(\Delta)\mu(\Delta)$ are omitted at the right-hand side of this identity. Moreover, also a second order error is committed if $n\mu_n(\Delta)^2$ is replaced by $\mu_n(\Delta)$, because it has second order small probability that there are at least two sample points in the small set Δ. On the other hand, $n\mu_n(\Delta)^2 = \mu_n(\Delta)$ if Δ contains only zero or one sample point. The above considerations suggest that $\nu_n(\,dx)^2 = \mu_n(\,dx) = \mu(\,dx) + \frac{1}{\sqrt{n}}[\sqrt{n}(\mu_n(\,dx) - \mu(\,dx))] = \mu(\,dx) + \frac{1}{\sqrt{n}}\nu_n(\,dx)$. (This means that in the 'identity' expressing the square $\nu_n(\,dx)^2$ of a normalized empirical measure a correcting term $\frac{1}{\sqrt{n}}\nu_n(\,dx)$ appears. If the sample size $n \to \infty$, then the normalized empirical measure tends to a white noise with counting measure μ, and this correcting term disappears.)

In paper [31] the diagram formula for the product of two multiple integrals with respect to a normalized empirical measure is proved with a different notation. It says that the identity suggested by the above heuristic argument really holds. This result may also help in the proof of the diagram formula for degenerate U-statistics. But a direct proof of this result seems to be simpler.

In the proof of the diagram formula for the product of two degenerate U-statistics first we write this product as the sum of U-statistics. Then by applying Hoeffding's decomposition for each term in this sum the product of two degenerate U-statistics can also be written as a sum of degenerate U-statistics. Actually we apply a slightly refined version of the Hoeffding decomposition where we exploit that we took the product of two *degenerated* U-statistics. Such a calculation yields the diagram formula for the

product of two degenerate U-statistics. With the help of a good notation and some additional work also the product of several degenerate U-statistics can be written as the sum of appropriate degenerate U-statistics. In such a way we get the general form of the diagram formula for the product of degenerate U-statistics.

In this section I formulate the diagram formula for the product of two and finitely many degenerate U-statistics together with an estimate about the L_2-norm of the kernel functions of the degenerate U-statistics appearing in the diagram formula, and a formula about the expectation of products of degenerate U-statistics. To formulate these results some new notations have to be introduced. The proofs of the results in this section are postponed to the next section.

In the formulation of the diagram formula for the product of degenerate U-statistics a more general class of diagrams have to be considered than in the case of multiple Wiener–Itô integrals. We shall define these new diagrams under the name coloured diagrams, and the kernel functions of the U-statistics appearing in the diagram formula will be introduced with their help.

A class of coloured diagrams $\Gamma(k_1, \dots, k_m)$ will be defined whose vertices will be the pairs (p, r), $1 \le p \le m$, $1 \le r \le k_p$, and the set of vertices (p, r), $1 \le r \le k_p$, with a fixed number p will be called the p-th row of the diagram. To define the coloured diagrams of the class $\Gamma(k_1, \dots, k_m)$ first the notions of chain and coloured chain will be introduced. A sequence $\beta = \{(p_1, r_1), \dots, (p_s, r_s)\}$ with $1 \le p_1 < p_2 < \cdots < p_s \le m$ and $1 \le r_u \le k_{p_u}$ for all $1 \le u \le s$ will be called a chain. The number s of the pairs (p_u, r_u) in this sequence, denoted by $\ell(\beta)$, will be called the length of the chain β. Chains of length $\ell(\beta) = 1$, i.e. chains consisting only of one element (p_1, r_1) are also allowed. We shall define a function $c(\beta) = \pm 1$ which will be called the colour of the chain β, and the pair $(\beta, c(\beta))$ will be called a coloured chain. We shall allow arbitrary colouring $c(\beta) = \pm 1$ of a chain with the only restriction that a chain of length 1 can only get the colour -1, i.e. $c(\beta) = -1$ if $\ell(\beta) = 1$.

A coloured diagram $\gamma \in \Gamma(k_1, \dots, k_m)$, $\gamma = \{\beta(l_1), \dots, \beta(l_s)\}$ is a partition of the set $\{(p, r)\colon 1 \le p \le m, 1 \le r \le k_p\}$ to the union of some coloured chains $\beta(l_1), \dots, \beta(l_s)$, i.e. each vertex (p, r) is the element of exactly one chain $\beta(l_j) \in \gamma$. Beside this, each chain $\beta(l_j)$ of a diagram γ has a colour $c_\gamma(\beta(l_j)) = \pm 1$. The set $\Gamma(k_1, \dots, k_m)$ consists of all partitions of the set of vertices $\{(p, r), 1 \le p \le m, 1 \le r \le k_p\}$ to coloured chains, where an arbitray colouring of the chains with the numbers ± 1 is allowed with the only restriction that for a chain $\beta \in \gamma$ of length $\ell(\beta) = 1$ of a diagram $\gamma \in \Gamma(k_1, \dots, k_m)$ $c_\gamma(\beta) = -1$. Here we introduced an indexation of the chains $\beta \in \gamma$ of a diagram γ with some integers $1 \le l_1 < l_2 < \cdots < l_s$. It turned out useful to fix such an indexation (depending on the diagram γ of these chains, and to define the objects we need in our later considerations with its help. It was also useful to allow more general indexation with numbers $l_1, \dots, l_s$ and not with the numbers $1, \dots, s$.

We shall also introduce an enumeration of the vertices of a coloured diagram $\gamma \in \Gamma(k_1, \dots, k_m)$ with the help of the enumaration of its chains. Given a coloured diagram $\gamma = (\beta(l_1), \dots, \beta(l_s)) \in \Gamma(k_1, \dots, k_m)$ we define the indices $\alpha_\gamma(p, r)$ of its vertices in the following way. Put $\alpha_\gamma(p, r) = l_j$ to a vertex (p, r) if $(p, r) \in \beta(l_j)$. We shall split the set of indices $\{l_1, \dots, l_s\}$ of the chains contained in a coloured diagram γ into two disjoint

sets $O(\gamma) = \{l_j\colon\ 1 \le j \le s,\ c_\gamma(\beta(l_j)) = -1\}$, called the set of open indices of the diagram γ and $C(\gamma) = \{l_j\colon\ 1 \le j \le s,\ c_\gamma(\beta(l_j)) = 1\}$, called the set of closed indices of the diagram γ. We shall also list the elements of $O(\gamma)$ in an increasing order, i.e. write $O(\gamma) = \{\bar{l}_1, \dots, \bar{l}_{|O(\gamma)|}\}$, $\bar{l}_1 < \bar{l}_2 < \dots < \bar{l}_{|O(\gamma)|}$. (We shall denote the cardinality of a finite set A by $|A|$ in the sequel.) We defined the coloured diagrams and introduced their open and closed indices, because, as we shall see, in the diagram formula such degenerate U-statistics appear which are defined with the help of these coloured diagrams, and the indices of the arguments of the U-statistic corresponding to the coloured diagram γ are closely related to the chains of γ with colour -1, hence to the open indices of γ.

In the diagram formula we express the product $\prod_{p=1}^{m} I_{n,k_p}(f_p)$ of degenerate U-statistics with canonical kernel functions f_p of k_p variables as the sum of appropriate degenerate U-statistics. The kernel functions of the degenerate U-statistics appering in this representation of the product of degenerate U-statistics will depend on the above defined coloured diagrams γ, and they will be denoted by F_γ, $\gamma \in \Gamma(k_1, \dots, k_m)$. In the definition of these functions F_γ we shall apply the operators introduced below.

Given a function $h(x_{u_1}, \dots, x_{u_r})$ with coordinates in the space $(X, \mathcal{X})$ (the indices $u_1, \dots, u_r$ are all different, otherwise they can be chosen in an arbitrary way) and a probability measure μ on the space $(X, \mathcal{X})$ let us introduce its transforms $P_{u_j}h$ and $Q_{u_j}h$, $1 \le j \le r$, by the formulas

$$(P_{u_j}h)(x_{u_l}\colon\ u_l \in \{u_1, \dots, u_r\} \setminus \{u_j\}) = \int h(x_{u_1}, \dots, x_{u_r})\mu(\,dx_{u_j}), \quad 1 \le j \le r, \tag{11.1}$$

and

$$(Q_{u_j}h)(x_{u_1}, \dots, x_{u_r}) = h(x_{u_1}, \dots, x_{u_r}) - \int h(x_{u_1}, \dots, x_{u_r})\mu(\,dx_{u_j}), \quad 1 \le j \le r. \tag{11.2}$$

(These formulas are very similar to the definition of the operators P_j and Q_j introduced in formula (9.1) before the proof of the Hoeffding decomposition.)

First we consider the product of two degenerate U-statistics, i.e. the case $m = 2$. Let us have a measurable space $(X, \mathcal{X})$ with a probability measure μ on it together with two measurable functions $f_1(x_1, \dots, x_{k_1})$ and $f_2(x_1, \dots, x_{k_2})$ of k_1 and k_2 variables on this space which are canonical with respect to the measure μ. Let $\xi_1, \xi_2, \dots$ be a sequence of $(X, \mathcal{X})$ valued, independent and identically distributed random variables with distribution μ. We want to express the product $I_{n,k_1}(f_1)I_{n,k_2}(f_2)$ of degenerate U-statistics defined with the help of the above random variables and kernel functions f_1 and f_2 as a sum of degenerate U-statistics. For this goal we introduce some notations.

Given two functions $f_1(x_1, \dots, x_{k_1})$ and $f_2(x_1, \dots, x_{k_2})$ and a coloured diagrams $\gamma \in \Gamma(k_1, k_2)$ consisting of s coloured chains $\beta(l_1), \dots, \beta(l_s)$ we define the function

$$\overline{(f_1 \circ f_2)}_\gamma(x_{l_1}, \dots, x_{l_s}) = f_1(x_{\alpha_\gamma(1,1)}, \dots, x_{\alpha_\gamma(1,k_1)})f_2(x_{\alpha_\gamma(2,1)}, \dots, x_{\alpha_\gamma(2,k_2)}), \tag{11.3}$$

where $\alpha_\gamma(p, r)$ denotes the index of the vertex (p, r) of the diagram γ in their above defined enumeration α_γ. (In formula (11.3) all arguments of the functions f_1 and f_2

have different indices. But the indices $\alpha_\gamma(1,j)$ and $\alpha_\gamma(2,j')$ may agree for some pairs (j,j'). This happens if the vertices $(1,j)$ and $(2,j')$ belong to the same chain $\beta \in \gamma$ of length 2.) Let us also define the function

$$(f_1 \circ f_2)_\gamma(x_{l_p},\ l_p \in O(\gamma)) = \left(\prod_{p \in C(\gamma)} P_p \prod_{p \in O_2(\gamma)} Q_p \right) \overline{(f_1 \circ f_2)}_\gamma(x_{l_1}, \dots, x_{l_s}), \quad (11.4)$$

with the operators P_p and Q_p defined (with a different indexation) in formulas (11.1) and (11.2), where $C(\gamma)$ is the set of indices of the closed diagrams of γ, and $O_2(\gamma) \subset O(\gamma)$, defined as $O_2(\gamma) = \{l\colon\ c_\gamma(\beta_l) = -1,$ and $\ell(\beta(l) = 2\}$, is the set of indices of the chains of γ with colour -1 and length 2. are the above defined sets of open and closed indices of the diagram γ. The function $(f_1 \circ f_2)_\gamma$ depends only on the arguments indexed by open vertices of the diagram γ. Let us also remark that the operators P_p and Q_p in formula (11.4) are exchangeable, hence it is not important in what order we apply them.

The function $F_\gamma(f_1, f_2)$ we apply in the formulation of the diagram formula in the special case when the product of two degenerate U-statistics is considered is similar to the function $(f_1 \circ f_2)_\gamma$. We need a small technical step for its definition. We want to work with such a function whose variables are indexed with the numbers $1, 2, \dots, |O(\gamma)|$ while the indices of the function $f_1 \circ f_2)_\gamma$ are the elements of the set $O(\gamma) = \{\bar{l}_1, \dots, \bar{l}_{|O(\gamma)|}\}$. Hence we define the function $t = t_\gamma$ on the set $O(\gamma)$ defined by the formula $t(\bar{l}_j) = j$, $1 \le j \le |O(\gamma)|$, and introduce the function

$$F_\gamma(f_1, f_2)(x_1, x_2, \dots, x_{|O(\gamma)|}) = (f_1 \circ f_2)_\gamma(x_{t(l_p)},\ l_p \in O(\gamma)) \quad (11.5)$$

Next we formulate the diagram formula for the product of two degenerate U-statistics with the help of the above defined quantities.

Theorem 11.1. (The diagram formula for the product of two degenerate U-statistics). *Let a sequence of independent and identically distributed random variables $\xi_1, \xi_2, \dots$ be given with some distribution μ on a measurable space $(X, \mathcal{X})$ together with two bounded canonical functions $f_1(x_1, \dots, x_{k_1})$ and $f_2(x_1, \dots, x_{k_2})$ with respect to the probability measure μ on the product spaces $(X^{k_1}, \mathcal{X}^{k_1})$ and $(X^{k_2}, \mathcal{X}^{k_2})$ respectively. Let us take the class of coloured diagrams $\Gamma(k_1, k_2)$ introduced above together with the functions $F_\gamma(f_1, f_2)$ defined in formulas (11.1)—(11.5).*

For all $\gamma \in \Gamma$ $F_\gamma(f_1, f_2)_\gamma$ is a canonical function with respect to the measure μ with $|O(\gamma)|$ arguments, where $O(\gamma)$ and $C(\gamma)$ denote the set of open and closed indices of the diagram γ. The product of the degenerate U-statistics $I_{n,k_1}(f_1)$ and $I_{n,k_2}(f_2)$, $n \ge \max(k_1, k_2)$, defined in (8.7) can be expressed as

$$\begin{aligned}(n^{-k_1/2} k_1! I_{n,k_1}(f_1))(n^{-k_2/2} k_2! I_{n,k_2}(f_2)) \\ = \sum_{\gamma \in \Gamma(k_1,k_2)}{}' \prod_{j=1}^{|C(\gamma)|} \left(\frac{n - s(\gamma) + j}{n} \right) n^{-W(\gamma)/2} \cdot n^{-|O(\gamma)|/2} |O(\gamma)|! I_{n,|O(\gamma)|}(F_\gamma(f_1, f_2))\end{aligned} \quad (11.6)$$

with $W(\gamma) = k_1 + k_2 - |O(\gamma)| - 2|C(\gamma)|$ *and* $s(\gamma) = |O(\gamma)| + |C(\gamma)|$ *(which equals the number of coloured diagrams in* γ*), where* $\sum^{\prime(n)}$ *means that summation is taken only for such coloured diagrams* $\gamma \in \Gamma(k_1, k_2)$ *which satisfy the inequality* $s(\gamma) \le n$*, and* $\prod_{j=1}^{|C(\gamma)|}$ *equals 1 in the case* $|C(\gamma)| = 0$*. The term* $I_{n,|O(\gamma)|}(F_\gamma(f_1, f_2))$ *can be replaced by* $I_{n,|O(\gamma)|}(\mathrm{Sym} F_\gamma(f_1, f_2))$ *in formula (11.6).*

Consider the L_2*-norm of the functions* $F_\gamma(f_1, f_2)$ *defined by the formula*

$$\|F_\gamma(f_1, f_2)\|_2^2 = \|(f_1 \circ f_2)_\gamma\|_2^2 = \int (f_1 \circ f_2)_\gamma^2(x_{l_p},\, l_p \in O(\gamma)) \prod_{l_p \in O(\gamma)} \mu(\, dx_{l_p}).$$

The inequality

$$\|F_\gamma(f_1, f_2)\|_2 = \|(f_1 \circ f_2)_\gamma\|_2 \le \|f_1\|_2 \|f_2\|_2 \quad \textit{if } W(\gamma) = 0 \tag{11.7}$$

holds for this norm. The condition $W(\gamma) = 0$ *in formula (11.7) means that the diagram* $\gamma \in \Gamma(k_1, k_2)$ *has no chains* β *of length* $\ell(\beta) = 2$ *with colour* $c_\gamma(\beta) = -1$*. In the case of a general diagram* $\gamma \in \Gamma(k_1, k_2)$ *the inequality*

$$\|F_\gamma(f_1, f_2)\|_2 = \|(f_1 \circ f_2)_\gamma\|_2 \le 2^{W(\gamma)} \min(\|f_1\|_2, \|f_2\|_2) \tag{11.8}$$

holds if the L_∞*-norm of the functions* f_1 *and* f_2 *satisfies the inequalities* $\|f_1\|_\infty \le 1$ *and* $\|f_2\|_\infty \le 1$*. Relations (11.7) and (11.8) also hold for non-canonical functions* f_1 *and* f_2*.*

Inequality (11.7) is actually a repetition of estimate (10.11) about the diagrams appearing in the case of Wiener–Itô integrals. Inequality (11.8) yields a weaker bound about the L_2-norm $\|F_\gamma(f_1, f_2)\|_2 = \|(f_1 \circ f_2)_\gamma\|_2$ for a general diagram γ. In particular, it depends not only on the L_2-norm, but also on the L_∞-norm of the functions f_1 and f_2. This is closely related to the fact that in the estimates on the distribution of U-statistics, — unlike the case of Wiener–Itô integrals, — a condition is imposed not only on the L_2-norm of the kernel function f, but also on its L_∞-norm. I return to this question later.

Remark 1. The expression $W(\gamma) = k_1 + k_2 - |O(\gamma)| - 2|C(\gamma)|$ appearing in formulas (11.6), (11.7) and (11.8) has the following content. It equals the number of those diagrams $\beta(l_j) \in \gamma$ for which $\ell(\beta(l_j)) = 2$, and $c_\gamma(\beta(l_j)) = -1$. Indeed, if $W(\gamma)$ denotes the number of such chains, and $\bar{W}(\gamma)$ equals the number of chains $\beta(l_j) \in \gamma$ for which $\ell(\beta(l_j)) = 1$ (and as a consequence $c_\gamma(\beta(l_j)) = -1$), then $W(\gamma) + \bar{W}(\gamma) = |O(\gamma)|$, and $2W(\gamma) + \bar{W}(\gamma) + 2|C(\gamma)| = k_1 + k_2$. These identities imply the statement of this remark.

Remark 2. The term $I_{n,|O(\gamma)|}(F_\gamma(f_1, f_2))$ appeared in the sum at the right-hand side of (11.6) only if the condition $s(\gamma) \le n$ was satisfied. This restriction in the summation had a technical character, which has no great importance in our investigations. It is related to the fact that a U-statistic $I_{n,k}(f)$ exists only if $n \ge k$. As a consequence,

some U-statistics disappear at the right-hand side of (11.6) if the sample size n of the U-statistics is relatively small. The term $I_{n,|O(\gamma)|}(F_\gamma(f_1, f_2))$ appeared in (11.6) through the Hoeffding decomposition of a U-statistic with kernel function $\overline{(f_1 \circ f_2)}_\gamma$ defined in (11.3). This function has $s(\gamma)$ arguments, and the U-statistic corresponding to it appears in our calculations only if the sample size n is not smaller than this number.

Let us recall the convention introduced after the definition of canonical degenerate U-statistics by which $I_{n,0}(c)$ is a degenerate U-statistic of order zero, and $I_{n,0}(c) = c$ for a constant c. By applying this convention we write $F_\gamma((f_1, f_2) = f_1 \circ f_2$ in relation (11.6) for those diagrams γ for which $|O(\gamma)| = 0$, i.e. $c_\gamma(\beta) = 1$ for all chains $\beta \in \gamma$. We shall introduce another convention which implies that Theorem 11.1 is valid also in the degenerate case when the function $f_{k_1} = c$ with a constant c, and $k_1 = 0$. In this case $\Gamma(k_1, k_2)$ consists of only one diagram γ containing the chains $\beta_j = \{j\}$ of length one and colour $c_\gamma(\{j\}) = -1$, $1 \le j \le k_2$. We define $I(F_\gamma(f_1, f_2)) = cf_2$ in this case. Beside this, we have $W(\gamma) = k_1 + k_2 - |O(\gamma)| - 2|C(\gamma)| = 0$, $|O(\gamma)| = k_2$, and $|C(\gamma)| = 0$. Hence formula (11.6) remains valid also in the case $k_1 = 0$. We have introduced this convention because the following inductive argument leading to the proof of the diagram formula for the product of degenerate U-statistics in the general case is valid under such a convention.

Let us turn to the formulation of the general form of the diagram formula for the product of degenerate U-statistics. First I define a function $F_\gamma = F_\gamma(f_1, \dots, f_m)$ for each coloured diagram $\gamma \in \Gamma(k_1, \dots, k_m)$ and collection of canonical functions (with respect to a probability measure μ on a measurable space $(X, \mathcal{X})$) $f_1, \dots, f_m$ with $k_1, \dots,$ and k_m variables. These functions F_γ will be the kernel functions of the degenerate U-statistics at the right-hand side of the diagram formula.

These functions F_γ will be defined by induction with respect to the number m of the components in the product. For $m = 2$ we have already defined the function $F_\gamma(f_1, f_2)$. Let the functions $F_\gamma(f_1, \dots, f_{m-1})$ be defined for each coloured diagram $\gamma \in \Gamma(k_1, \dots, k_{m-1})$. To define $F_\gamma(f_1, \dots, f_m)$ for a coloured diagram $\gamma \in \Gamma(k_1, \dots, k_m)$ first we define the predecessor $\gamma_{pr} = \gamma_{pr}(\gamma) \in \Gamma(k_1, \dots, k_{m-1})$ of γ. We shall define the coloured diagram γ_{pr} together with an appropriate indexation of its element with the help of the enumeration of the elements of γ. Roughly speaking, the elements of γ_{pr} are the the restrictions of the chains contained in γ to the first $m-1$ rows of the diagram, i.e. to the set $\{(p, r)\colon 1 \le p \le m-1, 1 \le r \le k_p\}$.

To define precisely the predecessor γ_{pr} of γ let us divide first the chains of the coloured diagram $\gamma = \{\beta(l_1), \dots, \beta(l_s)\} \in \Gamma(k_1, \dots, k_m)$ into two disjoint subsets $\gamma = \gamma_1 \cup \gamma_2$, defined as $\gamma_1 = \{\beta(l_j)\colon \beta(l_j) \in \gamma, \beta(l_j) \cap \{(m,1), \dots, (m,k_m)\} \neq \emptyset\}$ and $\gamma_2 = \{\beta(l_j)\colon \beta(l_j) \in \gamma, \beta(l_j) \cap \{(m,1), \dots, (m,k_m)\} = \emptyset\}$, i.e. a coloured chain $\beta \in \gamma$ belongs to γ_1 if it contains a vertex from the last row $\{(m,1), \dots, (m,k_m)\}$ of the diagram, and it belongs to γ_2 if it does not contain such a vertex. We define with the help of the chains $\beta(l_j) \in \gamma_1$ the chains $\beta_{pr}(l_j) = \beta(l_j) \setminus \{(m,1), \dots, (m,k_m)\}$ and with the help of the chains $\beta(l_j) \in \gamma_2$ the chains $\beta_{pr}(l_j) = \beta(l_j)$. (For those chains $\beta(l_j) \in \gamma_1$ which consist only of one vertex of the form (m, r), $1 \le r \le k_m$, the corresponding chain $\beta_{pr}(l_j)$ would be the empty set. These empty sets are omitted from the set of chains

$\beta_{pr}(l_j) \in \gamma_{pr}$.) The set of all above defined chains $\beta_{pr}(l_j)$ provides a partition of the set of vertices $\{(p,r)\colon\ 1 \le p \le m-1,\ 1 \le r \le k_p\}$. The diagram γ_{pr} will consist of these chains $\beta_{pr}(l_j)$. To complete the definition of the coloured diagram γ_{pr} we still have to define the colour $c_{\gamma_{pr}}(\beta_{pr}(l_j))$ of these chains.

We define the colour of these chains by the formulas $c_{\gamma_{pr}}(\beta_{pr}(l_j)) = -1$ if $\beta(l_j) \in \gamma_1$, and $c_{\gamma_{pr}}(\beta_{pr}(l_j)) = c_\gamma(\beta(l_j))$ if $\beta(l_j) \in \gamma_2$. In such a way we defined the predecessor $\gamma_{pr} \in \Gamma(k_1, \dots, k_{m-1})$ of the diagram $\gamma \in \Gamma(k_1, \dots, k_m)$. Moreover we gave an indexation of the chains of γ_{pr} with the help of the indexation of the chains of γ.

With the help of the coloured diagram $\gamma_{pr} \in \Gamma(k_1, \dots, k_{m-1})$ we can define the function $F_{\gamma_{pr}} = F_{\gamma_{pr}}(f_1, \dots, f_{m-1})$ which is a function of $|O(\gamma_{pr})|$ variables $x_1, \dots, x_{|O(\gamma_{pr})|}$. We shall define the function $F_\gamma = F_\gamma(f_1, \dots, f_m)$ similarly to the definition of $F_\gamma(f_1, f_2)$ given by formulas (11.3), (11.4) and (11.5) in the case $m = 2$. In this case $F_{\gamma_{pr}}$ plays the role of the function f_1 and f_m the role of the function f_2. To define the function $F_\gamma(f_1, \dots, f_m)$ we still have to define a coloured diagram $\gamma_{cl} = \gamma_{cl}(\gamma) \in \Gamma(|O(\gamma_{pr})|, k_m)$ that we shall call the closing diagram of γ. The heuristic content of the diagram γ_{cl} is that it contains the additional information we need to reconstruct the diagram $\gamma \in \Gamma(k_1, \dots, k_m)$ if we know its predecessor γ_{pr}. We shall define it together with an enumeration of its chains that depends on the enumeration of the chains of the diagram γ.

To define the diagram γ_{cl} let us first consider the listing $O(\gamma_{pr}) = \{\bar{l}_1, \dots, \bar{l}_{|O(\gamma_{pr})|}\}$, $1 \le \bar{l}_1 < \bar{l}_2 < \dots < l_{|O(\gamma_{pr}|}$, of the indices of the open indices of the diagram γ_{pr} in increasing order. Let us fix a vertex $(1, j)$, $1 \le j \le |O(\gamma_{pr})|$ in the first row of γ_{cl}. We shall denote the chain of γ_{cl} containing this vertex by $\beta_{cl}(\bar{l}_j)$, and define it together with its colour in the following way. Let us consider the (open) chain $\beta_{pr}(\bar{l}_j)$ together with its 'continuation' $\beta(\bar{l}_j)$. Clearly, $\beta_{pr}(\bar{l}_j) \subset \beta(\bar{l}_j)$. If $\beta(\bar{l}_j) \in \gamma_1$, then $\beta(\bar{l}_j) = \beta_{pr}(\bar{l}_j) \cup \{(m, r_j)\}$ with some integer $1 \le r_j \le k_m$. In this case we define the chain containing the vertex $(1, j)$ as the diagram $\beta_{cl}(\bar{l}_j) = \{(1, j), (2, r_j)\}$ with this number r_j, and it gets the colour $c_{\gamma_{cl}}(\beta_{cl}(\bar{l}_j)) = c_\gamma(\beta(\bar{l}_j))$. If $\beta(\bar{l}_j) \in \gamma_2$, then $\beta_{pr}(\bar{l}_j) = \beta(\bar{l}_j)$, and we define the chain containing the vertex $(1, j)$ as the chain $\beta_{cl}(\bar{l}_j) = \{(1, j)\}$ of length 1 and with colour $c_{\gamma_{cl}}(\beta_{cl}(\bar{l}_j)) = -1$.

We still have to consider those vertices $(2, r)$ of $\Gamma(|O(\gamma_{pr})|, k_m)$, $1 \le r \le k_m$, for which there exists a chain $\beta(l_{j(r)}) \in \gamma$ such that $\beta(l_{j(r)}) = \{m, r)\}$, because these are the vertices of the of the set of vertices $\{(1, j)\colon\ 1 \le j \le |O(\gamma_{pr})| \cup \{(2, r)\colon\ 1 \le r \le k_m\}$ which are not contained in the previously defined chains $\beta_{cl}(\bar{l}_j)$. To cover these vertices with an (appropriately indexed) chain of γ_{cl} let us define the chains $\beta_{cl}(l_{j(r)}) = \{(2, r)\}$ with the colour $c_{\gamma_{cl}}(\beta_{cl}(l_{j(r)})) = -1$ for such vertices $(2, r)$. The above defined coloured chains provide a partition of the set $\{(1, j)\colon\ 1 \le j \le |O(\gamma_{pr})| \cup \{(2, r)\colon\ 1 \le r \le k_m\}$, and they are the elements of the coloured diagram γ_{cl}.

We shall define the function $F_\gamma(f_1, \dots, f_m)$ with the help of the above introduced diagrams γ_{pr} and γ_{cl} in the following way. Put, similarly to formula (11.3),

$$\begin{aligned}&\overline{(F_{\gamma_{pr}}(f_1, \dots, f_{m-1}) \circ f_m)}_\gamma(x_{l_1}, \dots, x_{l_s})\\ &\quad = F_{\gamma_{cl}}(x_{\alpha_{\gamma_{cl}}(1,1)}, \dots, x_{\alpha_{\gamma_{cl}}(1,|O(\gamma_{pr})|)}) f_m(x_{\alpha_{\gamma_{cl}}(2,1)}, \dots, x_{\alpha_{\gamma_{pr}}(2,k_m)}),\end{aligned} \tag{11.9}$$

where $s = s(\gamma_{cl})$ is the number of the chains contained in γ_{cl}. The indices $l_1, l_2 \dots$, and l_s of the variables at the left-hand side of (11.9) agree with the indices of the chains of the diagram γ_{cl}, and $\alpha_{\gamma_{cl}}(p, r)$ denotes the index of the vertex (p, r) of the diagram γ_{cl} which is induced by the enumeration of the indices of the chains in γ_{cl}. Next we define with the help of formula (11.9), similarly to the relation (11.4), the function

$$\begin{aligned}&(F_{\gamma_{pr}}(f_1, \dots, f_{m-1}) \circ f_m)_\gamma(x_p,\ p \in O(\gamma_{cl})) \\ &\quad = \left(\prod_{p \in C(\gamma_{cl})} P_p \prod_{p \in O_2(\gamma_{cl})} Q_p \right) \overline{(F_{\gamma_{pr}}(f_1, \dots, f_{m-1} \circ f_m)_\gamma}(x_p,\ p \in O(\gamma_{cl}) \cup C(\gamma_{cl}))\end{aligned} \tag{11.10}$$

with the operators P_p and Q_p defined (with a different indexation) in formulas (11.1) and (11.2), where the sets $O(\gamma_{cl})$ and $C(\gamma_{cl})$ are the sets of open and closed indices of the diagram γ_{cl}, and the set $O_2(\gamma_{cl})$ (for a general diagram with two rows) was defined after formula (11.4). The function $(F_{\gamma_{pr}}(f_1, \dots, f_{m-1}) \circ f_m)_\gamma$ depends only on the arguments indexed by the open indices of the diagram γ_{cl}.

The function $F_\gamma(f_1, \dots, f_m)$ will be defined by means of a reindexation of the arguments of the function $(F_{\gamma_{pr}}(f_1, \dots, f_{m-1}) \circ f_m)_\gamma(x_{l_p},\ l_p \in O(\gamma_{cl}))$ which will be made to get a function with arguments $x_1, x_2, \dots, x_{|O(\gamma_{cl})|}$. It is defined, similarly to formula (11.5), as

$$F_\gamma(f_1, \dots, f_m)(x_1, x_2, \dots, x_{|O(\gamma_{cl})|}) = (F_{\gamma_{pr}}(f_1, \dots, f_{m-1}) \circ f_m)_\gamma(x_{t(l_p)},\ l_p \in O(\gamma_{cl})), \tag{11.11}$$

where the indices $t(l_p)$ are defined in the following way. We list the open indices of the diagram γ_{cl} in an increasing order as $O(\gamma_{cl}) = \{\bar{l}_1, \dots, \bar{l}_{|O(\gamma_{cl})|}\}$, $\bar{l}_1 < \bar{l}_2 < \dots < \bar{l}_{|O(\gamma_{cl})|}$, and define the function $t(\cdot)$ on the set $O(\gamma_{cl})$ as $t(\bar{l}_p) = p$ for $1 \le p \le |O(\gamma_{cl})|$.

To complete the definition of the function $F_\gamma(f_1, \dots, f_m)$ observe that $|O(\gamma_{cl})| = |O(\gamma)|$. (Even the sets $O(\gamma_{cl})$ and $O(\gamma)$ agree with the enumeration of the chains of these two diagrams we have chosen.) Hence we can write

$$F_\gamma(f_1, \dots, f_m)(x_1, x_2, \dots, x_{|O(\gamma_{cl})|}) = F_\gamma(f_1, \dots, f_m)(x_1, x_2, \dots, x_{|O(\gamma)|}). \tag{11.12}$$

To formulate the general form of the diagram formula for the product of degenerate U-statistics we introduce some quantities which will be the version of the quantites appearing in the coefficients of the right-hand side of (11.6) in Theorem 11.1. Put

$$W(\gamma) = \sum_{l_p \in O(\gamma)} (\ell(\beta(l_p)) - 1) + \sum_{l_p \in C(\gamma)} (\ell(\beta(l_p)) - 2), \quad \gamma \in \Gamma(k_1, \dots, k_m), \tag{11.13}$$

where $\ell(\beta)$ denotes the length of the chain β.

To define the next quantity we need let us first introduce the following notation. Given a chain $\beta = \{(p_1, r_1), \dots, (p_l, r_l)\}$, $1 \le p_1 < p_2 < \dots < p_l \le m$, in the set $\{(p, r)\colon 1 \le p \le m,\ 1 \le r \le k_p\}$ let us define its upper level $u(\beta) = p_1$, and its deepest level $d(\beta) = l_p$. Let us define with their help for all diagrams $\gamma \in \Gamma(k_1, \dots, k_m)$ and

integers p, $1 \le p \le m$, the sets $\mathcal{B}_1(\gamma,\beta) = \{\beta\colon\ \beta \in \gamma,\, c_\gamma(\beta) = 1,\, d(\beta) = p\}$, and $\mathcal{B}_2(\gamma,p) = \{\beta\colon\ \beta \in \gamma, c_\gamma(\beta) = -1,\, d(\beta) \le p\} \cup \{\beta\colon\ \beta \in \gamma,\, u(\beta) \le p,\, d(\beta) > p\}$, i.e. $\mathcal{B}_1(\gamma,p)$ consists of those chains $\beta \in \Gamma$ which have colour 1, all their vertices are in the first p rows of the diagram, and contain a vertex in the p-th row, while $\mathcal{B}_2(\gamma,p)$ consists of those chains $\beta \in \gamma$ which have either colour -1, and all their vertices are in the first p rows of the diagram, or they have (with an arbitrary colour) a vertex both in the first p rows both in the remaining rows of the diagram. Put $B_1(\gamma,p) = |\mathcal{B}_1(\gamma,p)|$ and $B_2(\gamma,p) = |\mathcal{B}_2(\gamma,p)|$. With the help of these numbers we define

$$J_n(\gamma,p) = \begin{cases} \prod_{j=1}^{B_1(\gamma,p)} \left(\frac{n - B_1(\gamma,p) - B_2(\gamma,p) + j}{n} \right) & \text{if } B_1(\gamma,p) \ge 1 \\ 1 \quad \text{if } B_1(\gamma,p) = 0 \end{cases} \tag{11.14}$$

for all $2 \le p \le m$ and diagrams $\gamma \in \Gamma(k_1,\dots,k_m)$.

Theorem 11.2 will be formulated with the help of the above notations.

Theorem 11.2 (The diagram formula for the product of several degenerate U-statistics). *Let a sequence of independent and identically distributed random variables $\xi_1, \xi_2, \dots$ be given with some distribution μ on a measurable space $(X, \mathcal{X})$ together with $m \ge 2$ bounded functions $f_p(x_1,\dots,x_{k_p})$ on the spaces $(X^{k_p}, \mathcal{X}^{k_p})$, $1 \le p \le m$, canonical with respect to the probability measure μ. Let us consider the class of coloured diagrams $\Gamma(k_1,\dots,k_m)$ together with the functions $F_\gamma = F_\gamma(f_1,\dots,f_m)$, $\gamma \in \Gamma(k_1,\dots,k_m)$, defined in formulas (11.9)—(11.12) and the constants $W(\gamma)$ and $J_n(\gamma,p)$, $1 \le p \le m$, given in formulas (11.13) and (11.14).*

The functions $F_\gamma(f_1,\dots,f_m)$ are canonical with respect to the measure μ with $|O(\gamma)|$ variables, and the product of the degenerate U-statistics $I_{n,k_p}(f_p)$, $1 \le p \le m$, $n \ge \max\limits_{1\le p\le m} k_p$, defined in (8.7) can be expressed as

$$\begin{aligned} \prod_{p=1}^m n^{-k_p/2} k_p! I_{n,k_p}(f_{k_p}) = {} & \sum_{\gamma\in\Gamma(k_1,\dots,k_m)}{}' \left(\prod_{p=2}^m J_n(\gamma,p) \right) n^{-W(\gamma)/2} \\ & n^{-|O(\gamma)|/2} |O(\gamma)|! I_{n,|O(\gamma)|}(F_\gamma(f_1,\dots,f_m)), \end{aligned} \tag{11.15}$$

where $\sum'^{(n,m)}$ means that summation is taken for those $\gamma \in \Gamma(k_1,\dots,k_m)$ which satisfy the relation $B_1(\gamma,p) + B_2(\gamma,p) \le n$ for all $2 \le p \le m$ with the quantities $B_1(\gamma,p)$ and $B_2(\gamma,p)$ introduced before the definition of $J_n(\gamma,p)$ in (11.14). The terms $I_{n,|O(\gamma)|}(F_\gamma(f_1,\dots,f_m))$ at the right-hand side of formula (11.15) can be replaced by $I_{n,|O(\gamma)|}(\mathrm{Sym}\, F_\gamma(f_1,\dots,f_m))$.

In Theorem 11.2 the product of such degenerate U-statistics were considered, whose kernel functions were bounded. This also implies that all functions F_γ appearing at the right-hand side of (11.15) are well-defined (i.e. the integrals appearing in their definition are convergent) and bounded. In the applications of Theorem 11.2 it is useful to have

more information about the behaviour of the functions F_γ. We shall need some good bound on their L_2-norm. Such a result is formulated in the following

Lemma 11.3. (Estimate about the L_2-norm of the kernel functions of the U-statistics appearing in the diagram formula). *Let m functions $f_p(x_1, \dots, x_{k_p})$ be given on the products $(X^{k_p}, \mathcal{X}^{k_p})$ of some measurable space $(X, \mathcal{X})$, $1 \le p \le m$, with a probability measure μ on it, which satisfy inequalities (8.1) and (8.2) (if the index k is replaced by the index k_p in them), but these functions need not be canonical. Let us take a coloured diagram $\gamma \in \Gamma(k_1, \dots, k_m)$, and consider the function $F_\gamma(f_1, \dots, f_m)$ defined by formulas (11.9)—(11.12). The L_2-norm of the function $F_\gamma(f_1, \dots, f_m)$ (with respect to the power of the measure μ to the space where $F_\gamma(f_1, \dots, f_m)$ is defined) satisfies the inequality*

$$\|F_\gamma(f_1, \dots, f_m)\|_2 \le 2^{W(\gamma)} \prod_{p \in U(\gamma)} \|f_p\|_2,$$

where $W(\gamma)$ is given in (11.13), and the set $U(\gamma) \subset \{1, \dots, m\}$ is defined in the following way. Let us define for a coloured chain $\beta = \{(l_1, r_1), (l_2, r_2), \dots, (l_s, r_s)\} \in \gamma$ with $1 \le l_1 < \dots < l_s \le m$ the set of its interior levels as and $\mathrm{Int}\,(\beta) = \{l_2, \dots, l_{s-1}, l_s\}$ if $c_\gamma(\beta) = -1$ and $\mathrm{Int}\,(\beta) = \{l_2, \dots, l_{s-1}\}$ if $c_\gamma(\beta) = 1$. Then we define $U(\gamma) = \{1, \dots, m\} \setminus \left(\bigcup_{\beta \in \gamma} \mathrm{Int}\,(\beta) \right)$.

The last result of this section is a corollary of Theorem 11.2. In this corollary we give an estimate on the expected value of product of degenerate U-statistics. To formulate this result we introduce the following terminology. Let us call a (coloured) diagram $\gamma \in \Gamma(k_1, \dots, k_m)$ closed if $c_\gamma(\beta) = 1$ for all chains $\beta \in \gamma$. Let us denote the set of all closed diagrams by $\bar{\Gamma}(k_1, \dots, k_m)$. Observe that $F_\gamma(f_1, \dots, f_m)$ is constant (a function of zero variable) for all closed diagram $\gamma \in \bar{\Gamma}(k_1, \dots, k_m)$, and $I_{n,k(\gamma)}(F_\gamma(f_1, \dots, f_m)) = F_\gamma(f_1, \dots, f_m)$ in this case. Now we formulate the following result.

Corollary of Theorem 11.2 about the expectation of a product of degenerate U-statistics. *Let a finite sequence of functions $f_p(x_1, \dots, x_{k_p})$, $1 \le p \le m$, be given on the products $(X^{k_p}, \mathcal{X}^{k_p})$ of some measurable space $(X, \mathcal{X})$ together with a sequence of independent and identically distributed random variables with value in the space $(X, \mathcal{X})$ which satisfy the conditions of Theorem 11.2.*

Let us apply the notation of Theorem 11.2 together with the notion of the above introduced class of closed diagrams $\bar{\Gamma}(k_1, \dots, k_m)$. The identity

$$E\left(\prod_{p=1}^m k_p! n^{-k_p/2} I_{n,k_p}(f_{k_p}) \right) = \sum_{\gamma \in \bar{\Gamma}(k_1, \dots, k_m)}{}' \left(\prod_{p=1}^m J_n(\gamma, p) \right) n^{-W(\gamma)/2} \cdot F_\gamma(f_1, \dots, f_m) \tag{11.16}$$

holds. This identity has the consequence

$$\left| E\left(\prod_{p=1}^m k_p! n^{-k_p/2} I_{n,k_p}(f_{k_p}) \right) \right| \le \sum_{\gamma \in \bar{\Gamma}(k_1, \dots, k_m)} n^{-W(\gamma)/2} |F_\gamma(f_1, \dots, f_m)|. \tag{11.17}$$

Beside this, if $\|f_p\|_2 \le \sigma$ *for all* $1 \le p \le m$, *then the numbers* $F_\gamma(f_1, \dots, f_m)$ *at the right-hand side of (11.17) satisfy the inequality*

$$|F_\gamma(f_1, \dots, f_m)| \le 2^{W(\gamma)} \sigma^{|U(\gamma)|} \quad \textit{for all } \gamma \in \bar{\Gamma}(k_1, \dots, k_m). \tag{11.18}$$

In formula (11.18) the same number $W(\gamma)$ *and set* $U(\gamma)$ *appear as in Lemma 11.3. The only difference is that in the present case* $c_\gamma(\beta) = 1$ *for all chains* $\beta \in \gamma$ *which appear in the definition of* $U(\gamma)$.

Remark: We have applied a different terminology for diagrams in this section and in Section 10, where the the theory of Wiener–Itô integrals was discussed. But there is a simple relation between the terminology of these sections. If we take only those diagrams from the diagrams considered in this section which contain only chains of length 1 or 2, and beside this the chains of length 1 have colour -1, and the chains of length 2 have colour 1, then we get the diagrams considered in the previos section. Moreover, the functions $F_\gamma = F_\gamma(f_1, \dots, f_m)$ are the same in the two cases. Hence formula (10.18) in the Corollary of Theorem 10.2 and formula (11.17) in the Corollary of Theorem 11.2 make possible to compare the moments of Wiener–Itô integrals and degenerate U-statistics.

The main difference between these estimates is that formula (11.17) contains some additional terms. They are the contributions of those diagrams $\gamma \in \bar{\Gamma}(k_1, \dots, k_m)$ which contain chains $\beta \in \gamma$ with length $\ell(\beta) > 2$. These are those diagrams $\gamma \in \bar{\Gamma}(k_1, \dots, k_m)$ for which $W(\gamma) > 1$. The estimate (11.18) given for the terms F_γ corresponding to such diagrams is weaker, than the estimate given for the terms F_γ with $W(\gamma) = 0$, since $|U(\gamma)| < m$ if $W(\gamma) \ge 1$, while $|U(\gamma)| = m$, if $W(\gamma) = 0$. On the other hand, such terms have a coefficient $n^{-W(\gamma)/2}$ at the right-hand side of formula (11.17). A closer study of these formulas may explain the relation between the estimates given for the tail distrtibution of Wiener–Itô integrals and degenerate U-statistics.

12. The proof of the diagram formula for U-statistics.

In this section the results of the previous section will be proved. First I prove its main result, the diagram formula for the product of two degenerate U-statistics.

Proof of Theorem 11.1. In the first step of the proof the product $k_1!I_{n,k_1}(f_1)k_2!I_{n,k_2}(f_2)$ of two degenerate U-statistics will be rewritten as a sum of not necessarily degenerate U-statistics. In this step a term by term multiplication is carried out for the product $k_1!I_{n,k_1}(f_1)k_2!I_{n,k_2}(f_2)$, and the terms of the sum obtained in such a way are put in different classes indexed by the (non-coloured) diagrams with two rows of length k_1 and k_2. This step is very similar to the heuristic argument leading to formulas (10.13) and (10.13a) in our explanation about the diagram formula for Wiener-Itô integrals.

To carry out this step of the proof consider all sets of pairs

$$\{(u_1, u_1'), \dots, (u_r, u_r')\}, \quad 1 \le r \le \min(k_1, k_2),$$

with the following properties: $1 \le u_1 < u_2 < \cdots < u_r \le k_1$, the numbers $u_1', \dots, u_r'$ are different, and $1 \le u_s' \le k_2$, for all $1 \le s \le r$. To a set of pairs $\{(u_1, u_1'), \dots, (u_r, u_r')\}$ with these properties let us correspond the following diagram $\bar\gamma((u_1, u_1'), \dots, (u_r, u_r')) \in \bar\Gamma(k_1, k_2)$, where $\bar\Gamma(k_1, k_2)$ denotes the set of (non-coloured) diagrams with two rows of length k_1 and k_2. The diagram $\bar\gamma((u_1, u_1'), \dots, (u_r, u_r'))$ has two rows, $\{1, \dots, k_1\}$, and $\{2, \dots, k_2\}$, its chains of length 2 are the sets $\{(1, u_s), (2, u_s')\}$, $1 \le s \le r$, and beside this it contains the chains $\{(1, r)\}$, $r \in \{1, \dots, k_1\} \setminus \{u_1, \dots, u_r\}$, and $\{(2, r)\}$, $r \in \{1, \dots, k_2\} \setminus \{u_1', \dots, u_r'\}$ of length 1. All (non-coloured) diagrams $\bar\gamma \in \bar\Gamma(k_1, k_2)$ can be represented in the form $\bar\gamma = \bar\gamma((u_1, u_1'), \dots, (u_r, u_r'))$ with the help of a set of pairs $\{(u_1, u_1'), \dots, (u_r, u_r')\}$, $1 \le r \le \min(k_1, k_2)$, with the above properties in a unique way.

To make the notation in the subsequent discussion simpler we fix, similarly to the case of coloured diagrams, an indexation of the chains of a diagram $\bar\gamma \in \bar\Gamma(k_1, k_2)$, and we define with its help an indexation of the vertices of this diagram $\bar\gamma$, too. Let us take the following natural indexation. Consider the diagram $\bar\gamma = \bar\gamma((u_1, u_1'), \dots, (u_r, u_r')) \in \bar\Gamma(k_1, k_2)$ which has $s(\bar\gamma) = k_1 + k_2 - r$ chains. The chain $\beta \in \bar\gamma$ containing the vertex $(1, j)$ gets the index j, i.e. $(1, j) \in \beta(j)$ for $1 \le j \le k_1$. To define the index of the remaining chains of $\bar\gamma$ which are chains of length 1 of the form $(2, j)$ with $j \in \{1, \dots, k_2\} \setminus \{u_1', \dots, u_r'\}$ let us take the list $\{\bar l_1, \dots, \bar l_{k_2 - r}\}$, $1 \le \bar l_1 < \cdots < \bar l_{k_2 - r}$, of the elements of the set $\{1, \dots, k_2\} \setminus \{u_1', \dots, u_r'\}$ in an increasing order. Then we define the indices of the remaining chains by the formula $\beta(k_2 + j) = \{(2, \bar l_j)\})$, $1 \le j \le k_2 - r$. After this we define the indexation of the vertices of the diagram γ by the formula $\alpha_{\bar\gamma}(p, r) = l$ with that index l for which $(p, r) \in \beta(l)$. Let us also define the sets $V_1 = V_1(\bar\gamma) = \{1, \dots, k_1 + k_2 - r\} \setminus \{u_1, \dots, u_r\}$ and $V_2 = V_2(\bar\gamma) = \{u_1, \dots, u_r\}$, i.e. V_1 is the set of indices of the chains of $\bar\gamma$ of length 1, and V_2 is the set of indices of the chains of $\bar\gamma$ of length 2.

Let us consider the product $k_1!I_{n,k_1}(f_1)k_2!I_{n,k_2}(f_2)$, and rewrite it in the form of the sum we get by carrying out a term by term multiplication in this expression. We put the terms obtained in such a way into disjoint classes indexed by the the diagrams $\bar\gamma \in \bar\Gamma(k_1, k_2)$ in the following way: A product $f_1(\xi_{j_1}, \dots, \xi_{j_{k_1}}) f_2(\xi_{j_1'}, \dots, \xi_{j_{k_2}'})$

belongs to the class indexed by the diagram $\bar\gamma((u_1,u_1'),\dots,(u_r,u_r'))$ with the parameters $(u_1,u_1'),\dots,(u_r,u_r')$, $1\le r\le\min(k_1,k_2)$, where $1\le u_1<u_2<\dots<u_r\le k_1$, the numbers $u_1',\dots,u_r'$ are different, and $1\le u_s'\le k_2$, for all $1\le s\le r$ if the indices $j_1,\dots,j_{k_1},j_1',\dots,j_{k_2}'$ in the arguments of the variables in $f_1(\cdot)$ and $f_2(\cdot)$ satisfy the relation $j_{u_s}=j'_{u_s'}$, $1\le s\le r$, and there is no more coincidence between the indices $j_1,\dots,j_{k_1},j_1',\dots,j_{k_2}'$.

It is not difficult to see by applying the above partition of the terms in the product $k_1!I_{n,k_1}(f_1)k_2!I_{n,k_2}(f_2)$, and exploiting that each diagram of $\bar\Gamma(k_1,k_2)$ can be written in the form $\bar\gamma((u_1,u_1'),\dots,(u_r,u_r'))$ in a unique way that the identity

$$n^{-k_1/2}k_1!I_{n,k_1}(f_1)k_2!n^{-k_2/2}I_{n,k_2}(f_2)=\sum_{\bar\gamma\in\bar\Gamma(k_1,k_2)}{}'^{(n)}\ n^{-(k_1+k_2)/2}s(\bar\gamma)!I_{n,s(\bar\gamma)}((\overline{f_1\circ f_2})_{\bar\gamma}) \tag{12.1}$$

holds, where the functions $(\overline{f_1\circ f_2})_{\bar\gamma}$ are defined in formula (11.3), $s(\bar\gamma)$ denotes the number of chains (of length 1 or 2) in $\bar\gamma$, and the notation $\sum'^{(n)}$ means that summation is taken only for such diagrams $\bar\gamma\in\bar\Gamma(k_1,k_2)$ for which $n\ge s(\bar\gamma)$. (Let me remark that although formula (11.3) was defined for coloured diagrams, the colours of the chains played no role in it.)

Relation (12.1) is not appropriate for our purposes, since the functions $(\overline{f_1\circ f_2})_{\bar\gamma}$ in it may be non-canonical. To get the desired formula, Hoeffding's decomposition will be applied for the U-statistics $I_{n,s(\bar\gamma)}((\overline{f_1\circ f_2})_{\bar\gamma})$ appearing at the right-hand side of formula (12.1). This decomposition becomes slightly simpler because of some special properties of the function $(\overline{f_1\circ f_2})_{\bar\gamma}$ related to the canonical property of the initial functions f_1 and f_2.

To carry out this procedure let us observe that a function $f(x_{u_1},\dots,x_{u_k})$ is canonical if and only if $P_{u_s}f(x_{u_1},\dots,x_{u_k})=0$ with the operator P_{u_s} defined in (11.1) for all indices u_s, $1\le s\le k$. Beside this, the condition that the functions f_1 and f_2 are canonical implies the relation $P_v(\overline{f_1\circ f_2})_{\bar\gamma}=0$ for $v\in V_1(\bar\gamma)$ for all diagrams $\bar\gamma\in\bar\Gamma(k_1,k_2)$, and this relation remains valid if the function $(\overline{f_1\circ f_2})_{\bar\gamma}$ is replaced by such functions which we get by applying the product of some transforms $P_{v'}$ and $Q_{v'}$, $v'\in P_2$, with the transforms P and Q defined in formulas (11.1) and (11.2).

Beside this, the transforms P_v or Q_v are exchangeable with the operators $P_{v'}$ or $Q_{v'}$ if $v\ne v'$, $P_v+Q_v)=I$, where I denotes the identity operator, and $P_vQ_v=0$, since $P_vQ_v=P_v-P_v^2=0$. The above relations make possible the following decomposition of the function $(\overline{f_1\circ f_2})_{\bar\gamma}$ for all $\bar\gamma\in\bar\Gamma(k_1,k_2)$ to the sum of canonical functions (just as it was done in the Hoeffding decomposition):

$$\begin{aligned}(\overline{f_1\circ f_2})_{\bar\gamma}&=\prod_{v\in V_2}(P_v+Q_v)(\overline{f_1\circ f_2})_{\bar\gamma}\\&=\sum_{A\subset V_2}\left(\prod_{v\in A}P_v\prod_{v\in V_2\setminus A}Q_v\right)(\overline{f_1\circ f_2})_{\bar\gamma}=\sum_{\gamma\in\Gamma(\bar\gamma)}(f_1\circ f_2)_\gamma,\end{aligned} \tag{12.2}$$

where the function $(f_1\circ f_2)_\gamma$ is defined in formula (11.4), and $\Gamma(\bar\gamma)$ denotes the set of those coloured diagrams $\gamma\in\Gamma(k_1,k_2)$ which consist of those chains (with a colour ±1) as

the non-coloured diagram $\bar\gamma$. (Clearly, $s(\gamma)=s(\bar\gamma)$ for the number of chains of γ and $\bar\gamma$ if $\gamma\in\Gamma(\bar\gamma)$.) Indeed, given a set $A\subset V_2$, we have $(\prod_{v\in A}P_v\prod_{v\in V_2\setminus A}Q_v)(\overline{f_1\circ f_2})_{\bar\gamma}=(f_1\circ f_2)_\gamma$ with that coloured diagram $\gamma\in\Gamma(\bar\gamma)$ whose chains with colour -1 are the chains $\beta(l)\in\bar\gamma$ with $l\in V_2\setminus A$, and summing up this identity for all $A\subset V_2$ we get relation (12.2). The function $(f_1\circ f_2)_\gamma$ corresponding to the coloured diagram obtained with the help of the set A has $|O(\gamma)|=k_1+k_2-|A|$ variables, where $|O(\gamma)|$ is the number of open indices in γ.

Let us consider the functions $F_\gamma(f_1,f_2)$, $\gamma\in\Gamma(k_1,k_2)$, defined in (11.5) which means a reindexation of the functions $(f_1\circ f_2)_\gamma$ to get functions with variables $x_1,\dots,x_{|O(\gamma)|}$. We claim that

$$\begin{aligned}&n^{-(k_1+k_2)/2}|O(\bar\gamma)|!I_{n,\bar s(\bar\gamma)}\left((\overline{f_1\circ f_2})_{\bar\gamma}\right)\\&\quad=\sum_{\gamma\in\Gamma(\bar\gamma)}n^{-(k_1+k_2)/2}n^{|C(\gamma)|}J_n(\gamma)|O(\gamma)|!I_{n,|O(\gamma)|}\left(F_\gamma(f_1,f_2)\right)\end{aligned}\tag{12.3}$$

with $J_n(\gamma)=1$ if $|C(\gamma)|=0$, and

$$J_n(\gamma)=\prod_{j=1}^{|C(\gamma)|}\left(\frac{n-s(\gamma)+j}{n}\right)\quad\text{if }\ |C(\gamma)|>0.\tag{12.4}$$

for all $\bar\gamma\in\bar\Gamma(k_1,k_2)$.

Since $I_{n,|O(\gamma)|}\left(F_\gamma(f_1,f_2)\right)=I_{n,|O(\gamma)|}\left(f_1\circ f_2\right)_\gamma)$ relation (12.3) follows from relation (12.2) just as formula (9.3) follows from formula (9.2) in the proof of the Hoeffding decomposition. Let us understand why the coefficient $n^{|C(\gamma)}J_n(\gamma)$ appears at the right-hand side of (12.3).

This coefficient can be calculated in the following way. Take a general term $(f_1\circ f_2)_\gamma(\xi_{j_{l_u}},\,l_u\in O(\gamma))$ in the U-statistic $|O(\gamma)|!I_{n,|O(\gamma)|}((f_1\circ f_2)_\gamma)$, and calculate the number of terms $(\overline{f_1\circ f_2})_{\bar\gamma}(\xi_{j'_1},\xi_{j'_2}\dots,\xi_{j'_{s(\bar\gamma)}})$ in the U-statistic $|O(\bar\gamma)|!I_{n,\bar s(\bar\gamma)}((\overline{f_1\circ f_2})_{\bar\gamma})$ for which the sequence of indices $(j'_1,\dots,j'_{s(\bar\gamma)})$ satisfies the relation $j'_{l_u}=j_{l_u}$ for all $l_u\in O(\gamma)$. I claim that it equals $n^{|C(\gamma)|}J_n(\gamma)$. It can be seen that this number $n^{|C(\gamma)|}J_n(\gamma)$ appears as the coefficient at right-hand side of (12.3).

Indeed, we have to calculate the number of such sequences $j'_1,j'_2,\dots,j'_{s(\bar\gamma)}$ for which the value $j'_{l_u}=j_{l_u}$ is prescribed for the indices $l_u\in O(\gamma)$, and the other elements of the sequence can take arbitrary integer value between 1 and n with the only restriction that all elements of the sequence $j'_1,j'_2,\dots,j'_{s(\bar\gamma)}$ must be different. The number of such sequences equals $(n-|O(\gamma)|)(n-|O(\gamma)|-1)\cdots(n-|C(\gamma)|-|O(\gamma)|+1)=J_n(\gamma)n^{|C(\gamma)|}$. (In this calculation we exploited the fact that $|O(\gamma)|+|C(\gamma)|=s(\gamma)$.)

Let us observe that $k_1+k_2-2|C(\gamma)|=|O(\gamma)|+W(\gamma)$ with the number $W(\gamma)$ introduced in the formulation of Theorem 11.1. Hence

$$n^{-(k_1+k_2)/2}n^{|C(\gamma)|}=n^{-W(\gamma)/2}n^{-|O(\gamma)|/2}.$$

Let us replace the left-hand side of the last identity by its right-hand side in (12.3), and let us sum up the identity we get in such a way for all $\bar\gamma \in \bar\Gamma(k_1, k_2)$ such that $s(\bar\gamma) \le n$. The identity we get in such a way together with formulas (12.1) and (12.4) imply the identity (11.6). Clearly, $I_{n,|O(\gamma)|}(F_\gamma(f_1, f_2)) = I_{n,|O(\gamma)|}(\mathrm{Sym}F_\gamma(f_1, f_2))$, hence the term $I_{n,|O(\gamma)|}(F_\gamma(f_1, f_2))$ can be replaced by $I_{n,|O(\gamma)|}(\mathrm{Sym}F_\gamma(f_1, f_2))$ in formula (11.6). We still have to prove inequalities (11.7) and (11.8).

Inequality (11.7), the estimate of the L_2-norm of the function $(f_1 \circ f_2)_\gamma$ follows from the Schwarz inequality, and actually it agrees with inequality (10.11), proved at the start of Appendix B. Hence its proof is omitted here.

To prove inequality (11.8) let us introduce, similarly to formula (11.2), the operators

$$\tilde Q_{u_j} h(x_{u_1}, \dots, x_{u_r}) = h(x_{u_1}, \dots, x_{u_r}) + \int h(x_{u_1}, \dots, x_{u_r}) \mu(\, dx_{u_j}), \quad 1 \le j \le r, \tag{12.5}$$

in the space of functions $h(x_{u_1}, \dots, x_{u_r})$ with coordinates in the space $(X, \mathcal{X})$. (The indices $u_1, \dots, u_r$ are all different.) Observe that both the operators $\tilde Q_{u_j}$ and the operators P_{u_j} defined in (11.1) are positive, i.e. these operators map a non-negative function to a non-negative function. Beside this, $Q_{u_j} \le \tilde Q_{u_j}$, and the norms of the operators $\frac{\tilde Q_{u_j}}{2}$ and P_{u_j} are bounded by 1 both in the $L_1(\mu)$, the $L_2(\mu)$ and the supremum norm.

Let us define the function

$$\widetilde{(f_1 \circ f_2)}_\gamma(x_j,\ j \in O(\gamma)) = \left(\prod_{j \in C(\gamma)} P_j \prod_{j \in O_2(\gamma)} \tilde Q_j \right) \overline{(f_1 \circ f_2)}_\gamma(x_j,\ j \in C(\gamma) \cup O(\gamma)) \tag{12.6}$$

with the notation of Section 11. The function $\widetilde{(f_1 \circ f_2)}_\gamma$ was defined with the help of $\overline{(f_1 \circ f_2)}_\gamma$ similarly to $(f_1 \circ f_2)_\gamma$ defined in (11.4), only the operators Q_j were replaced by $\tilde Q_j$ in its definition.

In the proof of (11.8) it may be assumed that $\|f_1\|_2 \le \|f_2\|_2$. The properties of the operators P_{u_j} and $\tilde Q_{u_j}$ listed above together with the condition $\sup |f_2(x_1, \dots, x_k)| \le 1$ imply that

$$|(f_1 \circ f_2)_\gamma| \le (\widetilde{|f_1|} \circ |f_2|)_\gamma \le (\widetilde{|f_1|} \circ 1)_\gamma, \tag{12.7}$$

where '$\le$' means that the function at the right-hand side is greater than or equal to the function at the left-hand side in all points, and the term 1 in (12.7) denotes the function which equals identically 1. Because of the identity $\|F_\gamma(f_1, f_2)\|_2 = \|(f_1 \circ f_2)_\gamma\|_2$ and relation (12.7) it is enough to show that

$$\begin{aligned} \|(\widetilde{|f_1|} \circ 1)_\gamma\|_2 &= \left\| \left(\prod_{j \in C(\gamma)} P_j \prod_{j \in O_2(\gamma)} \tilde Q_j \right) |f_1(x_{\alpha_\gamma(1,1)}, \dots, x_{\alpha_\gamma(1,k_1)})| \right\|_2 \\ &\le 2^{W(\gamma)} \|f_1\|_2. \end{aligned} \tag{12.8}$$

to prove relation (11.8). But this inequality trivially holds, since the norm of all operators P_j in formula (12.8) is bounded by 1, the norm of all operators $\tilde{Q}_j$ is bounded by 2 in the $L_2(\mu)$ norm, and $|O_2(\gamma)| = W(\gamma)$.

Proof of Theorem 11.2. Theorem 11.2 will be proved with the help of Theorem 11.2 by induction with respest to the number of degenerate U-statistics $k_p! I_{n,k_p}(f_p)$, $1 \le p \le m$. Formula (11.15) holds for $m = 2$ by Theorem 11.1. To prove it for a general parameter m let us first fix a coloured diagram $\bar{\gamma} \in \Gamma(k_1, \dots, k_{m-1})$ and consider the set of diagrams of m rows which are its 'continuation', i.e. let

$$\Gamma(\bar{\gamma}) = \{\gamma\colon\ \gamma \in \Gamma(k_1, \dots, k_m),\ \gamma_{pr} = \bar{\gamma}\}.$$

(Here we work with the diagrams γ_{pr} and γ_{cl} introduced for a diagram $\gamma \in \Gamma(k_1, \dots, k_m)$ in the previous section.) I claim that

$$\begin{aligned} & n^{-|O(\bar{\gamma})|/2} |O(\bar{\gamma})|! I_{n,|O(\bar{\gamma})|}(F_{\bar{\gamma}}(f_1, \dots, f_{m-1})) \cdot n^{-k_m/2} k_m! I_{n,k_m}(f_m) \\ & \quad = \sum_{\gamma \in \Gamma(\bar{\gamma})}{}' \prod_{j=1}^{|C(\gamma_{cl})|} \left(\frac{n - s(\gamma_{cl}) + j}{n} \right) n^{-W(\gamma_{cl})/2} \\ & \qquad n^{-|O(\gamma)|/2} |O(\gamma))|! I_{n,|O(\gamma)|}(F_{\gamma}(f_1, \dots, f_m)), \end{aligned} \tag{12.9}$$

where $\sum'$ means that summation is taken for such $\gamma \in \Gamma(\bar{\gamma})$ for which $s(\gamma_{cl}) \le n$, and $\prod_{j=1}^{|C(\gamma_{cl})|}$ equals 1, if $|C(\gamma_{cl})| = 0$.

Relation (12.9) can be checked by applying Theorem 11.1 for the pair of functions $F_{\bar{\gamma}}(f_1, \dots, f_{m-1})$ and f_m. To get it we have to understand that there is a mutual correspondence between the coloured diagrams $\gamma \in \Gamma(|O(\bar{\gamma})|, k_m)$ and the class of diagrams $\{\gamma_{cl}\colon\ \gamma \in \Gamma(\bar{\gamma})\}$. Indeed, for each $\gamma \in \Gamma(\bar{\gamma})$ there corresponds a diagram $\gamma_{cl} \in \Gamma(|O(\bar{\gamma})|, k_m)$. On the other hand, given a fixed enumeration of the chains of the diagram $\bar{\gamma}$ we can correspond to all diagrams $\gamma' \in \Gamma(|O(\bar{\gamma})|, k_m)$ a diagram $\gamma(\gamma') \in \Gamma(\bar{\gamma})$ such that $\gamma(\gamma')_{cl} = \gamma'$.

This diagram $\gamma(\gamma')$ can be defined in the following way. Let $\bar{l}_1, \bar{l}_2, \dots, \bar{l}_{|O(\bar{\gamma})|}$ be the indices of the chains with colour -1 of the diagram $\bar{\gamma}$. Then the the chains of colour 1 of $\bar{\gamma}$ will be chains of colour 1 of $\gamma(\gamma')$, too. If the vertex $(1, j)$ of the diagram γ' is contained in a chain of length 1, then the diagram $\gamma(\gamma')$ contains the chain $\beta(\bar{l}_j)$ with colour -1. If this vertex is contained in a chain $\{(1, j), (2, r_j)\} \in \gamma'$ of length 2, then $\gamma(\gamma')$ contains the diagram $\beta(\bar{l}_j) \cup \{(m, r_j)\}$ with the same colour as the chain $\{(1, j), (2, r_j)\}$ has in γ'. Finally, if the vertex $(2, r)$ is contained in the chain $\{(2, r)\}$ of length 1 in γ', then $\{(m, r)\}$ will be a chain of length 1 of $\gamma(\gamma')$ with colour -1. In such a way we get such a diagram $\gamma(\gamma') \in \Gamma(\bar{\gamma})$ for which $\gamma(\gamma')_{cl} = \gamma'$. (The above construction of the set $\gamma(\gamma')$ depends on the enumeration of the chains of the diagram $\bar{\gamma}$, but we get the same class of diagrams for two different enumerations of $\bar{\gamma}$ if we take the diagrams $\gamma(\gamma')$ for all $\gamma' \in \Gamma(|O(\bar{\gamma})|, k_m)$.)

We get relation (12.9) by applying Theorem 11.1 for the product

$$n^{-|O(\bar\gamma)|/2}|O(\bar\gamma)|!I_{n,|O(\bar\gamma)|}(F_{\bar\gamma}(f_1,\dots,f_{m-1}))\cdot n^{-k_m/2}k_m!I_{n,k_m}(f_m)$$

and writing all diagrams $\gamma'\in\Gamma(|O(\gamma)|,k_m)$ in the form γ_{cl}, where γ_{cl} is the closing diagram of the diagram $\gamma(\gamma')\in\Gamma(\bar\gamma)$ defined in the previous paragraph.

Relation (11.15) for the parameter m can be proved with the help of relation (12.9) and the inductive assumption by which it holds for $m-1$. Indeed, let us multiply formula (12.9) by $\prod_{p=2}^{m-1}J_n(\bar\gamma,p)n^{-W(\bar\gamma)/2}$, and sum up this identity for all such diagrams $\bar\gamma\in\Gamma(k_1,\dots,k_{m-1})$ for which $B_1(\gamma,p)+B_2(\bar\gamma,p)\le n$ for all $2\le p\le m-1$. Then the sum of the terms at the left-hand side equals the left-hand side of formula (11.15) for parameter m.

I claim that the sum of the terms at the right-hand side equals the right-hand side of formula (11.15) for parameter m. To see this it is enough to check that for all $\gamma\in\Gamma(\bar\gamma)$ we have $W(\bar\gamma)+W(\gamma_{cl})=W(\gamma_{pr})+W(\gamma_{cl})=W(\gamma)$, $\prod_{p=2}^{m-1}J_n(\gamma_{pr},p)\prod_{j=1}^{|C(\gamma_{cl})|}\left(\frac{n-s(\gamma_{cl})+j}{n}\right)=\prod_{p=2}^{m}J_n(\gamma,p)$, where $\prod_{j=1}^{|C(\gamma_{cl})|}=1$, if $|C(\gamma_{cl})|=0$, and the relation $B_1(\gamma,p)+B_2(\gamma,p)\le n$ holds for all $2\le p\le m$ if and only if $B_1(\gamma_{pr},p)+B_2(\gamma_{pr},p)\le n$ for all $2\le p\le m-1$, and $s(\gamma_{cl})\le n$. But these relations can be simply checked. The identity about the function $W(\cdot)$ can be checked by taking into account the definition of the diagrams γ_{pr} and γ_{cl}, in particular the colouring of the chains in these diagrams. The remaining relations can be proved with the help of the observation that for a diagram $\gamma\in\Gamma(k_1,\dots,k_m)$ $B_1(\gamma_{pr},p)=B_1(\gamma,p)$ and $B_2(\gamma_{pr},p)=B_2(\gamma,p)$ for all $2\le p\le m-1$. Beside this $|C(\gamma_{cl})|=B_1(\gamma,m)$ and $|O(\gamma_{cl})|=B_2(\gamma,m)$. Theorem 11.2 is proved.

Proof of Lemma 11.3. The proof is similar to that of formula (11.8) at the end of Theorem 11.1. Let us define the functions $\tilde F_\gamma(f_1,\dots,f_p)$, $\gamma\in\Gamma(k_1,\dots,k_p)$, recursively for all $2\le p\le m$ similarly to the definition of the functions $F_\gamma(f_1,\dots,f_p)$ with the difference that the operator $Q_{u_j}=I-P_{u_j}$ is replaced by $\tilde Q_{u_j}=I+P_{u_j}$ in the new definition. Then we have $|F_\gamma(f_1,\dots,f_m)|\le\tilde F_\gamma(|f_1|,\dots,|f_m|)$ in all points. Hence $\|F_\gamma(f_1,\dots,f_m)\|_2\le\|\tilde F_\gamma(f_1,\dots,f_m)\|_2$, and to prove Lemma 11.3 it is enough to show that

$$\|\tilde F_\gamma(|f_1|,\dots,|f_m|)\|_2\le 2^{W(\gamma)}\prod_{p\in U(\gamma)}\|f_p\|_2\quad\text{if }\gamma\in\Gamma(k_1,\dots,k_m)\tag{12.10}$$

with the same number $W(\gamma)$ and set $U(\gamma)$ which were considered in Lemma 11.3. Relation (12.10) will be proved by induction with respect to m.

Relation (12.10) holds for $m=2$. Indeed, if $W(\gamma)=0$, then $U(\gamma)=\{1,2\}$, we have $\tilde F_\gamma=F_\gamma$, and formula (11.7) supplies the estimate. If $W(\gamma)\ge1$, then $U(\gamma)=\{1\}$, and actually in the proof of relation (11.8) we proved this relation.

In the case $m > 2$ this inequality will be proved by induction with the help of the identity

$$\|\tilde{F}_\gamma(|f_1|,\dots,|f_m|)\|_2 = \Bigg\| \Bigg(\prod_{p\in C(\gamma_{cl})} P_p \prod_{p\in O_2(\gamma_{cl})} \tilde{Q}_p \Bigg) \overline{(\tilde{F}_{\gamma_{pr}}(|f_1|,\dots,|f_{m-1}|\circ|f_m|)}_{\gamma_{cl}}(x_p,\ p\in O(\gamma_{cl})\cup C(\gamma_{cl}))\Bigg\|_2 . \tag{12.11}$$

In the case $W(\gamma_{cl}) = 0$ we have $U(\gamma) = U(\gamma_{pr}) \cup \{m\}$, $W(\gamma) = W(\gamma_{cr})$, and formula (2.11) contains no operator $\tilde{Q}_p$. In this case inequality (12.10) follows from the representation of $\|\tilde{F}_\gamma(|f_1|,\dots,|f_m|)\|_2$ given in (12.11), relation (11.7) and from the inductive hypothesis by which inequality (12.10) holds for $\|(\tilde{F}_{\gamma_{pr}}(|f_1|,\dots,|f_{m-1}|)\|_2$.

In the case $W(\gamma_{cl}) > 0$ we have $U(\gamma) = U(\gamma_{pr})$, $W(\gamma) = W(\gamma_{pr}) + W(\gamma_{cl})$, and inequality (12.10) can be proved similarly to the case $W(\gamma_{cl}) = 0$ with the only difference that in this case instead of (11.7) we have to apply that strengthened version of (11.8) which is contained in formula (12.10) in the special case $m = 2$. Lemma 11.3 is proved.

The corollary of Theorem 11.2 is a simple consequence of Theorem 11.2 and Lemma 11.3.

Proof of the corollary of Theorem 11.2. Observe that F_γ is a function of $|O(\gamma)|$ arguments. Hence a coloured diagram $\gamma \in \Gamma(k_1,\dots,k_m)$ is in the class of closed diagrams, i.e. $\gamma \in \bar{\Gamma}(k_1,\dots,k_m)$ if and only if $F_\gamma(f_1,\dots,f_m)$ is a constant. Thus formula (11.16) is a simple consequence of relation (11.15) and the observation that $EI_{n,|O(\gamma)|}(F_\gamma(f_1,\dots,f_m)) = 0$ if $|O(\gamma)| \ge 1$, i.e. if $\gamma \notin \bar{\Gamma}(k_1,\dots,k_m)$, and

$$I_{n,O(\gamma)}(F_\gamma(f_1,\dots,f_m)) = F_\gamma(f_1,\dots,f_m) \quad \text{if } \gamma \in \bar{\Gamma}(k_1,\dots,k_m).$$

Relations (11.17) and (11.18) follow from relation (11.16) and Lemma 11.3.

13. The proof of Theorems 8.3, 8.5 and Example 8.7.

This section contains the proof of the estimates on the distribution of a multiple Wiener–Itô integral or degenerate U-statistic formulated in Theorems 8.5 and 8.3 together with the proof of Example 8.7. Beside this, also a multivariate version of Hoeffding's inequality (Theorem 3.4) will be proved here. The latter result is useful in the estimation of the supremum of degenerate U-statistics. The estimate on the distribution of a multiple random integral with respect to a normalized empirical distribution given in Theorem 8.1 is omitted, because, as it was shown in Section 9, this result follows from the estimate of Theorem 8.3 on degenerate U-statistics. This section will be finished with a separate part Section 13 B, where the results proved in this section are discussed together with the method of their proofs and some recent results.

The proof of Theorems 8.5 and 8.3 is based on a good estimate on high moments of Wiener–Itô integrals and degenerate U-statistics. These estimates follow from the corollaries of Theorems 10.2 and 11.2. Such an approach slightly differs from the classical proof in the one-variate case. The natural one-variate version of the problems discussed here is an estimate about the tail distribution of a sum of independent random variables. The latter estimate is generally proved by giving a good bound on the moment generating function of the sum. Such a method does not always work in the multivariate case, because, as later calculations will indicate, there is no good moment-generating function estimate for U-statistics or multiple Wiener–Itô integrals of order $k \geq 3$. Actually, the moment-generating function of a Wiener–Itô integral of order $k \geq 3$ is always divergent, because the tail behaviour of such a random integral is similar to that of the k-th power of a Gaussian random variable. On the other hand, good bounds on the moments EZ^{2M} of a random variable Z for all positive integers M (or at least for a sufficiently rich class of parameters M) together with the application of the Markov inequality for Z^{2M} and an appropriate choice of the parameter M yield a good estimate on the distribution of Z.

Propositions 13.1 and 13.2 give estimates on the moments of Wiener–Itô integrals and degenerate U-statistics.

Proposition 13.1. (Estimate of the moments of Wiener–Itô integrals). *Let $f(x_1, \dots, x_k)$ be a function of k variables on some measurable space $(X, \mathcal{X})$ that satisfies formula (8.12) with some σ-finite measure μ. Take the k-fold Wiener–Itô integral $Z_{\mu,k}(f)$ of this function with respect to a white noise μ_W with reference measure μ. The inequality*

$$E\left(k!|Z_{\mu,k}(f)|\right)^{2M} \leq 1 \cdot 3 \cdot 5 \cdots (2kM-1)\sigma^{2M} \quad \text{for all } M = 1, 2, \dots \tag{13.1}$$

holds.

By Stirling's formula Proposition 13.1 implies that

$$E(k!|Z_{\mu,k}(f)|)^{2M} \leq \frac{(2kM)!}{2^{kM}(kM)!}\sigma^{2M} \leq A\left(\frac{2}{e}\right)^{kM}(kM)^{kM}\sigma^{2M} \tag{13.2}$$

for any $A > \sqrt{2}$ if $M \ge M_0 = M_0(A)$. Formula (13.2) can be considered as a simpler, better applicable version of Proposition 13.1. It can be better compared with the moment estimate on degenerate U-statistics given in (13.3).

Proposition 13.2 provides a similar, but weaker inequality for the moments of normalized degenerate U-statistics.

Proposition 13.2. (Estimate on the moments of degenerate U-statistics). *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 exist 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*

$$E\left(n^{-k/2}k!I_{n,k}(f)\right)^{2M} \le A\,(1+C\sqrt{\eta})^{2kM}\left(\frac{2}{e}\right)^{kM}(kM)^{kM}\,\sigma^{2M} \quad \text{for all integers } M \text{ such that } kM_0 \le kM \le \eta n\sigma^2. \tag{13.3}$$

In formula (13.3) such a constant $C = C(k)$ can be chosen which does not depend on the order k of the U-statistic $I_{n,k}(f)$. For instance $C = 4$ is an appropriate choice.

Theorem 13.2 yields a good estimate on $E\left(n^{-k/2}k!I_{n,k}(f)\right)^{2M}$ with a fixed exponent $2M$ with the choice $\eta = \frac{kM}{n\sigma^2}$. With such a choice of the number η formula (13.3) yields an estimate on the moments $E\left(n^{-k/2}k!I_{n,k}(f)\right)^{2M}$ comparable with the estimate on the corresponding Wiener–Itô integral if $M \le n\sigma^2$, while it yields a much weaker estimate if $M \gg n\sigma^2$.

Now I turn to the proof of these propositions.

Proof of Proposition 13.1. Proposition 13.1 can be simply proved by means of the Corollary of Theorem 10.2 with the choice $m = 2M$, and $f_p = f$ for all $1 \le p \le 2M$. Formulas (10.18) and (10.19) yield that

$$E\left(k!Z_{\mu,k}(f)^{2M}\right) \le \left(\int f^2(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(dx_k)\right)^M |\Gamma_{2M}(k)| \le |\Gamma_{2M}(k)|\sigma^{2M},$$

where $|\Gamma_{2M}(k)|$ denotes the number of closed diagrams γ in the class $\bar{\Gamma}(\underbrace{k,\dots,k}_{2M \text{ times}})$ introduced in the corollary of Theorem 10.2. Thus to complete the proof of Proposition 13.1 it is enough to show that $|\Gamma_{2M}(k)| \le 1\cdot 3\cdot 5\cdots(2kM-1)$. But this can easily be seen with the help of the following observation. Let $\bar{\Gamma}_{2M}(k)$ denote the class of all graphs with vertices (l,j), $1 \le l \le 2M$, $1 \le j \le k$, such that from all vertices (l,j) exactly one edge starts, all edges connect different vertices, but edges connecting vertices (l,j) and (l,j') with the same first coordinate l are also allowed. Let $|\bar{\Gamma}_{2M}(k)|$ denote the number of graphs in $\bar{\Gamma}_{2M}(k)$. Then clearly $|\Gamma_{2M}(k)| \le |\bar{\Gamma}_{2M}(k)|$. On the other hand, $|\bar{\Gamma}_{2M}(k)| = 1\cdot 3\cdot 5\cdots(2kM-1)$. Indeed, let us list the vertices of the graphs from

$\bar{\Gamma}_{2M}(k)$ in an arbitrary way. Then the first vertex can be paired with another vertex in $2kM-1$ way, after this the first vertex from which no edge starts can be paired with $2kM-3$ vertices from which no edge starts. By following this procedure the next edge can be chosen $2kM-5$ ways, and by continuing this calculation we get the desired formula.

Proof of Proposition 13.2. Relation (13.3) will be proved by means of relations (11.17) and (11.18) in the Corollary of Theorem 11.2 with the choice $m = 2M$ and $f_p = f$ for all $1 \le p \le 2M$. The class of closed coloured diagrams $\Gamma(k, M) = \bar{\Gamma}(\underbrace{k, \dots, k}_{2M \text{times}})$ will be partitioned into subclasses $\Gamma(k, M, r)$, $1 \le r \le kM$, where $G(M, k, r)$ contains those closed diagrams $\gamma \in \Gamma(k, M)$ for which $W(\gamma) = 2r$. Let us recall that $W(\gamma)$ was defined in (11.13), and in the case of closed diagrams $W(\gamma) = \sum\limits_{\beta \in \gamma} (\ell(\beta) - 2)$. For a diagram $\gamma \in \Gamma(k, M)$, $W(\gamma)$ is an even number, since $W(\gamma) + 2s(\gamma) = 2kM$, where $s(\gamma)$ denotes the the number of chains in γ.

First we prove an estimate about the cardinality of $\Gamma(M, k, r)$. We claim that there exist some constant $A = A(k) > 0$ and threshold index $M_0 = M_0(k)$ depending only the order k of the U-statistic $In, k(f)$ such that

$$|\Gamma(k, M, r)| \le A \binom{2kM}{2r} \left(\frac{2}{e}\right)^{kM} (kM)^{kM+r} 2^{2r} \quad \text{for all } 0 \le r \le kM \tag{13.4}$$

if $A \ge A_0(k)$ and $M \ge M_0(k)$.

To prove formula (13.4) map all diagrams $\gamma \in \Gamma(k, M, r)$ to a paired diagram $T(\gamma)$ in such a way that $T(\gamma) \ne T(\gamma')$ if $\gamma \ne \gamma'$, and the number of paired diagrams $T(\gamma)$, $\gamma \in \Gamma(k, M, r)$, obtained in such a way can be well bounded. (We call a diagram a paired diagram, if all its chains have length 2, i.e. the set of its vertices is partitioned into pairs $\{(p, r), (p', r')\}$, with $p \ne p'$. To define these paired diagrams first we introduce the set $\mathcal{W}(\gamma) = \bigcup\limits_{\beta \in \gamma} \{(p_2(\beta), r_2(\beta)), \dots, (p_{s-1}(\beta), r_{s-1}(\beta))\}$, where $\beta = \{(p_1(\beta), r_1(\beta)), \dots, (p_s(\beta), r_s(\beta))\}$ with $1 \le p_1(\beta) < p_2(\beta) < \dots < p_s(\beta) \le 2M$ for all $\beta \in \gamma$, i.e. $\mathcal{W}(\gamma)$ is the set of vertices we get by omitting the first and last vertices of all chains $\beta \in \gamma$, and then taking the union of the vertices of these diminished chains. Observe that $|\mathcal{W}(\gamma)| = W(\gamma)$ for a closed diagram.

We take a copy (p, r, C) of all elements $(p, r) \in \mathcal{W}(\gamma)$ of a diagram $\gamma \in \Gamma(k, M, r)$, and define the set of vertices $V(T(\gamma))$ of the paired diagram $T(\gamma)$ as a set of vertices consisting of $2M$ rows, and the p-th row of this set is $\{(p, 1), \dots, (p, k_p)\} \cup \{(p, r, C)\colon\ (p, r) \in \mathcal{W}(\gamma)\}$ for all $1 \le p \le 2M$. Then we define the paired diagram $T(\gamma)$ on the set $V(T(\gamma))$ in the follwing way. Given a chain $\beta = \{(p_1(\beta), r_1(\beta)), \dots, (p_s(\beta), r_s(\beta))\} \in \gamma$, with $1 \le p_1(\beta) < p_2(\beta) < \dots < p_s(\beta) \le 2M$, we correspond to it the following sets of pairs (chains of length 2) in $V(T(\gamma))$:

$$\{((p_1(\beta), r_1(\beta)), ((p_2(\beta), r_2(\beta), C)\}, \{((p_2(\beta), r_2(\beta)), ((p_3(\beta), r_3(\beta), C)\}, \dots,$$
$$\{((p_{s-2}(\beta), r_{s-2}(\beta)), ((p_{s-1}(\beta), r_{s-1}(\beta), C)\}, \{((p_{s-1}(\beta), r_{s-1}(\beta)), ((p_s(\beta), r_s(\beta)\}.$$

(In the case $\ell(\beta) = 2$, we map β to itself.) Defining these pairs of vertices for all $\beta \in \gamma$ we get the paired diagram $T(\gamma)$ with the desired properties.

The number of the above defined sets $V(T(\gamma))$, $\gamma \in \Gamma(k, M, r)$, is less than or equal to $\binom{2kM}{2r}$, and each of these sets $V(T(\gamma))$ has $2kM + 2r$ vertices. Hence the number of paired diagrams with vertices in a fixed set $V(T(\gamma))$ is bounded by $1 \cdot 3 \cdot \cdots \cdot 5 \cdot (2kM - 2r - 1)$. The above considerations provide the bound

$$|\Gamma(k, M, r)| \le \binom{2kM}{2r} 1 \cdot 3 \cdot 5 \cdots (2kM + 2r - 1) = \binom{2kM}{2r} \frac{(2kM + 2r)!}{2^{kM+r}(kM + r)!}. \quad (13.5)$$

Stirling's formula yields that $\frac{(2kM+2r)!}{2^{kM+r}(kM+r)!} \le A \left(\frac{2}{e}\right)^{kM+r} (kM + r)^{kM+r}$ with some constant $A > \sqrt{2}$ if $M \ge M_0$ with some $M_0 = M_0(A)$. Since $r \le kM$ we can write $(kM + r)^{kM+r} \le (kM)^{kM} \left(1 + \frac{r}{kM}\right)^{kM} (2kM)^r \le (kM)^{kM+r} e^r 2^r$. The above calculation together with (13.5) imply inequality (13.4).

For a diagram $\gamma \in \Gamma(k, M, r)$ $W(\gamma) = 2r$, and beside this the cardinality of the set $U(\gamma)$ defined in the formulation of Lemma 11.3 satisfies the inequality $|U(\gamma)| \ge 2M - W(\gamma) = 2M - 2r$. Hence by relation (11.18) $n^{-W(\gamma)/2}|F_\gamma| \le 2^{2r} n^{-r} \sigma^{|U(\gamma)|} \le 2^{2r} \left(n\sigma^2\right)^{-r} \sigma^{2M} \le \eta^r 2^{2r} (kM)^{-r} \sigma^{2M}$ for $\gamma \in \Gamma(k, M, r)$ if $kM \le \eta n \sigma^2$ and $\sigma^2 \le 1$.

This estimate together with relation (11.17) imply that for $kM \le \eta n \sigma^2$

$$E\left(n^{-k/2} k! I_{n,k}(f_k)\right)^{2M} \le \sum_{\gamma \in \Gamma(k,M)} n^{-W(\gamma)/2} \cdot |F_\gamma| \le \sum_{r=0}^{kM} |\Gamma(k, M, r)| \eta^r 2^{2r} (kM)^{-r} \sigma^{2M}.$$

Hence by formula (13.4)

$$\begin{aligned} E\left(n^{-k/2} k! I_{n,k}(f_k)\right)^{2M} &\le A \left(\frac{2}{e}\right)^{kM} (kM)^{kM} \sigma^{2M} \sum_{r=0}^{kM} \binom{2kM}{2r} (4\sqrt{\eta})^{2r} \\ &\le A \left(\frac{2}{e}\right)^{kM} (kM)^{kM} \sigma^{2M} \left(1 + 4\sqrt{\eta}\right)^{2kM} \end{aligned}$$

if $kM_0 \le kM \le \eta n \sigma^2$. Thus we have proved Proposition 13.2 with $C = 4$.

It is not difficult to prove Theorem 8.5 with the help of Proposition 13.1.

Proof of Theorem 8.5. By formula (13.2) which is a consequence of Proposition 13.1 and the Markov inequality

$$P\left(|k! Z_{\mu,k}(f)| > u\right) \le \frac{E\left(k! Z_{\mu,k}(f)\right)^{2M}}{u^{2M}} \le A \left(\frac{2kM\sigma^{2/k}}{eu^{2/k}}\right)^{kM} \quad (13.6)$$

with some constant $A > \sqrt{2}$ if $M \ge M_0$ with some constant $M_0 = M_0(A)$, and M is an integer.

Put $\bar{M} = \bar{M}(u) = \frac{1}{2k}\left(\frac{u}{\sigma}\right)^{2/k}$, and $M = M(u) = [\bar{M}]$, where $[x]$ denotes the integer part of a real number x. Choose some number u_0 such that $\frac{1}{2k}\left(\frac{u_0}{\sigma}\right)^{2/k} \ge M_0 + 1$. Then relation (13.6) can be applied with $M = M(u)$ for $u \ge u_0$, and it yields that

$$\begin{aligned} P\left(|k!Z_{\mu,k}(f)| > u\right) &\le A\left(\frac{2kM\sigma^{2/k}}{eu^{2/k}}\right)^{kM} \le e^{-kM} \le Ae^k e^{-k\bar{M}} \\ &= Ae^k \exp\left\{-\frac{1}{2}\left(\frac{u}{\sigma}\right)^{2/k}\right\} \quad \text{if } u \ge u_0. \end{aligned} \tag{13.7}$$

Relation (13.7) means that relation (8.14) holds for $u \ge u_0$ with the pre-exponential coefficient Ae^k. By enlarging this coefficient if it is needed it can be guaranteed that relation (8.14) holds for all $u > 0$. Theorem 8.5 is proved.

Theorem 8.3 can be proved similarly by means of Proposition 13.2. Nevertheless, the proof is technically more complicated, since in this case the optimal choice of the parameter in the Markov inequality cannot be given in such a direct form as in the proof of Theorem 8.5. In this case the Markov inequality is applied with an only almost optimal choice of the parameter M.

Proof of Theorem 8.3. The Markov inequality and relation (13.3) with $\eta = \frac{kM}{n\sigma^2}$ imply that

$$\begin{aligned} P(k!n^{-k/2}|I_{n,k}(f)| > u) &\le \frac{E\left(k!n^{-k/2}I_{n,k}(f)\right)^{2M}}{u^{2M}} \\ &\le A\left(\frac{1}{e}\cdot 2kM\left(1 + C\frac{\sqrt{kM}}{\sqrt{n}\sigma}\right)^2\left(\frac{\sigma}{u}\right)^{2/k}\right)^{kM} \end{aligned} \tag{13.8}$$

for all integers $M \ge M_0$ with some $M_0 = M_0(A)$.

Relation (8.10) will be proved with the help of estimate (13.8) first in the case $D \le \frac{u}{\sigma} \le n^{k/2}\sigma^k$ with a sufficiently large constant $D = D(k, C) > 0$ depending on k and the constant C in (13.8). To this end let us introduce the number $\bar{M}$ by means of the formula

$$k\bar{M} = \frac{1}{2}\left(\frac{u}{\sigma}\right)^{2/k}\frac{1}{1 + B\frac{\left(\frac{u}{\sigma}\right)^{1/k}}{\sqrt{n}\sigma}} = \frac{1}{2}\left(\frac{u}{\sigma}\right)^{2/k}\frac{1}{1 + B\left(un^{-k/2}\sigma^{-(k+1)}\right)^{1/k}}$$

with a sufficiently large number $B = B(C) > 0$ and $M = [\bar{M}]$, where $[x]$ means the integer part of the number x.

Observe that $\sqrt{k\bar{M}} \le \left(\frac{u}{\sigma}\right)^{1/k}$, $\frac{\sqrt{k\bar{M}}}{\sqrt{n}\sigma} \le \left(un^{-k/2}\sigma^{-(k+1)}\right)^{1/k} \le 1$, and

$$\left(1 + C\frac{\sqrt{k\bar{M}}}{\sqrt{n}\sigma}\right)^2 \le 1 + B\frac{\sqrt{k\bar{M}}}{\sqrt{n}\sigma} \le 1 + B\left(un^{-k/2}\sigma^{-(k+1)}\right)^{1/k}$$

with a sufficiently large $B = B(C) > 0$ if $\frac{u}{\sigma} \le n^{k/2}\sigma^k$. Hence

$$\begin{aligned}\frac{1}{e} \cdot 2kM \left(1 + C\frac{\sqrt{kM}}{\sqrt{n}\sigma}\right)^2 \left(\frac{\sigma}{u}\right)^{2/k} &\le \frac{1}{e} \cdot 2k\bar{M} \left(1 + C\frac{\sqrt{k\bar{M}}}{\sqrt{n}\sigma}\right)^2 \left(\frac{\sigma}{u}\right)^{2/k} \\ &\le \frac{1}{e} \cdot \frac{\left(1 + C\frac{\sqrt{k\bar{M}}}{\sqrt{n}\sigma}\right)^2}{1 + B\left(un^{-k/2}\sigma^{-(k+1)}\right)^{1/k}} \le \frac{1}{e}\end{aligned} \tag{13.9}$$

if $\frac{u}{\sigma} \le n^{k/2}\sigma^k$. If the inequality $D \le \frac{u}{\sigma}$ also holds with a sufficiently large $D = D(B,k) > 0$, then $M \ge M_0$, and the conditions of inequality (13.8) hold. This inequality together with inequality (13.9) yield that

$$P(k!n^{-k/2}|I_{n,k}(f)| > u) \le Ae^{-kM} \le Ae^k e^{-k\bar{M}}$$

if $D \le \frac{u}{\sigma} \le n^{k/2}\sigma^k$, i.e. inequality (8.10) holds in this case with a pre-exponential constant Ae^k. By increasing the pre-exponential constant Ae^k in this inequality we get that relation (8.10) holds for all $0 \le \frac{u}{\sigma} \le n^{k/2}\sigma^k$. Theorem 8.3 is proved.

Example 8.7 is a relatively simple consequence of Itô's formula for multiple Wiener–Itô integrals.

Proof of Example 8.7. We may restrict our attention to the case $k \ge 2$. Itô's formula for multiple Wiener-Itô integrals, more explicitly relation (10.21), implies that the random variable $k!Z_{\mu,k}(f)$ 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

$$P\left(|\eta^k| - D|\eta^{k-2}| > \frac{u}{\sigma}\right) \ge P\left(|\eta^k| > \frac{u}{\sigma} + A\left(\frac{u}{\sigma}\right)^{(k-2)/k}\right) \quad \text{if } \frac{u}{\sigma} > B.$$

Hence

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

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 (8.16) holds.

In this section also the multivariate version of Hoeffding's inequality will be proved. Before its formulation some notations will be introduced.

Let us fix two positive integers k and n and some real numbers $a(j_1, \dots, j_k)$ for all sequences of arguments $\{j_1, \dots, j_k\}$ such that $1 \le j_l \le n$, $1 \le l \le k$, and $j_l \ne j_{l'}$ if $l \ne l'$.

With the help of the above real numbers $a(\cdot)$ and a sequence of independent random variables $\varepsilon_1, \dots, \varepsilon_n$, $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = \frac{1}{2}$, $1 \le j \le n$, the random variable

$$V = \sum_{\substack{(j_1,\dots,j_k)\colon 1 \le j_l \le n \text{ for all } 1 \le l \le k, \\ j_l \ne j_{l'} \text{ if } l \ne l'}} a(j_1, \dots, j_k)\varepsilon_{j_1} \cdots \varepsilon_{j_k} \tag{13.10}$$

and number

$$S^2 = \sum_{\substack{(j_1,\dots,j_k)\colon 1 \le j_l \le n \text{ for all } 1 \le l \le k, \\ j_l \ne j_{l'} \text{ if } l \ne l'}} a^2(j_1, \dots, j_k). \tag{13.11}$$

will be introduced.

With the help of the above notations the following result can be formulated.

Theorem 13.3. (The multivariate version of Hoeffding's inequality). *The random variable V defined in formula (13.10) satisfies the inequality*

$$P(|V| > u) \le C \exp\left\{-\frac{1}{2}\left(\frac{u}{S}\right)^{2/k}\right\} \quad \text{for all } u \ge 0 \tag{13.12}$$

with the constant S defined in (13.11) and some constants $C > 0$ depending only on the parameter k in the expression V.

Theorem 13.3 will be proved by means of two simple lemmas. Before their formulation the random variable

$$Z = \sum_{\substack{(j_1,\dots,j_k)\colon 1 \le j_l \le n \text{ for all } 1 \le l \le k, \\ j_l \ne j_{l'} \text{ if } l \ne l'}} |a(j_1, \dots, j_k)|\eta_{j_1} \cdots \eta_{j_k} \tag{13.13}$$

will be introduced, where $\eta_1, \dots, \eta_n$ are independent random variables with standard normal distribution, and the numbers $a(j_1, \dots, j_k)$ agree with those in formula (13.10). The following lemmas will be proved.

Lemma 13.4. *The random variables V and Z introduced in (13.10) and (13.13) satisfy the inequality*

$$EV^{2M} \le EZ^{2M} \quad \text{for all } M = 1, 2, \dots. \tag{13.14}$$

Lemma 13.5. *The random variable Z defined in formula (13.13) satisfies the inequality*

$$EZ^{2M} \le 1 \cdot 3 \cdot 5 \cdots (2kM - 1)S^{2M} \quad \text{for all } M = 1, 2, \dots \tag{13.15}$$

with the constant S defined in formula (13.11).

Proof of Lemma 13.4. We can write, by carrying out the multiplications in the expressions EV^{2M} and EZ^{2M}, by exploiting the additive and multiplicative properties of the expectation for sums and products of independent random variables together with the identities $E\varepsilon_j^{2k+1} = 0$ and $E\eta_j^{2k+1} = 0$ for all $k = 0, 1, \dots$ that

$$EV^{2M} = \sum_{\substack{(j_1,\dots,j_l,\, m_1,\dots,m_l):\\ 1\le j_s\le n,\ m_s\ge 1,\ 1\le s\le l,\ m_1+\dots+m_l=kM}} A(j_1,\dots,j_l,m_1,\dots,m_l) E\varepsilon_{j_1}^{2m_1}\cdots E\varepsilon_{j_l}^{2m_l} \quad (13.16)$$

and

$$EZ^{2M} = \sum_{\substack{(j_1,\dots,j_l,\, m_1,\dots,m_l):\\ 1\le j_s\le n,\ m_s\ge 1,\ 1\le s\le l,\ m_1+\dots+m_l=kM}} B(j_1,\dots,j_l,m_1,\dots,m_l) E\eta_{j_1}^{2m_1}\cdots E\eta_{j_l}^{2m_l} \quad (13.17)$$

with some coefficients $A(j_1,\dots,j_l,m_1,\dots,m_l)$ and $B(j_1,\dots,j_l,m_1,\dots,m_l)$ such that

$$|A(j_1,\dots,j_l,m_1,\dots,m_l)| \le B(j_1,\dots,j_l,m_1,\dots,m_l). \quad (13.18)$$

The coefficients $A(\cdot,\cdot,\cdot)$ and $B(\cdot,\cdot,\cdot)$ could be expressed explicitly, but we do not need such a formula. What is important for us is that $A(\cdot,\cdot,\cdot)$ can be expressed as the sum of certain terms, and $B(\cdot,\cdot,\cdot)$ as the sum of the absolute value of the same terms. Hence relation (13.18) holds. Since $E\varepsilon_j^{2m} \le E\eta_j^{2m}$ for all parameters j and m formulas (13.16), (13.17) and (13.18) imply Lemma 13.4.

Proof of Lemma 13.5. Let us consider a white noise $W(\cdot)$ on the unit interval $[0,1]$ with the Lebesgue measure λ on $[0,1]$ as its reference measure, i.e. let us take a set of Gaussian random variables $W(A)$ indexed by the measurable sets $A \subset [0,1]$ such that $EW(A) = 0$, $EW(A)W(B) = \lambda(A \cap B)$ with the Lebesgue measure λ for all measurable subsets of the interval $[0,1]$. Let us introduce n orthonormal functions $\varphi_1(x), \dots, \varphi_n(x)$ with respect to the Lebesgue measure on the interval $[0,1]$, and define the random variables $\eta_j = \int \varphi_j(x) W(\,dx)$, $0 \le j \le n$. Then $\eta_1, \dots, \eta_n$ are independent random variables with standard normal distribution, hence we may assume that they appear in the definition of the random variable Z in formula (13.13). Beside this, the identity $\eta_{j_1} \cdots \eta_{j_k} = \int \varphi_{j_1}(x_1) \cdots \varphi_{j_k}(x_k) W(\,dx_1) \dots W(\,dx_k)$ holds for all k-tuples $(j_1, \dots, j_k)$, such that $1 \le j_s \le n$ for all $1 \le s \le k$, and the indices $j_1, \dots, j_s$ are different. This identity follows from Itô's formula for multiple Wiener–Itô integrals formulated in formula (10.20) of Theorem 10.3.

Hence the random variable Z defined in (13.13) can be written in the form

$$Z = \int f(x_1, \dots, x_k) W(\,dx_1) \dots W(\,dx_k)$$

with the function

$$f(x_1, \dots, x_k) = \sum_{\substack{(j_1,\dots,j_k):\ 1\le j_l\le n \text{ for all } 1\le l\le k,\\ j_l\ne j_{l'} \text{ if } l\ne l'}} |a(j_1, \dots, j_k)| \varphi_{j_1}(x_1) \cdots \varphi_{j_k}(x_k).$$

Because of the orthogonality of the functions $\varphi_j(x)$

$$S^2 = \int_{[0,1]^k} f^2(x_1, \dots, x_k)\, dx_1 \dots\, dx_k.$$

Lemma 13.5 is a straightforward consequence of the above relations and formula (13.1) in Proposition 13.1.

Proof of Theorem 13.3. The proof of Theorem 13.3 with the help of Lemmas 13.4 and 13.5 is an almost word for word repetition of the proof of Theorem 8.5. By Lemma 13.4 inequality (13.15) remains valid if the random variable Z is replaced by the random variable V at its left-hand side. Hence the Stirling formula yields that

$$EV^{2M} \le EZ^{2M} \le \frac{(2kM)!}{2^{kM}(kM)!} S^{2M} \le C \left(\frac{2}{e}\right)^{kM} (kM)^{kM} S^{2M}$$

for any $C \ge \sqrt{2}$ if $M \ge M_0(A)$. As a consequence, by the Markov inequality the estimate

$$P(|V| > u) \le \frac{EV^{2M}}{u^{2M}} \le C \left(\frac{2kM}{e} \left(\frac{S}{u}\right)^{2/k} \right)^{kM} \tag{13.19}$$

holds for all $C > \sqrt{2}$ if $M \ge M_0(C)$. Put $k\bar{M} = k\bar{M}(u) = \frac{1}{2}\left(\frac{u}{S}\right)^{2/k}$ and $M = M(u) = [\bar{M}]$, where $[x]$ denotes the integer part of the number x. Let us choose a threshold number u_0 by the identity $\frac{1}{2k}\left(\frac{u_0}{S}\right)^{2/k} = M_0(C) + 1$. Formula (13.19) can be applied with $M = M(u)$ for $u \ge u_0$, and it yields that

$$P(|V| > u) \le Ce^{-kM} \le Ce^k e^{-k\bar{M}} = Ce^k \exp\left\{-\frac{1}{2}\left(\frac{u}{S}\right)^{2/k}\right\} \quad \text{if } u \ge u_0.$$

The last inequality means that relation (13.12) holds for $u \ge u_0$ if the constant C is replaced by Ce^k in it. With the choice of a sufficiently large constant C relation (13.12) holds for all $u \ge 0$. Theorem 13.3 is proved.

13. B) A short discussion about the methods and results.

A comparison of Theorem 8.5 and Example 8.7 shows that the estimate (8.15) is sharp. At least no essential improvement of this estimate is possible which holds for *all* Wiener–Itô integrals with a kernel function f satisfying the conditions of Theorem 8.5. This fact also indicates that the bounds (13.1) and (13.2) on high moments of Wiener–Itô integrals are sharp. It is worth while comparing formula (13.2) with the estimate of Proposition 13.2 on moments of degenerate U-statistics.

Let us consider a normalized k-fold degenerate U-statistic $n^{-k/2}k!I_{n,k}(f)$ with some kernel function f and a μ-distributed sample of size n. Let us compare its moments with those of a k-fold Wiener–Itô integral $\mathrm{k!}Z_{\mu,k}(f)$ with the same kernel function f with respect to a white noise μ_W with reference measure μ. Let σ denote the L_2-norm of the

kernel function f. If $M \le \varepsilon n\sigma^2$ with a small number $\varepsilon > 0$, then Proposition 13.2 (with an appropriate choice of the parameter η which is small in this case) provides an almost as good bound on the $2M$-th moment of the normalized U-statistic as Proposition 13.1 provides on the $2M$-th moment of the corresponding Wiener–Itô integral. In the case $M \le Cn\sigma^2$ with some fixed (not necessarily small) number $C > 0$ the $2M$-th moment of the normalized U-statistic can be bounded by $C(k)^M$ times the natural estimate on the $2M$-th moment of the Wiener–Itô integral with some constant $C(k) > 0$ depending only on the number C. This can be so interpreted that in this case the estimate on the moments of the normalized U-statistic is weaker than the estimate on the moments of the Wiener–Itô integral, but they are still comparable. Finally, in the case $M \gg n\sigma^2$ the estimate on the $2M$-th moment of the normalized U-statistic is much worse than the estimate on the $2M$-th moment of the Wiener–Itô integral.

A similar picture arises if the distribution of the normalized degenerate U-statistic

$$F_n(u) = P(n^{-k/2}k!|I_{n,k}(f)| > u)$$

is compared to the distribution of the Wiener–Itô integral

$$G(u) = P(k!|Z_{\mu,k}(f)| > u).$$

A comparison of Theorems 8.3 and 8.5 shows that for $0 \le u \le \varepsilon n^{k/2}\sigma^{k+1}$ with a small $\varepsilon > 0$ an almost as good estimate holds $F_n(u)$ as for $G(u)$. In the case $0 \le u \le n^{k/2}\sigma^{k+1}$ the behaviour of $F_n(u)$ and $G(u)$ is similar, only in the exponent of the estimate on $F_n(u)$ in formula (8.10) a worse constant appears. Finally, if $u \gg n^{k/2}\sigma^{k+1}$, then — as Example 8.8 shows, at least in the case $k = 2$, — the (tail) distribution function $F_n(u)$ satisfies a much worse estimate than the function $G(u)$. Thus a similar picture arises as in the case when the estimate on the tail-distribution of normalized sums of independent random variables, discussed in Section 3, is compared to the behaviour of the standard normal distribution in the neighbourhood of infinity. To understand this similarity better it is useful to recall Theorem 10.4, the limit theorem about normalized degenerate U-statistics. Theorems 8.3 and 8.5 enable us to compare the tail behaviour of normalized degenerate U-statistics with their limit presented in the form of multiple Wiener–Itô integrals, while the one-variate versions of these results compare the distribution of sums of independent random variables with their Gaussian limit.

The above results show that good bounds on the moments of degenerate U-statistics and multiple Wiener–Itô also provide a good estimate on their distribution. To understand the behaviour of high moments of degenerate U-statistics it is useful to have a closer look at the simplest case $k = 1$, when the moments of sums of independent random variables with expectation zero are considered.

Let us consider a sequence of independent and identically distributed random variables $\xi_1, \dots, \xi_n$ with expectation zero, take their sum $S_n = \sum_{j=1}^{n} \xi_j$, and let us try to give a good estimate on the moments ES_n^{2M} for all $M = 1, 2, \dots$. Because of the

independence of the random variables ξ_j and the condition $E\xi_j = 0$ the identity

$$ES_n^{2M} = \sum_{\substack{(j_1,\dots,j_s,l_1,\dots,l_s) \\ j_1+\dots+j_s=2M,\, j_u\ge 2,\text{ for all } 1\le u\le s \\ l_u\ne l_{u'} \text{ if } u\ne u'}} E\xi_{l_1}^{j_1}\cdots E\xi_{l_s}^{j_s} \tag{13.20}$$

holds. Simple combinatorial considerations show that a dominating number of terms at the right-hand side of (13.20) are indexed by a vector $(j_1,\dots,j_M; l_1,\dots,l_M)$ such that $j_u = 2$ for all $1 \le u \le M$, and the number of such vectors is equal to $\binom{n}{M}\frac{(2M)!}{2^M} \sim n^M\frac{(2M)!}{2^M M!}$. The last asymptotic relation holds if the number n of terms in the random sum S_n is sufficiently large. The above considerations suggest that under not too restrictive conditions $ES_n^{2M} \sim \left(n\sigma^2\right)^M \frac{(2M)!}{2^M M!} = E\eta_{n\sigma^2}^{2M}$, where $\sigma^2 = E\xi^2$ is the variance of the terms in the sum S_n, and η_u denotes a random variable with normal distribution with expectation zero and variance u. The question arises when the above heuristic argument gives a right estimate.

For the sake of simplicity let us restrict our attention to the case when the absolute value of the random variables ξ_j is bounded by 1. Let us observe that even in this case the above heuristic argument holds only under the condition that the variance σ^2 of the random variables ξ_j is not too small. Indeed, let us consider such random variables ξ_j, for which $P(\xi_j = 1) = P(\xi_j = -1) = \frac{\sigma^2}{2}$, $P(\xi_j = 0) = 1 - \sigma^2$. Then these random variables ξ_j have variance σ^2, and the contribution of the terms $E\xi_j^{2M}$, $1 \le j \le n$, to the sum in (13.20) equals $n\sigma^2$. If σ^2 is very small, then it may happen that $n\sigma^2 \gg \left(n\sigma^2\right)^M \frac{(2M)!}{2^M M!}$, and the approximation given for ES_n^{2M} in the previous paragraph does not hold any longer. Hence the asymptotic relation for a very high moment ES_n^{2M} suggested by the above heuristic argument may only hold if the variance σ^2 of the summands satisfies an appropriate lower bound.

In the proof of Proposition 13.2 a similar picture appears in a hidden way. In the calculation of the moments of a degenerate U-statistic the contribution of certain (closed) diagrams, more precisely of some integrals defined with their help, has to be estimated. Some of these diagrams (those in which all chains have length 2) appear also in the calculation of the moments of multiple Wiener–Itô integrals. In the calculation of the moments of sums of independent random variables the terms consisting of products of second moments play such a role in the sum in formula (13.20) as the 'nice' diagrams consisting of chains of length 2 play in the calculation of the moments of degenerate U-statistics in formula (11.17). In nice cases the remaining diagrams do not give a much greater contribution than these 'nice' diagrams, and we get an almost as good bound for the moments of a normalized degenerate U-statistic as for the moments of the corresponding multiple Wiener–Itô integral. The proof of Proposition 13.2 shows that such a situation appears under very general conditions.

Let me also remark that there is an essential difference between the tail behaviour of Wiener–Itô integrals and normalized degenerate U-statistics. A good estimate can be given on the tail distibution of Wiener–Itô integrals which depends only on the L_2-norm of the kernel function, while in the case of normalized degenerate U-statistics the

corresponding estimate depends not only on the L_2-norm but also on the L_∞ norm of the kernel function. In Theorem 8.3 such an estimate is proved. Moreover, it can be shown that this dependence of the estimate on the L_∞ norm of the kernel function is essential, it appears not only in this result.

For $k \geq 2$ the distribution of k-fold Wiener-Itô integrals are not determined by the L_2-norm of their kernel functions. This is an essential difference between Wiener–Itô integrals of order $k \geq 2$ and $k = 1$. In the case $k = 1$ a Wiener–Itô integral is a Gaussian random variable with expectation zero, and its variance equals the square of the L_2-norm of its kernel function. Hence its distribution is completely determined by the L_2-norm of its kernel function. On the other hand, the distribution of a Wiener–Itô integral of order $k \geq 2$ is not determined by its variance. Theorem 8.5 yields a 'worst case' estimate on the distribution of Wiener–Itô integrals if we have a bound on their variance. In the statistical problems which provided the main motivation for this work such estimates are needed, but it may be interesting to know what kind of estimates are known about the distribution of a multiple Wiener–Itô integral or degenerate U-statistic if we have some additional information about its kernel function. Some results will be mentioned in this direction, but several technical details will be omitted from their discussion.

H. P. Mc. Kean proved the following lower bound on the distribution of multiple Wiener–Itô integrals. (See [28] or [41].)

Theorem 13.6. (Lower bound on the distribution of Wiener–Itô integrals). *All k-fold Wiener–Itô integrals $Z_{\mu,k}(f)$ satisfy the inequality*

$$P(|Z_{\mu,k}(f)| > u) > Ke^{-Au^{2/k}} \tag{13.21}$$

with some numbers $K = K(f, \mu) > 0$ and $A = A(f, \mu) > 0$.

The constant A in the exponent $Au^{2/k}$ of formula (13.21) is always finite, but Mc. Kean's proof yields no explicit upper bound on it. The following example shows that in certain cases if we fix the constant K in relation (13.21), then this inequality holds only with a very large constant $A > 0$ even if the variance of the Wiener–Itô integral equals 1.

Take a probability measure μ and a white noise μ_W with reference measure μ on a measurable space $(X, \mathcal{X})$, and let $\varphi_1, \varphi_2, \dots$ be a sequence of orthonormal functions on $(X, \mathcal{X})$ with respect to this measure μ. Define for all $L = 1, 2, \dots$, the function

$$f(x_1, \dots, x_k) = f_L(x_1, \dots, x_k) = (k!)^{1/2} L^{-1/2} \sum_{j=1}^{L} \varphi_j(x_1) \cdots \varphi_j(x_k) \tag{13.22}$$

and the Wiener–Itô integral

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

Then $EZ^2_{\mu,k}(f) = 1$, and the high moments of $Z_{\mu,k}(f)$ can be well estimated. For a large parameter L these moments are much smaller, than the quantities suggested by

Proposition 13.1. (The calculation leading to the estimation of the moments of $Z_{\mu,k}(f)$ will be omitted.) These moment estimates also imply that if the parameter L is large, then for not too large numbers u the probability $P(|Z_{\mu,k}(f)| > u)$ has a much better estimate than that given in Theorem 8.5. As a consequence, for a large number L and fixed number K relation (13.21) may hold only with a very big number $A > 0$.

We can expect that if we take a Gaussian random polynomial $P(\xi_1, \dots, \xi_n)$ whose arguments are Gaussian random variables $\xi_1, \dots, \xi_n$, and which is the sum of many small almost independent terms, then a similar picture arises as in the case of a Wiener–Itô integral with kernel function (13.22) with a large parameter L. Such a random polynomial has an almost Gaussian distribution by the central limit theorem, and we can also expect that its not too high moments behave so as the corresponding moments of a Gaussian random variable with expectation zero and the same variance as the Gaussian random polynomial we consider. Such a bound on the moments has the consequence that the estimate on the probability $(P(\xi_1, \dots, \xi_n) > u)$ given in Theorem 8.5 can be improved if the number u is not too large. A similar picture arises if we consider Wiener–Itô integrals whose kernel function satisfies some 'almost independence' properties. The problem is to find the right properties under which we can get a good estimate that exploits the almost independence property of a Gaussian random polynomial or of a Wiener–Itô integral. The main result of R. Latała's paper [26] can be considered as a response to this question. I describe this result below.

To formulate Latała's result some new notions have to be introduced. Given a finite set A let $\mathcal{P}(A)$ denote the set of all its partitions. If a partition $P = \{B_1, \dots, B_s\} \in \mathcal{P}(A)$ consists of s elements then we say that this partition has order s, and write $|P| = s$. In the special case $A = \{1, \dots, k\}$ the notation $\mathcal{P}(A) = \mathcal{P}_k$ will be used. Given a measurable space $(X, \mathcal{X})$ with a probability measure μ on it together with a finite set $B = \{b_1, \dots, b_j\}$ let us introduce the following notations. Take j different copies $(X_{b_r}, \mathcal{X}_{b_r})$ and μ_{b_r}, $1 \le r \le j$, of this measurable space and probability measure indexed by the elements of the set B, and define their product $(X^{(B)}, \mathcal{X}^{(B)}, \mu^{(B)}) = \left(\prod_{r=1}^{j} X_{b_r}, \prod_{r=1}^{j} \mathcal{X}_{b_r}, \prod_{r=1}^{j} \mu_{b_r} \right)$. The points $(x_{b_1}, \dots, x_{b_j}) \in X^{(B)}$ will be denoted by $x^{(B)} \in X^{(B)}$ in the sequel. With the help of the above notations I introduce the quantities needed in the formulation of the following Theorem 13.7.

Let a function $f = f(x_1, \dots, x_k)$ be given on the k-fold product $(X^k, \mathcal{X}^k, \mu^k)$ of a measurable space $(X, \mathcal{X})$ with a probability measure μ. For all partitions $P = \{B_1, \dots, B_s\} \in \mathcal{P}_k$ of the set $\{1, \dots, k\}$ consider the functions $g_r\left(x^{(B_r)}\right)$ on the space $X^{(B_r)}$, $1 \le r \le s$, and define with their help the quantities

$$\alpha(P) = \alpha(P, f, \mu) = \sup_{g_1, \dots, g_s} \int f(x_1, \dots, x_k) g_1\left(x^{(B_1)}\right) \cdots g_s\left(x^{(B_s)}\right) \mu(dx_1) \dots \mu(dx_k);$$

$$\text{where supremum is taken for such functions } g_1, \dots, g_s, \quad g_r \colon X^{B_r} \to R^1$$

$$\text{for which } \int g_r^2\left(x^{(B_r)}\right) \mu^{(B_r)}\left(dx^{(B_r)}\right) \le 1 \quad \text{for all } 1 \le r \le s, \tag{13.23}$$

and put

$$\alpha_s = \max_{P \in \mathcal{P}_k,\, |P|=s} \qquad 1 \le s \le k. \tag{13.24}$$

In Latałas estimation of Wiener–Itô integrals of order k the quantities α_s, $1 \le s \le k$, play a similar role as the number σ^2 in Theorem 8.5. Observe that in the case $|P| = 1$, i.e. if $P = \{1, \dots, k\}$ the identity $\alpha^2(P) = \int f^2(x_1, \dots, x_k)\mu(\,dx_1) \dots \mu(\,dx_k)$ holds, which means that $\alpha_1 = \sigma$. The following estimate is valid for Wiener–Itô integrals of general order.

Theorem 13.7. Latała's estimate about the tail-distribution of Wiener–Itô integrals. *Let a k-fold Wiener–Itô integral $Z_{\mu,k}(f)$, $k \ge 1$, be defined with the help of a white noise μ_W with a non-atomic reference measure μ and a kernel function f of k-variable such that $\int f^2(x_1, \dots, x_k)\mu(\,dx_1) \dots \mu(\,dx_k) < \infty$. There is some universal constant $C(k) < \infty$ depending only of the order k of the random integral such that the inequalities*

$$E(Z_{\mu,k}(f))^{2M} \le \left(C(k) \max_{1 \le s \le k} (M^{s/2}\alpha_s) \right)^{2M}, \tag{13.25}$$

and

$$P(|Z_{\mu,k}(f)| > u) \le C(k) \exp\left\{ -\frac{1}{C(k)} \min_{1 \le s \le k} \left(\frac{u}{\alpha_s} \right)^{2/s} \right\} \tag{13.26}$$

hold for all $M = 1, 2, \dots$ and $u > 0$ with the quantities α_s, defined in formulas (13.23) and (13.24).

Inequality (13.26) is a simple consequence of (13.25). In the special case when $\alpha_s \le M^{-(s-1)/2}$ for all $1 \le s \le k$, then inequality (13.25) says that the moment $EZ_{\mu,k}(f)^{2M}$ has the same magnitude as the $2M$-th moment of a standard Gaussian random variable multiplied by a constant, and it implies a good estimate on $P(|Z_{\mu,k}(f)| > u)$ given in (13.26). Actually the result of Theorem 13.7 can be reduced to the special case when $\alpha_s \le M^{-(s-1)/2}$ for all $1 \le s \le k$. Thus it can be interpreted so that if the quantities α_s of a k-fold Wiener–Itô integral are sufficiently small, then these 'almost independence' conditions imply that the $2M$-th moment of this integrals behaves like a one-fold Wiener–Itô integral with the same variance.

Actually Latała formulated his result in a different form, and he proved a slightly weaker result. He considered Gaussian polynomials of the following form:

$$P(\xi_j^{(s)},\ 1 \le j \le n,\ 1 \le s \le k) = \frac{1}{k!} \sum_{(j_1,\dots,j_k)\colon 1 \le j_s \le n,\ 1 \le s \le k} a(j_1, \dots, j_k)\xi_{j_1}^{(1)} \cdots \xi_{j_k}^{(k)}, \tag{13.27}$$

where $\xi_j^{(s)}$, $1 \le j \le n$ and $1 \le s \le k$, are independent standard normal random variables. Latała gave an estimate about about the moments and tail-distribution of such random polynomials.

The problem about the behaviour of such random polynomials can be reformulated as a problem about the behaviour of Wiener–Itô integrals in the following way: Take a

measurable space $(X, \mathcal{X})$ with a non-atomic measure μ on it. Let Z_μ be a white noise with reference measure μ, let us choose a set of orthogonal functions $h_j^{(s)}(x)$, $1 \le j \le n$, $1 \le s \le k$, on the space $(X, \mathcal{X})$ with respect to the measure μ, and define the function

$$f(x_1, \dots, x_k) = \frac{1}{k!} \sum_{(j_1,\dots,j_k)\colon 1 \le j_s \le n,\, 1 \le s \le k} a(j_1, \dots, j_k) h_{j_1}^{(1)}(x_1) \cdots h_{j_k}^{(k)}(x_k) \qquad (13.28)$$

together with the Wiener–Itô integral $Z_{\mu,k}(f)$. Since the random integrals $\bar{\xi}_j^{(s)} = \int h_j^{(s)}(x) Z_\mu(\,dx)$, $1 \le j \le n$, $1 \le s \le k$, are independent, standard Gaussian random variables, it is not difficult to see with the help of Itô's formula (Theorem 10.3 in this work) that the distributions of the random polynomial $P(\xi_j^{(s)},\ 1 \le j \le n, 1 \le s \le k)$ and $Z_{\mu,k}(f)$ agree. Here we reformulated Latała's estimates about random polynomials of the form (13.27) to estimates about Wiener–Itô integrals with kernel function of the form (13.28).

These estimates are equivalent to Latała's result if we restrict our attention to the special class of Wiener–Itô integrals with kernel functions of the form (13.28). But we have formulated our result for Wiener–Itô integrals with a general kernel function. Latała's proof heavily exploits the special structure of the random polynomials given in (13.27), the independence of the random variables $\xi_j^{(s)}$ for different parameters s in it. (It would be interesting to find a proof which does not exploit this property.) On the other hand, this result can be generalized to the case discussed in Theorem 13.7. This generalization can be proved by exploiting the theorem of de la Peña and Montgomery Smith about the comparision of U-statistics and decoupled U-statistics (formulated in Theorem 14.3 of this work) and the properties of the Wiener–Itô integrals. I omit the details of the proof.

Latała also proved a converse estimate in [26] about random polynomials of Gaussian random polynomials which shows that the estimates of Theorem 13.7 are sharp. We formulate it in its original form, i.e. we restrict our attention to the case of Wiener–Itô integrals with kernel functions of the form (13.28).

Theorem 13.8. A lower bound about the tail distribution of Wiener–Itô integrals. *A random integral $Z_{\mu,k}(f)$ with a kernel function of the form (13.28) satisfies the inequalities*

$$E(Z_{\mu,k}(f))^{2M} \ge \left(C(k) \max_{1 \le s \le k} (M^{s/2} \alpha_s) \right)^{2M},$$

and

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

for all $M = 1, 2, \dots$ and $u > 0$ with some universal constant $C(k) > 0$ depending only on the order k of the integral and the quantities α_s, defined in formula (13.23) and (13.24).

Let me finally remark that there is a counterpart of Theorem 13.7 about degenerate U-statistics. Such a result can be found in paper [1] of Adamczak. Here we do not

discuss this result, because it is far from the main topic of this work. We only remark that some new quantities has to be introduce the formulate this result. The appearance of these conditions is related to the fact that in an estimate about the tail-behaviour of a degenerate U-statistic we need an estimate not only on the L_2-norm but also on the supremum norm of the kernel function. In a sharp estimate the estimate about the supremum of the kernel function has to be replaced by a more complex system of conditions, just as the condition about the L_2-norm of the kernel function was replaced by a condition about the quantities α_s, $1 \le s \le k$, defined in formulas (13.23) and (13.24) in Theorem 13.7.

14. 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 was shown in Section 9 that Theorem 8.2 follows from Theorems 8.4. Hence it is enough to prove Theorem 8.4. It may be useful to study this problem together with its multiple Wiener–Itô integral version, Theorem 8.6.

Theorems 8.6 and 8.4 will be proved similarly to their one-variate versions, Theorems 4.2 and 4.1. In the proof of Theorem 8.6 the estimates on the tail distribution of a Gaussian random variable has to be replaced by the estimate of Theorem 8.5 about the tail distribution of multiple Wiener–Itô integrals. After this the same chaining argument can be applied as in the proof of Theorem 4.2. No new difficulties arise. On the other hand, in the proof of Theorem 8.4 several new difficulties have to be overcome. I start with the proof of Theorem 8.6.

Proof of Theorem 8.6. Fix a number $0 < \varepsilon < 1$, and 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 conditions of Theorem 8.6, a set $\mathcal{F}_p = \{f_{a(1,p)}, \dots, f_{a(m_p,p)}\} \subset \mathcal{F}$ of function with $m_p \le 2D\, 2^{(2p+4)L} \varepsilon^{-L} \sigma^{-L}$ elements in such a way that $\inf\limits_{1 \le j \le m_p} \int (f - f_{a(j,p)})^2 \, d\mu \le 2^{-4p-8} \varepsilon^2 \sigma^2$ for all $f \in \mathcal{F}$ with some let $f_{a(j.p)} \in \mathcal{F}_p$, and beside this $f_p \in \mathcal{F}_p$. 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 $\int |f_{a(j,p)} - f_{a(j',p-1)}|^2 \, d\mu \le \varepsilon^2 \sigma^2 2^{-4(p+1)}$. Theorem 8.5 with the choice $\bar{u} = \bar{u}(p) = 2^{-(p+1)} \varepsilon u$ and $\bar{\sigma} = \bar{\sigma}(p) = 2^{-2p-2} \varepsilon \sigma$ yields the estimates

$$\begin{aligned} P(A(j,p)) &= P\left(n^{-k/2} k! |Z_{n,k}(f_{a(j,p)} - f_{a(j',p-1)})| \ge 2^{-(1+p)} \varepsilon u \right) \\ &\le C \exp\left\{ -\frac{1}{2} \left(\frac{2^{p+1} u}{\sigma} \right)^{2/k} \right\}, \qquad 1 \le j \le m_p, \end{aligned} \tag{14.1}$$

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

$$P(B(s)) = P\left(n^{-k/2} k! |Z_{n,k}(f_{a(0,s)})| \ge \left(1 - \frac{\varepsilon}{2}\right) u \right) \le C \exp\left\{ -\frac{1}{2} \left(\frac{\left(1 - \frac{\varepsilon}{2}\right) u}{\sigma} \right)^{2/k} \right\}, \quad 1 \le s \le m_0. \tag{14.2}$$

Since all $f \in \mathcal{F}$ is the element of at least one set $\mathcal{F}_p$, $p = 0, 1, 2, \dots$, $(f_p \in \mathcal{F}_p)$, the definition of the predecessor of an index $a(j,p)$ and of the events $A(j,p)$ and $B(s)$ in formulas (14.1) and (14.2) together with the previous estimates imply that

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}} n^{-k/2}k!|Z_{n,k}(f)| \ge u\right) &\le P\left(\bigcup_{p=1}^{\infty}\bigcup_{j=1}^{m_p} A(j,p) \cup \bigcup_{s=1}^{m_0} B(s)\right) \\ &\le \sum_{p=1}^{\infty}\sum_{j=1}^{m_p} P(A(j,p)) + \sum_{s=1}^{m_0} P(B(s)) \qquad (14.3) \\ &\le \sum_{p=1}^{\infty} 2CD2^{(2p+4)L}\varepsilon^{-L}\sigma^{-L}\exp\left\{-\frac{1}{2}\left(\frac{2^{p+1}u}{\sigma}\right)^{2/k}\right\} \\ &\quad + 2^5 CD\varepsilon^{-L}\sigma^{-L}\exp\left\{-\frac{1}{2}\left(\frac{\left(1-\frac{\varepsilon}{2}\right)u}{\sigma}\right)^{2/k}\right\}. \end{aligned}$$

Standard calculation shows that if $u \ge ML^{k/2}\varepsilon^{-k/2}\sigma\left(\log^{k/2}\frac{2}{\sigma} + \log^{k/2}\frac{1}{\varepsilon}\right)$ with a sufficiently large constant M, then the inequalities

$$2^{(2p+4)L}\varepsilon^{-L}\sigma^{-L}\exp\left\{-\frac{1}{2}\left(\frac{2^{p+1}u}{\sigma}\right)^{2/k}\right\} \le 2^{-p}\left\{-\frac{1}{2}\left(\frac{(1-\varepsilon)u}{\sigma}\right)^{2/k}\right\}$$

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

$$\varepsilon^{-L}\sigma^{-L}\exp\left\{-\frac{1}{2}\left(\frac{\left(1-\frac{\varepsilon}{2}\right)u}{\sigma}\right)^{2/k}\right\} \le \exp\left\{-\frac{1}{2}\left(\frac{(1-\varepsilon)\,u}{\sigma}\right)^{2/k}\right\}.$$

These inequalities together with relation (14.3) imply relation (8.15). Theorem 8.6 is proved.

The proof of Theorem 8.4 is harder. In this case the chaining argument in itself does not supply the proof, since Theorem 8.3 gives a good estimate about the distribution of a degenerate U-statistic only if it has a not too small variance. The same difficulty appeared in the proof of Theorem 4.1, and the method applied in that case will be adapted to the present situation.

A multivariate version of Proposition 6.1 will be proved in Proposition 14.1, and Theorem 8.4 will be reduced to a simpler statement formulated in Proposition 14.2 with its help. This result is the natural multivariate version of Proposition 6.2. Proposition 6.2 was proved by an appropriate induction procedure with the help of some symmetrization argument. This procedure will be adapted to the present case, but at this point some new difficulties arise.

The symmetrization argument in the one-variate case was based on a symmetrization lemma (Lemma 7.1), where the distribution of a sum of independent random variables with expectation zero was bounded by the distribution of the difference of two independent copies of this sum. There exists a multivariate version of this result about degenerate U-statistics, but it has some unpleasant properties. Instead of an independent copy of the original U-statistic such modified versions of it have to be considered, where the random variables in some arguments of the kernel function of the original U-statistic have to be replaced by an independent copy, while in other arguments of the kernel function the original random variables have to be preserved. Several such expressions have to be handled simultaneously, and this causes some problems. This difficulty can be slightly diminished by introducing so-called decoupled U-statistics and working with them. There is a result of de la Peña and Montgomery–Smith which enables us to reduce the estimation of U-statistics to the estimation of decoupled U-statistics. The behaviour of decoupled U-statistics is very similar to that of the original U-statistics, but the application of the multivariate version of the symmetrization argument is simpler when decoupled U-statics are considered. In the next section the problems arising at the adaptation of the symmetrization argument to the present problem will be explained in more detail.

The notion of decoupled U-statistics will be introduced, and by means of a result of de la Peña and Montgomery–Smith Proposition 14.2 will be reduced to a version of it, Proposition 14.2′, which states a similar estimate about decoupled U-statistics. This result of de la Peña and Montgomery–Smith will be proved in Appendix D.

A result formulated in Proposition 14.1 can be proved in almost the same way as its one-variate version, Proposition 6.1. The only essential difference is that now we have to apply a multivariate version of the Bernstein's inequality. Theorem 14.1 contains the information we can get by applying Theorem 8.3 together with the chaining argument. It has a similar structure to Proposition 6.1. Its main content, inequality (14.4), yields a good estimate on the supremum of degenerated U-statistics if the supremum is taken for an appropriate finite subclass $\mathcal{F}_{\bar{\sigma}}$ of the original class of kernel functions $\mathcal{F}$. The class of kernel functions $\mathcal{F}_{\bar{\sigma}}$ is a relatively dense subclass of $\mathcal{F}$ in the L_2 norm, and it also has some other good properties.

In the formulation of Proposition 14.1 two parameters $\bar{A} > 2^k$ and $M \geq M_0(\bar{A}, k)$ will be introduced. Their introduction may seem at first sight unnatural, they make the notation more complicated. But they turned out to be useful quantities, they help to fit the parameters in Propositions 14.1 and 14.2 when we want to apply them simultaneously to reduce Theorem 8.4 to Proposition 14.2.

Proposition 14.1. *Let the k-fold power $(X^k, \mathcal{X}^k)$ of a measurable space $(X, \mathcal{X})$ be given together 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, $L \geq 1$, whose elements satisfy the following properties. 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 $0 < \sigma \leq 1$. Take a sequence of independent μ-distributed random variables $\xi_1, \dots, \xi_n$, $n \geq \max(k, 2)$, 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 sufficiently large numbers $M \ge M_0(\bar{A}, k)$ the following relation (depending on the numbers $\bar{A}$ and M) holds: For all numbers $u > 0$ such that $n\sigma^2 \ge \left(\frac{u}{\sigma}\right)^{2/k} \ge ML \log \frac{2}{\sigma}$ a number $\bar{\sigma} = \bar{\sigma}(u)$, $0 \le \bar{\sigma} \le \sigma \le 1$, 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 $\bigcup_{j=1}^{m} \mathcal{D}_j = \mathcal{F}$, and the (degenerate) U-statistics $I_{n,k}(f)$, $f \in \mathcal{F}_{\bar{\sigma}(u)}$, satisfy the inequality

$$P\left(\sup_{f \in \mathcal{F}_{\bar{\sigma}(u)}} n^{-k/2} |I_{n,k}(f)| \ge \frac{u}{\bar{A}}\right) \le 2CD \exp\left\{-\alpha \left(\frac{u}{10\bar{A}\sigma}\right)^{2/k}\right\} \tag{14.4}$$

$$\text{if} \quad n\sigma^2 \ge \left(\frac{u}{\sigma}\right)^{2/k} \ge ML \log \frac{2}{\sigma}$$

with the constants $\alpha = \alpha(k)$, $C = C(k)$ appearing in formula (8.10′) of the Corollary 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 14.1. 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 $\mathcal{F}_p = \{f_{a(1,p)}, \dots, f_{a(m_p,p)}\} \subset \mathcal{F}$ with $m_p \le D\, 2^{2pL} \sigma^{-L}$ elements in such a way that $\inf_{1 \le j \le m_p} \int (f - f_{a(j,p)})^2 \, d\mu \le 2^{-4p}\sigma^2$ 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 the inequalities $\int \left(\frac{f_{a(j,p)} - f_{a(j',p-1)}}{2}\right)^2 d\mu \le 4\sigma^2 2^{-4p}$ and $\sup_{x_j \in X,\, 1 \le j \le k} \left|\frac{f_{a(j,p)}(x_1,\dots,x_k) - f_{a(j',p-1)}(x_1,\dots,x_k)}{2}\right| \le 1$ hold. The Corollary of Theorem 8.3 yields that

$$\begin{aligned} P(A(j,p)) &= P\left(n^{-k/2} |I_{n,k}(f_{a(j,p)} - f_{a(j',p-1)})| \ge \frac{2^{-(1+p)}u}{\bar{A}}\right) \\ &\le C \exp\left\{-\alpha \left(\frac{2^p u}{8\bar{A}\sigma}\right)^{2/k}\right\} \quad \text{if} \quad 4n\sigma^2 2^{-4p} \ge \left(\frac{2^p u}{8\bar{A}\sigma}\right)^{2/k}, \\ &\qquad 1 \le j \le m_p,\ p = 1, 2, \dots, \end{aligned} \tag{14.5}$$

and

$$P(B(s)) = P\left(n^{-k/2} |I_{n,k}(f_{0,s})| \ge \frac{u}{2\bar{A}}\right) \le C \exp\left\{-\alpha \left(\frac{u}{2\bar{A}\sigma}\right)^{2/k}\right\}, \quad 1 \le s \le m_0,$$

$$\text{if} \quad n\sigma^2 \ge \left(\frac{u}{2\bar{A}\sigma}\right)^{2/k}. \tag{14.6}$$

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 $\mathcal{F}_{\bar{\sigma}} = \mathcal{F}_R$ (this is the class of functions $\mathcal{F}_p$ introduced at the start of the proof 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.) The cardinality m of the set $\mathcal{F}_{\bar{\sigma}}$ is clearly not greater than $D\bar{\sigma}^{-L}$, and $\bigcup_{j=1}^{m} \mathcal{D}_j = \mathcal{F}$. Beside this, the number R was chosen in such a way that the inequalities (14.5) and (14.6) hold for $1 \le p \le R$. Hence the definition of the predecessor of an index $a(j,p)$ implies that

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

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

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

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

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

and relation (14.4) holds.

The relations

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

hold. Hence $n\bar{\sigma}^2 \le 4\left(\frac{u}{\bar{A}\bar{\sigma}}\right)^{2/k}$. Beside this, 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 14.1 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 $2^{(4+2/k)R}\left(\frac{u}{\bar{A}\sigma}\right)^{2/k} \le 2^{2+6/k}n\sigma^2$ can be applied. This implies that $2^{-4R} \ge 2^{-4(2+6/k))/(4+2/k)}\left[\frac{\left(\frac{u}{\bar{A}\sigma}\right)^{2/k}}{n\sigma^2}\right]^{4/(4+2/k)}$, 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 $n\bar{\sigma}^2 \ge \frac{\bar{A}^{-4/3}}{50}(n\sigma^2)^{1/3}\left[\left(\frac{u}{\sigma}\right)^{2/k}\right]^{2/3} \ge \frac{\bar{A}^{-4/3}}{50}n^{1/9}\left(\frac{M}{3}\right)^{2/3}(L+\beta) \ge \frac{M^{2/3}(L+\beta)\log n}{1000\bar{A}^{4/3}}$.

A multivariate analog of Proposition 6.2 is formulated in Proposition 14.2, and it will be shown that Propositions 14.1 and 14.2 imply Theorem 8.4.

Proposition 14.2. *Let a probability measure μ be given 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 (with respect to the measure μ) kernel functions $f = f(x_1,\dots,x_k)$ with some parameter D and exponent L on the product space $(X^k,\mathcal{X}^k)$. Let all functions $f \in \mathcal{F}$ satisfy conditions (8.1) and (8.2) with some $0 < \sigma \le 1$. Let us 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*

$$P\left(\sup_{f\in\mathcal{F}}|n^{-k/2}I_{n,k}(f)| \ge An^{k/2}\sigma^{k+1}\right) \le \bar{C}e^{-\gamma A^{1/2k}n\sigma^2} \quad \text{if } A \ge A_0. \tag{14.7}$$

Proposition 14.2 yields an estimate for the tail distribution of the supremum of degenerate U-statistics at level $u \ge A_0n^{k/2}\sigma^{k+1}$, i.e. in the case when Theorem 8.3 does

not give a good estimate on the tail-distribution of the single degenerate U-statistics taking part in the supremum at the left-hand side of (14.7).

Formula (8.11) will be proved by means of Proposition 14.2 with the choice $\sigma = \bar\sigma = \bar\sigma(u)$ defined in Proposition 14.1 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 14.1, where f_j is the function appearing in the definition of the class of functions $\mathcal{D}_j$. Clearly,

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}} n^{-k/2} |I_{n,k}(f)| \ge u \right) &\le P\left(\sup_{f\in\mathcal{F}_{\bar\sigma}} n^{-k/2} |I_{n,k}(f)| \ge \frac{u}{\bar A} \right) \\ &\quad + \sum_{j=1}^m P\left(\sup_{g\in\mathcal{D}_j} n^{-k/2} \left| I_{n,k}\left(\frac{f_j - g}{2} \right) \right| \ge \left(\frac{1}{2} - \frac{1}{2\bar A} \right) u \right), \end{aligned} \tag{14.8}$$

where m is the cardinality of the set of functions $\mathcal{F}_{\bar\sigma}$ appearing in Proposition 14.1. We want to show that if first $\bar A$ and then $M \ge M_0(\bar A, k)$ are chosen sufficiently large in Proposition 14.1, then the second term at the right-hand side of formula (14.8) can be well bounded by means of Proposition 14.2, 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 $\gamma \bar A_0^{1/2k} \ge \frac{1}{K}$ with the numbers A_0, K and γ in Proposition 14.2, put $\bar A = \max(2^{k+2}\bar A_0, 2^k)$, and apply Proposition 14.1 with this number $\bar A$. Then by Proposition 14.1 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 $u \ge 4\bar A_0 n^{k/2}\bar\sigma^{k+1}$ with the number $\bar\sigma$ in Proposition 14.1. This implies that $\left(\frac{1}{2} - \frac{1}{2\bar A}\right) u \ge \frac{u}{4} \ge \bar A_0 n^{k/2}\bar\sigma^{k+1}$, $\bar A_0 \ge A_0$, and by replacing the expression $\left(\frac{1}{2} - \frac{1}{2\bar A}\right) u$ by $\bar A_0 n^{k/2}\bar\sigma^{k+1}$ in the probabilities of the sum in the second term at the right-hand side of (14.8) we enlarge them.

The numbers u considered in these estimations satisfy the condition $n\sigma^{2/k} \ge \left(\frac{u}{\sigma}\right)^{2/k} \ge M(L+\beta)^{3/2} \log \frac{2}{\sigma}$ imposed in Proposition 14.1 with some appropriately chosen constant M. Choose the number $M \ge M_0(\bar A, k)$ in Proposition 14.1 (it can also play the role of the number M in formula (8.11) 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 14.2. 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 14.2 can be applied to bound the terms in the sum at the right-hand side of (14.8). It yields the estimate

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

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 14.1 (relation (14.4) together

with the inequality $m \le D\bar{\sigma}^{-L}$) and formula (14.7) with $A = \bar{A}_0$ imply that

$$P\left(\sup_{f\in\mathcal{F}} n^{-k/2}|I_{n,k}(f)| \ge u\right) \le 2CD\exp\left\{-\alpha\left(\frac{u}{10\bar{A}\sigma}\right)^{2/k}\right\} + \bar{C}D\bar{\sigma}^{-L}e^{-\gamma\bar{A}_0^{1/2k}n\bar{\sigma}^2}. \tag{14.9}$$

To get the result of Theorem 8.4 from inequality (14.9) its second term at the right-hand side has to be replaced by a more appropriate expression where, in particular, the coefficient $\bar{\sigma}^{-L}$ disappears. 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 $\gamma\bar{A}_0^{1/2k}n\bar{\sigma}^2 \ge \frac{1}{K}n\bar{\sigma}^2 \ge L\log n \ge 2L\log\frac{1}{\bar{\sigma}}$, i.e. $\bar{\sigma}^{-L} \le e^{\gamma\bar{A}_0^{1/2k}n\bar{\sigma}^2/2}$. By the estimates of Proposition 14.1 $n\bar{\sigma}^2 \ge \frac{1}{64}\left(\frac{u}{\bar{A}\sigma}\right)^{2/k}$. 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 (14.9) yields that

$$P\left(\sup_{f\in\mathcal{F}} n^{-k/2}|I_{n,k}(f)| \ge u\right)$$
$$\le 2CD\exp\left\{-\frac{\alpha}{(10\bar{A})^2}\left(\frac{u}{\sigma}\right)^{2/k}\right\} + \bar{C}D\exp\left\{-\frac{\gamma}{128}\bar{A}_0^{1/2k}\bar{A}^{-2/k}\left(\frac{u}{\sigma}\right)^{2/k}\right\},$$

and this estimate implies Theorem 8.4.

Thus to complete the proof of Theorem 8.4 it is enough to prove Proposition 14.2, which is a multivariate analog of Proposition 6.2. The proof of Proposition 6.2 was based on a symmetrization argument. This argument can be better applied in the solution of the present problem if we work with so-called decoupled U-statistics. I introduce this notion together with its randomized version which will be useful in the subsequent part, where some symmetrization arguments will be applied.

Similarly to the one-variate case we shall also apply symmetrization type argument in the study of U-statistics. This will make possible the reduction of the estimation of (decoupled) U-statistics given in formula (14.10) to the estimation of randomized (decoupled) U-statistics given in formula (14.11).

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 in 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. The decoupled U-statistic $\bar{I}_{n,k}(f)$ 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*

$$\bar{I}_{n,k}(f) = \frac{1}{k!}\sum_{\substack{(l_1,\dots,l_k):\ 1\le l_j\le n,\ j=1,\dots,k,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} f\left(\xi_{l_1}^{(1)},\dots,\xi_{l_k}^{(k)}\right). \tag{14.10}$$

Let us have beside 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* $\xi_1^{(j)}, \dots, \xi_n^{(j)}$, $1 \le j \le k$. *The randomized decoupled* U*-statistic* $\bar{I}_{n,k}(f, \varepsilon)$ *(depending on the random sequences* $\xi_1^{(j)}, \dots, \xi_n^{(j)}$, $1 \le j \le k$, *the kernel function* f *and the randomizing sequence* $\varepsilon_1, \dots, \varepsilon_n$*) is defined by the formula*

$$\bar{I}_{n,k}^{\varepsilon}(f) = \frac{1}{k!} \sum_{\substack{(l_1,\dots,l_k)\colon 1 \le l_j \le n,\ j=1,\dots,k, \\ l_j \ne l_{j'} \text{ if } j \ne j'}} \varepsilon_{l_1} \cdots \varepsilon_{l_k} f\left(\xi_{l_1}^{(1)}, \dots, \xi_{l_k}^{(k)}\right). \tag{14.11}$$

A decoupled or randomized decoupled U-statistics (with real valued kernel function) will be called degenerate if its kernel function is canonical. This terminology is in full accordance with the definition of (usual) degenerate U-statistics.

A result of de la Peña and Montgomery–Smith will be formulated below. It gives an upper bound for the tail distribution of a U-statistic by means of the tail distribution of an appropriate decoupled U-statistic. It also has a generalization, where the supremum of U-statistics are bounded by the supremum of decoupled U-statistics. The theorem of de la Peña and Montgomery–Smith will be proved in Appendix D. It enables us to reduce Proposition 14.2 to a version formulated Proposition 14.2′, where the supremum of decoupled U-statistics has to be bounded. This problem is simpler than the original one.

Before the formulation of the theorem of de la Peña and Montgomery–Smith I make some remark about it. It considers more general U-statistics with kernel functions taking values in a separable Banach space, and it compares the norm of Banach space valued U-statistics and decoupled U-statistics. (Decoupled U-statistics were defined with general Banach space valued kernel functions, and the definition of U-statistics can also be generalized to separable Banach space valued kernel functions in a natural way.) This result was formulated in such a general form for a special reason. Its more general form helped a general form for a special reason. Its more general form helped to derive formula (14.13) of the subsequent theorem from formula (14.12). It can be exploited in the proof of formula (14.13) that the constants in the estimate (14.12) do not depend on the Banach space, where the kernel function f takes its values.

Theorem 14.3. (Theorem of de la Peña and Montgomery–Smith about the comparison of U-statistics and decoupled U-statistics). *Let us consider a sequence of independent and identically distributed random variables* $\xi_1, \dots, \xi_n$ *with values in a measurable space* $(X, \mathcal{X})$ *together with* k *independent copies* $\xi_1^{(j)}, \dots, \xi_n^{(j)}$, $1 \le j \le k$, *of this sequence. 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 in a separable Banach space* B. *Let us take 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. *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*

$$P\left(\|I_{n,k}(f)\| > u\right) \le \bar{C}P\left(\|\bar{I}_{n,k}(f)\| > \gamma u\right) \tag{14.12}$$

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*

$$P\left(\sup_{1\le s<\infty} \|I_{n,k}(f_s)\| > u\right) \le \bar{C}P\left(\sup_{1\le s<\infty} \left\|\bar{I}_{n,k}(f_s)\right\| > \gamma u\right). \tag{14.13}$$

Now I formulate the following version of Proposition 4.2.

Proposition 14.2′**.** *Let a probability measure* μ *be given 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 (with respect to the measure* μ*) kernel functions* $f = f(x_1, \dots, x_k)$ *with some parameter* D *and exponent* L *on the product space* $(X^k, \mathcal{X}^k)$*. Let all functions* $f \in \mathcal{F}$ *satisfy conditions (8.1) and (8.2) with some* $0 < \sigma \le 1$*. 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 in formula (14.10).*

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* $I_{n,k}(f)$*,* $f \in \mathcal{F}$*, 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 (14.7):*

$$P\left(\sup_{f\in\mathcal{F}} n^{-k/2}|\bar{I}_{n,k}(f)| \ge An^{k/2}\sigma^{k+1}\right) \le e^{-\gamma A^{1/2k} n\sigma^2} \quad \text{if } A \ge A_0. \tag{14.14}$$

It is clear that Proposition 14.2′ and Theorem 14.3, more explicitly formula (14.13) in it imply Proposition 14.2. Hence the proof of Theorem 8.4 was reduced to Proposition 14.2′ in this section. The proof of Proposition 14.2′ is based on a symmetrization argument. Its main ideas will be explained in the next section.

15. The strategy of the proof for the main result of this work.

In the previous section the proof of Theorem 8.4 was reduced to that of Proposition 14.2′. Proposition 14.2′ is a multivariate version of Proposition 6.2, and its proof is based on similar ideas. Proposition 6.2 was proved by means of Proposition 7.3 in which an inductive procedure was carried out. In the proof of Proposition 14.2′ this argument is applied in a more sophisticated situation. To understand how to apply it let us first observe that relation (14.14) in Proposition 14.2′ holds for $A > A_0 = \sigma^{-(k+1)}$ if the absolute value of the functions in the class $\mathcal{F}$ is bounded in the supremum norm by 1. Indeed, in this case $n^{-k/2}|\bar{I}_{n,k}(f)| \le n^{k/2}$ with probability 1 for all functions $f \in \mathcal{F}$, hence the probability at the left-hand side of (14.14) equals zero. A most important step of the proof of Proposition (14.2′) will be the right formulation of an appropriate inductive argument in Proposition 15.3, which is a natural multivariate anagogue of Proposition 7.2. In this proposition such classes of function $\mathcal{F}$ are considered which satisfy some nice properties. It will be shown that if all classes of functions with these properties satisfy relation (14.14) for all numbers $A \ge A_0$ with some sufficiently large threshold A_0, then they also satisfy this relation with a smaller threshold $A_0' < A_0$. The detailed formulation of Proposition 15.3 will be given later.

By a successive application of Proposition 15.3 it can be shown that the estimate (14.14) holds not only for $A_0 = \sigma^{-(k+1)}$, but also for a much smaller threshold number A_0. Proposition 14.2′ contains the estimate that can be obtained with the help of Proposition 15.3. This proposition is a natural multivariate version of Proposition 7.3, and also their proofs are similar. However, there is an essential difference between them. The proof of Proposition 7.3 contains a relatively simple symmetrization argument, but it is not clear how to adapt it to the case of Proposition 15.3. The greatest difficulties in the proof appeared at this point. In particular, the proof of Proposition 15.3 demanded the formulation of another inductive argument presented in Proposition 15.4. These two Propositions will be proved simultaneously.

This section contains the formulation of Propositions 15.3 and 15.4 together with two lemmas which are useful in their proof. It is also shown that Proposition 14.2′ follows from these Propositions.

To understand the difficulty about the adaptation of the symmetrization argument to the new case better it is useful to recall how it was applied in the proof of Proposition 6.2. In that result the supremum of a class of sums $\sum_{j=1}^{n} f(\xi_j)$ had to be bounded, where ξ_j, $1 \le j \le n$, were independent and identically distributed random variables, and the supremum was taken for the elements f of a nice class of functions $\mathcal{F}$. In the symmetrization argument of Proposition 6.2 this problem was reduced to the estimation of the supremum of appropriately randomized versions of the above sums. More explicitly, in the new problem the supremum of sums of the form $\sum_{j=1}^{n} \varepsilon_j f(\xi_j)$ was considered, where ε_j, $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = \frac{1}{2}$, are independent random variables, independent also of the random variables ξ_j, $1 \le j \le n$.

In the proof of Proposition 14.2′ we want to find such a multivariate version of this

argument where the estimation of the supremum of decoupled U-statistics defined in formula (14.10) is reduced to the estimation of the supremum of randomized decoupled U-statistics defined in formula (14.11). We want to do this by means of an appropriate adaptation of the method applied in the one-variate case.

The symmetrization argument in the proof of Proposition 6.2 had two important ingredients. The first one was the observation that if an independent copy $\xi_1', \dots, \xi_n'$ of the original random variables $\xi_1, \dots, \xi_n$ is taken, which is independent also of the randomizing sequence $\varepsilon_1, \dots, \varepsilon_n$, then the joint distribution of the sums $\sum_{j=1}^n [f(\xi_j) - f(\xi_j')]$ (depending on the class $\mathcal{F}$ of the functions f) and of the sums $\sum_{j=1}^n \varepsilon_j [f(\xi_j) - f(\xi_j')]$ agree. This was the step, where the randomizing terms ε_j appeared. The other ingredient was Lemma 7.1 that enabled us to compare the supremum of the sums $\sum_{j=1}^n [f(\xi_j) - f(\xi_j')]$ and $\sum_{j=1}^n f(\xi_j)$. Lemma 15.1 formulated below can be considered as the multivariate version of the first step in this proof. To formulate it some notations have to be introduced.

Let $\mathcal{V}_k$ denote the set of all sequences $\varepsilon(1), \dots, \varepsilon(k)\}$ of length k such that $\varepsilon(j) = +1$ or $\varepsilon(j) = -1$ for all $1 \le j \le k$. Let $m(v)$, $v = (\varepsilon(1), \dots, \varepsilon(k)) \in \mathcal{V}_k$, denote the number of digits -1 in the sequence v. Let a (real valued) function $f(x_1, \dots, x_k)$ of k variables be given on a measurable space $(X, \mathcal{X})$ together with a sequence of independent and identically distributed random variables $\xi_1, \dots, \xi_n$ with values in the space $(X, \mathcal{X})$ and $2k$ independent copies $\xi_1^{(j,1)}, \dots, \xi_n^{(j,1)}$ and $\xi_1^{(j,-1)}, \dots, \xi_n^{(j,-1)}$, $1 \le j \le k$, of this sequence. Let us have beside them another sequence $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$, $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = \frac{1}{2}$, of independent random variables, also independent of all previously introduced random variables. With the help of the above quantities we introduce the random variables

$$\tilde{I}_{n,k}(f) = \frac{1}{k!} \sum_{v \in \mathcal{V}_k} (-1)^{m(v)} \sum_{\substack{(l_1,\dots,l_k)\colon 1 \le l_r \le n,\ r=1,\dots,k, \\ l_r \ne l_{r'} \text{ if } r \ne r'}} f\left(\xi_{l_1}^{(1,v(1))}, \dots, \xi_{l_k}^{(k,v(k))}\right) \tag{15.1}$$

and

$$\tilde{I}_{n,k}^{\varepsilon}(f) = \frac{1}{k!} \sum_{v \in \mathcal{V}_k} (-1)^{m(v)} \sum_{\substack{(l_1,\dots,l_k)\colon 1 \le l_r \le n,\ r=1,\dots,k, \\ l_r \ne l_{r'} \text{ if } r \ne r'}} \varepsilon_{l_1} \cdots \varepsilon_{l_k} f\left(\xi_{l_1}^{(1,v(1))}, \dots, \xi_{l_k}^{(k,v(k))}\right) \tag{15.2}$$

The number $m(v)$ in the above formulas denotes the number of the digits -1 in the ± 1 sequence v of length k, hence it counts how many random variables $\xi_{l_j}^{(j,1)}$, $1 \le j \le k$, were replaced by the 'secondary copy' $\xi_{l_j}^{(j,-1)}$ for a $v \in \mathcal{V}_k$ in the inner sum in formulas (15.1) or (15.2).

The following result holds.

Lemma 15.1. *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 $\tilde{I}^{\varepsilon}_{n,k}(f)$ defined in formulas (15.1) and (15.2) 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 $\{\tilde{I}^{\varepsilon}_{n,k}(f);\ f \in \mathcal{F}\}$ agree.*

Formulas (15.1) and (15.2) show some similarity to the formula by which Stieltjes measures are defined in the k-dimensional space by means of a functions of k variables.

Proof of Lemma 15.1. I even claim that for any fixed sequence $u = (u(1), \dots, u(n))$, $u(l) = \pm 1$, $1 \le l \le n$, of length n, the conditional distribution of the field $\{\tilde{I}^{\varepsilon}_{n,k}(f);\ f \in \mathcal{F}\}$ 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 (15.1) are functions of a random vector with 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 the coordinates $(\xi_l^{(j)}, \bar{\xi}_l^{(j)}) = (\xi_l^{(j,1)}, \xi_l^{(j,-1)})$ with such indices (l, j) for which $u(l) = -1$ (and the index j is arbitrary) are replaced by $(\bar{\xi}_l^{(j)}, \xi_l^{(j)}) = (\xi_l^{(j,-1)}, \xi_l^{(j,1)})$, and the coordinates $(\xi_l^{(j)}, \bar{\xi}_l^{(j)})$ with such indices (l, j) for which $u(l) = 1$ are not changed. As a consequence, we carry out a measure preserving transformation by replacing the original vector $(\xi_l^{(j)}, \bar{\xi}_l^{(j)})$, $1 \le l \le n$, $1 \le j \le k$, in the definition of the expression $\tilde{I}_{n,k}(f)$ in (15.1) for all $f \in \mathcal{F}$ by this modified vector. On the other hand, I claim that the distribution of the random field we get by means of the above transformation of the field $\tilde{I}_{n,k}(f)$, $f \in \mathcal{F}$, agrees with the conditional distribution of the random field $\tilde{I}^{\varepsilon}_{n,k}(f)$, $f \in \mathcal{F}$, defined in (15.2) 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 $\tilde{I}^{\varepsilon}_{n,k}(f)$, $f \in \mathcal{F}$, 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 $\tilde{I}^{\varepsilon}_{n,k}(f)$. 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. Lemma 15.1 is proved.

In the proof of Proposition 6.2 which can be considered the one-variate version of Proposition 14.2′ we needed a symmetrization Lemma, formulated in Lemma 7.2. This lemma cannot be applied in the present case, because its (independence type) conditions do not hold in the problem we are considering. Hence we need a new, generalized version of this result which will be formulated in Lemma 15.2. It can be applied in the proof of Proposition 14.2′, too.

Lemma 15.2. (Generalized version of the Symmetrization Lemma.) *Let Z_p and $\bar{Z}_p$, $p = 1, 2, \dots$, 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

$$P(|\bar{Z}_p| \le \alpha|\mathcal{B})(\omega) \ge \beta \quad \textit{for all } p = 1, 2, \dots \textit{ if } \omega \in B \tag{15.3}$$

holds. Then

$$P\left(\sup_{1\le p<\infty} |Z_p| > \alpha + u\right) \le \frac{1}{\beta} P\left(\sup_{1\le p<\infty} |Z_p - \bar{Z}_p| > u\right) + (1 - P(B)) \quad \textit{for all } u > 0. \tag{15.4}$$

Proof of Lemma 15.2. Put $\tau = \min\{p\colon |Z_p| > \alpha + u)$ if there exists such an index $p \ge 1$, and put $\tau = 0$ otherwise. Then

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

Hence

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

Thus Lemma 15.2 is proved.

The proof of Lemma 15.2 was relatively simple, but its application may cause some problems. The main difficulty is to check condition (15.3) which is an analogue of condition (7.1) in Lemma 7.1. Let us recall that in condition (15.3) a conditional probability with respect to such a σ-algebra $\mathcal{B}$ is bounded in an appropriate way, for which the random variables Z_p, $p = 1, 2, \dots$, are $\mathcal{B}$-measurable. So the behaviour of the conditionial probability we have to estimate in an application of Lemma 15.2 also depends on the relation between the random variables Z_p and $\bar{Z}_p$.

Proposition 14.2′ will be proved by means of the following program. A backward induction procedure will be applied with the help of Proposition 15.3 formulated below which implies formula (14.14). To formulate Proposition 15.3 first we introduce a notion that we call the *good tail behaviour for a class of decoupled U-statistics*. Proposition 15.3 is a natural multivariate analog of Proposition 7.3. It will be proved by means of a symmetrization argument with the help of Lemma 15.2. This lemma will be applied with some appropriately defined random variables Z_p, $\bar{Z}_p$ and σ-algebra $\mathcal{B}$. In the proof

of Proposition 15.3 formula (15.3) has to be checked in an appropriate setting. This will be done by means of a good estimate of the conditional second moments $E(\bar{Z}_p^2|\mathcal{B})$.

In the proof of Proposition 7.3 a similar method was applied. In that proof formula (7.1) was checked for some random variables $\bar{Z}_n$ by means of the estimation of their second moments $E\bar{Z}_n^2$. But while the estimation of the second moments needed in the proof of formula (7.1) was simple, the estimation of the conditional second moments needed in the proof of Proposition 15.3 is the most difficult part of the proof. To carry out this estimation a new notion will be introduced under the name *good tail behaviour for a class of integrals of decoupled U-statistics*. We shall formulate another result in Proposition 15.4 which is related to this notion. Propositions 15.3 and 15.4 will be proved simultaneously.

To formulate Propositions 15.3 and 15.4 the following two notions will be introduced.

Definition of good tail behaviour for a class of decoupled U-statistics. *Let some measurable space $(X, \mathcal{X})$ be given together with a probability measure μ on it. Let us consider some countable 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 a 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 (14.10). 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 are fixed in the sequel) if*

$$P\left(\sup_{f\in\mathcal{F}} |n^{-k/2}\bar{I}_{n,k}(f)| \geq An^{k/2}\sigma^{k+1}\right) \leq \exp\left\{-A^{1/2k}n\sigma^2\right\} \quad \text{for all } A > T. \tag{15.5}$$

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 $n \geq k$ and a positive number $0 < \sigma \leq 1$, and consider some countable 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 $\xi_1^{(j)}, \dots, \xi_n^{(j)}$, $1 \leq j \leq k$, 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 $\xi_1^{(j)}, \dots, \xi_n^{(j)}$, $1 \leq j \leq k$, the kernel function $f_y(x_1, \dots, x_k) = f(x_1, \dots, x_k, y)$ and formula (14.10). Define with the help of these U-statistics $\bar{I}_{n,k}(f, y)$ the random integrals*

$$H_{n,k}(f) = \int \bar{I}_{n,k}(f, y)^2 \rho(\, dy), \quad f \in \mathcal{F}. \tag{15.6}$$

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

$$P\left(\sup_{f\in\mathcal{F}} n^{-k}H_{n,k}(f) \ge A^2 n^k \sigma^{2k+2}\right) \le \exp\left\{-A^{1/(2k+1)}n\sigma^2\right\} \quad \text{for all } A > T. \quad (15.7)$$

Propositions 15.3 and 15.4 will be formulated with the help of the above notions.

Proposition 15.3. *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 $L \ge 1$ and parameter D. Let us also assume that all functions $f \in \mathcal{F}$ satisfy the conditions $\sup_{x_j\in X, 1\le j\le k} |f(x_1, \dots, x_k)| \le 2^{-(k+1)}$, $\int f^2(x_1, \dots, x_k)\mu(\,dx_1)\dots\mu(\,dx_k) \le \sigma^2$, and $n\sigma^2 > K(L+\beta)\log n$ with an appropriately chosen fixed number $K = K(k)$ with $\beta = \max\left(\frac{\log D}{\log n}, 0\right)$.*

Choose the constant $K = K(k)$ (and the sample size n) in the condition $n\sigma^2 > K(L+\beta)\log n$ sufficiently large. Then there is some real number $A_0 = A_0(k) > 1$ such that if for all classes of functions $\mathcal{F}$ which satisfy the above conditions the sets of decoupled U-statistics $\bar{I}_{n,k}(f)$, $f \in \mathcal{F}$, 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 15.4. *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*

$$\sup_{x_j\in X, 1\le j\le k, y\in Y} |f(x_1, \dots, x_k, y)| \le 2^{-(k+1)} \quad (15.8)$$

and

$$\int f^2(x_1, \dots, x_k, y)\mu(\,dx_1)\dots\mu(\,dx_k)\rho(\,dy) \le \sigma^2 \quad \text{for all } f \in \mathcal{F}. \quad (15.9)$$

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

Then 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 (15.6) 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 15.4 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

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

Let me also remark that the estimate (15.7) we have imposed in the definition of the property '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 tail distribution of $\sup_{f \in \mathcal{F}} n^{-k} H_{n,k}(f)$ behaves similarly to that of const. $\left(\sigma \eta^k\right)^2$, where η is a standard normal random variable. Formula (15.7) 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. Formula (15.5) in the formulation of the property 'good tail behaviour for a class of decoupled U-statistics' has a similar interpretation. It says that $\sup_{f \in \mathcal{F}} |n^{-k/2} I_{n,k}(f)|$ behaves similarly to const. $\sigma|\eta^k|$ with a standard normal random variable η.

We wanted 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 Proposition 15.4 we have imposed beside formula (15.8) a fairly weak condition (15.9) about the L_2-norm of the function f. Most difficulties appear in the proof, because we did not want to work with a more restrictive condition.

It is not difficult to derive Proposition 14.2′ from Proposition 15.3. Indeed, let us observe that the set of decoupled U-statistics determined by a class of functions $\mathcal{F}$ satisfying the conditions of Proposition 15.3 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 (15.5) equals zero for $A > \sigma^{-(k+1)}$. Then we get from Proposition 15.3 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 15.3, 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 15.3, 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 14.2′, only it has to be applied 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 15.4 formulated below. Actually, we shall need this corollary of Proposition 15.4.

Corollary of Proposition 15.4. *If the class of functions $\mathcal{F}$ satisfies the conditions of Proposition 15.4, 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 (15.6) have a good tail behaviour at level $\bar{A}_0$.*

The main difficulty in the proof of Proposition 15.3 arises in the application of the symmetrization procedure corresponding to Lemma 7.2 in the one-variate case. This difficulty can be overcome by means of Proposition 15.4, more precisely by means of its corollary. It helps us to estimate the conditional variances of the decoupled U-statistics we have to handle in the proof of Proposition 15.3. The proof of Propositions 15.3 and 15.4 apply similar arguments, and they will be proved simultaneously. The following inductive procedure will be applied in their proof. First Proposition 15.3 and then Proposition 15.4 is proved for $k = 1$. If Propositions 15.3 and 15.4 are already proved for all $k' < k$ for some number k, then first we prove Proposition 15.3 and then Proposition 15.4 for this number k.

The proof both of Proposition 15.3 and 15.4 applies a symmetrization argument that will be proved in the next section. In the subsequent section Propositions 15.3 and 15.4 will be proved with its help. They imply Proposition 14.2′, hence also Theorem 8.4.

16. A symmetrization argument.

The proof of Propositions 15.3 and 15.4 applies some ideas similar to the argument in the proof of Proposition 6.2. But here some additional technical difficulties have to be overcome. As a first step, two results formulated in Lemma 16.1A and 16.1B will be proved. They can be considered as a symmetrization argument analogous to Lemma 7.2 in the proof of Propositions 6.2. Lemma 16.1A will be applied in the proof of Proposition 15.3 and Lemma 16.1B in the proof of Proposition 15.4. This section contains their proofs. Because of the inductive structure of our proofs we may assume in their proof for parameter k that Propositions 15.3 and 15.4 (and their consequences) hold for $k' < k$.

Lemma 16.1A is a natural multivariate version of Lemma 7.2. Lemma 7.2 enables us to replace the estimation of the distribution of the supremum of a class of sums of independent random variables to the estimation of the distribution of the supremum of the randomized version of these sums. Lemma 16.1 enables to reduce the estimation of the supremum of degenerate U-statistics to the estimation of the distribution of the distribution of the supremum of the randomized degenerate U-statistics corresponding to them. The supremum of the randomized degenerate U-statistics we have to bound to prove Theorem 15.3 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 13.3. The case of Lemma 16.1B is more complicated. In this result the probability investigated in Proposition 15.4 is bounded by the distribution of the supremum of some random

variables $\bar{W}(f)$, $f \in \mathcal{F}$, which will be defined in formula (16.7). The expressions $\bar{W}(f)$ are rather complicated, and it is worth while to study them more closely. This will be done in the proof of Corollary of Lemma 16.1B which yields a more appropriate bound for the expression we want to estimate in Proposition 15.4. This corollary will be applied in the sequel.

The proof of Lemmas 16.1A and 16.1B is similar to that of Lemma 7.2. First we introduce k additional independent copies $\bar{\xi}_n^{(j)}, \dots, \bar{\xi}_n^{(j)}$ of the k (independent and identically distributed) 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 16.1A and 16.1B. This statement will be formulated and proved in Lemmas 16.2A and 16.2B. 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 16.3A and 16.3B. In Lemma 16.3A the conditional variance of a random variable is estimated under some appropriate conditions. This estimate together with the generalized form of the symmetrization Lemma, Lemma 15.2, enable us to prove Lemma 16.1A. Lemma 16.1B can be proved similarly, but here the conditional distribution of a more complicated expression has to be estimated. This estimate can be proved with the help of Lemma 16.3B. In Lemma 16.3B the conditional expectation of the absolute value of an appropriate expression is bounded.

The main results of this section are the following two lemmas.

Lemma 16.1A. *Let $\mathcal{F}$ be a class of functions on the space $(X^k, \mathcal{X}^k)$ which satisfies the conditions of Proposition 15.3 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 $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$, $P(\varepsilon_l = 1) = P(\varepsilon_l = -1) = \frac{1}{2}$, $1 \le l \le n$, 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 (14.10) together with their randomized version $\bar{I}^{\varepsilon}_{n,k}(f)$ defined in formula (14.11).*

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

$$P\left(\sup_{f\in\mathcal{F}} n^{-k/2}\left|\bar{I}_{n,k}(f)\right| > An^{k/2}\sigma^{k+1}\right) < 2^{k+1}P\left(\sup_{f\in\mathcal{F}}\left|\bar{I}^{\varepsilon}_{n,k}(f)\right| > 2^{-(k+1)}An^k\sigma^{k+1}\right) + 2^k n^{k-1}e^{-A^{1/(2k-1)}n\sigma^2/k} \tag{16.1}$$

holds for all $A \ge A_0$.

It may be worth remarking that the second term at the right-hand side of formula (16.1) yields a small contribution to the upper bound in this relation because of the condition $n\sigma^2 \ge K(L+\beta)\log n$ with a sufficiently large constant $K = K(k)$.

To formulate Lemma 16.1B first some new quantities have to be introduced. Some of them will be used somewhat later. The quantities $\bar{I}^V_{n,k}(f, y)$ introduced in the subsequent formula (16.2) depend on the sets $V \subset \{1, \dots, k\}$, and they are the natural

adaptations of the inner sum terms in formula (15.1). Such expressions are needed in the formulation of the symmetrization result applied in the proof of Proposition 15.4. Their randomized versions $\bar{I}_{n,k}^{(V,\varepsilon)}(f,y)$, introduced in formula (16.5), correspond to the inner sum terms in formula (15.2). The integrals of these expressions will be also introduced in formulas (16.3) and (16.6).

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 15.4. 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 $(\varepsilon_1,\dots,\varepsilon_n)$, $P(\varepsilon_l=1)=P(\varepsilon_l=-1)=\frac{1}{2}$, $1\le l\le n$, 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

$$\bar{I}_{n,k}^{V}(f,y)=\frac{1}{k!}\sum_{\substack{(l_1,\dots,l_k)\colon 1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if }j\ne j'}} f\left(\xi_{l_1}^{(1,\delta_1(V))},\dots,\xi_{l_k}^{(k,\delta_k(V))},y\right), \tag{16.2}$$

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

$$H_{n,k}^{V}(f)=\int \bar{I}_{n,k}^{V}(f,y)^2\rho(\,dy),\quad f\in\mathcal{F}. \tag{16.3}$$

Put

$$\bar{I}_{n,k}(f,y)=\bar{I}_{n,k}^{\{1,\dots,k\}}(f,y),\quad H_{n,k}(f)=H_{n,k}^{\{1,\dots,k\}}(f), \tag{16.4}$$

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 $\xi_l^{(j)}=\xi_l^{(j,1)}$, $1\le j\le k$, $1\le l\le n$.

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

$$\bar{I}_{n,k}^{(V,\varepsilon)}(f,y)=\frac{1}{k!}\sum_{\substack{(l_1,\dots,l_k)\colon 1\le l_j\le n,\ j=1,\dots,k\\ l_j\ne l_{j'}\text{ if }j\ne j'}} \varepsilon_{l_1}\cdots\varepsilon_{l_k} f\left(\xi_{l_1}^{(1,\delta_1(V))},\dots,\xi_{l_k}^{(k,\delta_k(V))},y\right),\quad f\in\mathcal{F}, \tag{16.5}$$

and

$$H_{n,k}^{(V,\varepsilon)}(f)=\int \bar{I}_{n,k}^{(V,\varepsilon)}(f,y)^2\rho(\,dy),\quad f\in\mathcal{F}, \tag{16.6}$$

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

Let us also introduce the random variables

$$\bar{W}(f) = \int \left[\sum_{V \subset \{1,\dots,k\}} (-1)^{|V|} \bar{I}_{n,k}^{(V,\varepsilon)}(f,y) \right]^2 \rho(\,dy), \quad f \in \mathcal{F}. \tag{16.7}$$

With the help of the above notations Lemma 16.1B can be formulated in the following way.

Lemma 16.1B. *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 15.4 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 15 in the definition of good tail behaviour for a class of integrals of decoupled U-statistics before the formulation of Propositions 15.3 and 15.4), then the inequality

$$P\left(\sup_{f \in \mathcal{F}} |H_{n,k}(f)| > A^2 n^{2k} \sigma^{2(k+1)} \right) < 2P\left(\sup_{f \in \mathcal{F}} \left| \bar{W}(f) \right| > \frac{A^2}{2} n^{2k} \sigma^{2(k+1)} \right) + 2^{2k+1} n^{k-1} e^{-A^{1/2k} n \sigma^2 / k} \tag{16.8}$$

holds with the random variables $H_{n,k}(f)$ introduced in the second identity of relation (16.4) and with $\bar{W}(f)$ defined in formula (16.7) for all $A \ge T$.

A corollary of Lemma 16.1B will be formulated which can be better applied than the original lemma. Lemma 16.B is a little bit inconvenient, because the expression at the right-hand side of formula (16.8) contains a probability depending on $\sup_{f \in \mathcal{F}} |\bar{W}(f)|$, and $\bar{W}(f)$ is a too complicated expression. Some new formulas (16.9) and (16.10) will be introduced which enable us to rewrite $\bar{W}(f)$ in a slightly simpler form. These formulas yield such a corollary of Lemma 16.B which is more appropriate for our purposes. To work out the details first some diagrams will be introduced.

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 the following random variables $H_{n,k}(f|G, V_1, V_2)$ 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 $\bar{W}(f)$:

$$H_{n,k}(f|G, V_1, V_2) = \sum_{\substack{(l_1,\dots,l_k,l'_1,\dots,l'_k):\\ 1\le l_j\le n,\, l_j\ne l_{j'} \text{ if } j\ne j',\, 1\le j,j'\le k,\\ 1\le l'_j\le n,\, l'_j\ne l'_{j'} \text{ if } j\ne j',\, 1\le j,j'\le k,\\ l_j=l'_{j'} \text{ if } (j,j')\in e(G),\, l_j\ne l'_{j'} \text{ if } (j,j')\notin e(G)}} \prod_{j\in\{1,\dots,k\}\setminus v_1(G)} \varepsilon_{l_j} \prod_{j\in\{1,\dots,k\}\setminus v_2(G)} \varepsilon_{l'_j}$$
$$\frac{1}{k!^2} \int f(\xi_{l_1}^{(1,\delta_1(V_1))}, \dots, \xi_{l_k}^{(k,\delta_k(V_1))}, y)$$
$$f(\xi_{l'_1}^{(1,\bar{\delta}_1(V_2))}, \dots, \xi_{l'_k}^{(k,\bar{\delta}_k(V_2))}, y)\rho(\,dy), \tag{16.9}$$

where $\delta_j(V_1) = 1$ if $j \in V_1$, $\delta_j(V_1) = -1$ if $j \notin V_1$, and $\bar{\delta}_j(V_2) = 1$ if $j \in V_2$, $\bar{\delta}_j(V_2) = -1$ if $j \notin V_2$. (Let us observe that if the graph G contains s edges, then the product of the ε-s in (16.9) contains $2(k-s)$ terms, and the number of terms in the sum (16.9) is less than n^{2k-s}.) As the Corollary of Lemma 16.1B will indicate, in the proof of Proposition 15.4 the expression $H_{n,k}(f|G, V_1, V_2)$ has to be estimated. This can be done by means of Theorem 13.3, the multivariate version of Hoeffding's inequality. But the estimate we get in such a way will be rewritten in a form more appropriate for our inductive procedure. This will be done in the next section.

The identity

$$\bar{W}(f) = \sum_{G\in\mathcal{G},\, V_1,V_2\subset\{1,\dots,k\}} (-1)^{|V_1|+|V_2|} H_{n,k}(f|G, V_1, V_2) \tag{16.10}$$

will be proved.

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 (16.5). Let us rewrite this product as a sum of products of the form $\frac{1}{k!^2} \prod_{j=1}^{k} \varepsilon_{l_j} f(\cdots) \prod_{j=1}^{k} \varepsilon_{l'_j} f(\cdots)$ 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 the set of

indices of the remaining random variables ε_{l_j} is $\{l_j\colon\ j \in \{1,\dots,k\} \setminus v_1(G)\}$, the set of indices of the remaining random variables $\varepsilon_{l'_j}$ is $\{l'_{j'}\colon\ j \in \{1,\dots,k\} \setminus v_2(G)\}$, 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 (16.10).

Since the number of terms in the sum of formula (16.10) is less than $2^{4k}k!$, this relation implies that Lemma 16.1B has the following corollary:

Corollary of Lemma 16.1B. *Let a set of functions $\mathcal{F}$ satisfy the conditions of Proposition 15.4. 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*

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}} |H_{n,k}(f)| > A^2 n^{2k}\sigma^{2(k+1)}\right) \\ \le 2 \sum_{G\in\mathcal{G},\, V_1,V_2\subset\{1,\dots,k\}} P\left(\sup_{f\in\mathcal{F}} |H_{n,k}(f|G,V_1,V_2)| > \frac{A^2}{2^{4k+1}k!} n^{2k}\sigma^{2(k+1)}\right) \\ + 2^{2k+1} n^{k-1} e^{-A^{1/2k} n\sigma^2/k} \end{aligned} \tag{16.11}$$

holds with the random variables $H_{n,k}(f)$ and $H_{n,k}(f|G,V_1,V_2)$ defined in formulas (16.4) and (16.9) for all $A \ge T$.

In the proof of Lemmas 16.1A and 16.1B the result of the following Lemmas 16.2A and 16.2B will be applied.

Lemma 16.2A. *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 $\xi_1,\dots,\xi_n$ together with a sequence of independent random variables $(\varepsilon_1,\dots,\varepsilon_n)$, $P(\varepsilon_l = 1) = P(\varepsilon_l = -1) = \frac{1}{2}$, $1 \le l \le n$, which is also independent of the previous sequences.

Let $\mathcal{F}$ be a class of functions which satisfies the conditions of Proposition 15.3. 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

$$\bar{I}_{n,k}^{V}(f) = \frac{1}{k!} \sum_{\substack{(l_1,\dots,l_k)\colon\ 1\le l_j\le n,\ j=1,\dots,k,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} f\left(\xi_{l_1}^{(1,\delta_1(V))},\dots,\xi_{l_k}^{(k,\delta_k(V))}\right) \tag{16.12}$$

and its 'randomized version'

$$\bar I_{n,k}^{(V,\varepsilon)}(f)=\frac{1}{k!}\sum_{\substack{(l_1,\dots,l_k)\colon 1\le l_j\le n,\ j=1,\dots,k,\\ l_j\ne l_{j'}\text{ if }j\ne j'}}\varepsilon_{l_1}\cdots\varepsilon_{l_k}f\left(\xi_{l_1}^{(1,\delta_1(V))},\dots,\xi_{l_k}^{(k,\delta_k(V))}\right),\quad f\in\mathcal{F},\tag{16.12$'$}$$

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

Then the sets of random variables

$$S(f)=\sum_{V\subset\{1,\dots,k\}}(-1)^{|V|}\bar I_{n,k}^{V}(f),\quad f\in\mathcal{F},\tag{16.13}$$

and

$$\bar S(f)=\sum_{V\subset\{1,\dots,k\}}(-1)^{|V|}\bar I_{n,k}^{(V,\varepsilon)}(f),\quad f\in\mathcal{F},\tag{16.13$'$}$$

have the same joint distribution.

Lemma 16.2B. *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* $\xi_1,\dots,\xi_n$ *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 15.4. For all functions* $f\in\mathcal{F}$ *and* $V\in\{1,\dots,k\}$ *consider the decoupled* U*-statistics* $\bar I_{n,k}^{V}(f,y)$ *defined by formula (16.2) 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*

$$W(f)=\int\left[\sum_{V\subset\{1,\dots,k\}}(-1)^{|V|}\bar I_{n,k}^{V}(f,y)\right]^2\rho(\,dy),\quad f\in\mathcal{F}.\tag{16.14}$$

Then the random vectors $\{W(f)\colon\ f\in\mathcal{F}\}$ *defined in (16.14) and* $\{\bar W(f)\colon\ f\in\mathcal{F}\}$ *defined in (16.7) have the same distribution.*

Proof of Lemmas 16.2A and 16.2B. Lemma 16.2A actually agrees with the already proved Lemma 15.1, only the notation is different. The proof of Lemma 16.2B is very similar to the proof of Lemma 15.1. It can be shown 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 (16.5) and (16.7).

Beside this, 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 $(\xi_l^{(j,1)}, \xi_l^{(j,-1)})$, $1 \le l \le n$, $1 \le j \le k$. This implies that also the distribution of the field $W(f)$, $f \in \mathcal{F}$, defined in (16.14) agrees with the distribution of the field we obtain by carrying out the above transformation in the elements of the field $W(f)$, $f \in \mathcal{F}$. But the set of random variables obtained by means of this transformation agrees with the set of random variables introduced in the previous paragraph to describe the conditional distribution of $\bar{W}(f)$, $f \in \mathcal{F}$. (These random variables are defined by the same sums, only the terms in these sums are listed in a different order.) These facts imply Lemma 16.2B.

In the next step Lemma 16.3A will be formulated and proved.

Lemma 16.3A. *Let us consider a class of functions $\mathcal{F}$ satisfying the conditions of Proposition 15.3 with parameter k, and the random variables $\bar{I}_{n,k}^V(f)$, $f \in \mathcal{F}$, $V \subset \{1,\dots,k\}$, defined in formula (16.12). 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, $1 \le j \le k$, in the upper indices. For all $V \subset \{1,\dots,k\}$, $V \ne \{1,\dots,k\}$, there exists a number $A_0 = A_0(k) > 0$ such that the inequality*

$$P\left(\sup_{f\in\mathcal{F}} E\left(\bar{I}_{n,k}^V(f)^2 \middle|\, \mathcal{B}\right) > 2^{-(3k+3)} A^2 n^{2k} \sigma^{2k+2}\right) < n^{k-1} e^{-A^{1/(2k-1)} n\sigma^2/k}. \qquad (16.15)$$

holds for all $A \ge A_0$.

Proof of Lemma 16.3A. Let us first consider the case $V = \emptyset$. In this case the estimate $E\left(\bar{I}_{n,k}^{\emptyset}(f)^2 \middle|\, \mathcal{B}\right) = E\left(\bar{I}_{n,k}^{\emptyset}(f)^2\right) \le \frac{n^k}{k!}\sigma^2 \le n^{2k}\sigma^{2k+2}$ holds for all $f \in \mathcal{F}$. In the above calculation it was exploited that the functions $f \in \mathcal{F}$ are canonical, which implies certain orthogonalities, and beside this the inequality $n\sigma^2 \ge 1$ holds. The above relations imply that for $V = \emptyset$ the probability at the left-hand side of (16.15) equals zero if the number A_0 is chosen sufficiently large, i.e. the inequality (16.15) 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 (16.15) for such sets. For this goal let us introduce the random variables

$$\bar{I}_{n,k}^V(f, l_{u+1}, \dots, l_k) = \frac{1}{k!} \sum_{\substack{(l_1,\dots,l_u):\\ 1\le l_j\le n,\ j=1,\dots,u,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} f\left(\xi_{l_1}^{(1,1)}, \dots, \xi_{l_u}^{(u,1)}, \xi_{l_{u+1}}^{(u+1,-1)}, \dots, \xi_{l_k}^{(k,-1)}\right)$$

for all $f \in \mathcal{F}$, i.e. we fix the last $k-u$ coordinates $\xi_{l_{u+1}}^{(u+1,-1)}, \dots, \xi_{l_k}^{(k,-1)}$ of the random

variable $\bar{I}^V_{n,k}(f)$ and sum up with respect the first u coordinates. Then we can write

$$E\left(\bar{I}^V_{n,k}(f)^2\middle|\,\mathcal{B}\right) = E\left(\left(\sum_{\substack{(l_{u+1},\dots,l_k)\colon 1\le l_j\le n\ j=u+1,\dots,k,\\ l_j\ne l_{j'}\text{ if }j\ne j'}} \bar{I}^V_{n,k}(f,l_{u+1},\dots,l_k)\right)^2\middle|\,\mathcal{B}\right)$$
$$= \sum_{\substack{(l_{u+1},\dots,l_k)\colon 1\le l_j\le n,\ j=u+1,\dots,k,\\ l_j\ne l_{j'}\text{ if }j\ne j'}} E\left(\bar{I}^V_{n,k}(f,l_{u+1},\dots,l_k)^2\middle|\,\mathcal{B}\right). \tag{16.16}$$

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 $(l_{u+1},\dots,l_k)\ne(l'_{u+1},\dots,l'_k)$, which relation holds, since f is a canonical function.

It follows from relation (16.16) that

$$\left\{\omega\colon \sup_{f\in\mathcal{F}} E\left(\bar{I}^V_{n,k}(f)^2\middle|\,\mathcal{B}\right)(\omega) > 2^{-(3k+3)}A^2n^{2k}\sigma^{2k+2}\right\}$$
$$\subset \bigcup_{\substack{(l_{u+1},\dots,l_k)\colon\\ 1\le l_j\le n,\ j=u+1,\dots,k.\\ l_j\ne l_{j'}\text{ if }j\ne j'}} \left\{\omega\colon \sup_{f\in\mathcal{F}} E\left(\bar{I}^V_{n,k}(f,l_{u+1},\dots,l_k)^2\middle|\,\mathcal{B}\right)(\omega) > \frac{A^2n^{2k}\sigma^{2k+2}}{2^{(3k+3)}n^{k-u}}\right\}. \tag{16.17}$$

The probability of the events in the union at the right-hand side of (16.17) can be estimated with the help of the Corollary of Proposition 15.4 with parameter $u<k$ instead of k. (We may assume that Proposition 15.4 holds for $u<k$.) We claim that this corollary yields that

$$P\left(\sup_{f\in\mathcal{F}} E\left(\bar{I}^V_{n,k}(f,l_{u+1},\dots,l_k)^2\middle|\,\mathcal{B}\right) > \frac{A^2n^{k+u}\sigma^{2k+2}}{2^{(3k+3)}}\right) \le e^{-A^{1/(2u+1)}(n-u)\sigma^2}. \tag{16.18}$$

Let us show that if a class of functions $f\in\mathcal{F}$ satisfies the conditions of Proposition 15.3 then it also satisfies relation (16.18). For this goal 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 also introduce the class of those function $\bar{\mathcal{F}}$ in 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 $f(x_1,\dots,x_k)\in\mathcal{F}$. If the class of function $\mathcal{F}$ satisfies the conditions of Proposition 15.3 (with parameter k), then the class of functions $\bar{\mathcal{F}}$ satisfies the conditions of Proposition 15.4 with parameter $u<k$. Hence the Corollary of Proposition 15.4 can be applied for the class of functions $\bar{\mathcal{F}}$ by our inductive hypothesis. We shall apply

it for decoupled U-statistics with this class of kernel functions and parameters u and $n-u$ (instead of k and n), and we define the expressions $\bar{I}_{u,n-u}(f)$ and $H_{n-u,u}(f)$ with the help of the following u independent random sequences of independent μ-distributed random variables of length $n-u$: $\xi_l^{(j)} = \xi_l^{(j,1)}$, $1 \le j \le u$, $l \in \{1, \dots, n\} \setminus \{l_{u+1}, \dots, l_k\}$, where the set of numbers $\{l_{u+1}, \dots, l_k\}$ is the set of indices appearing in formula (16.18). (Actually to get a notation consistent with the definition of these expressions in Section 15 we have to reindex these random variables $\xi_l^{(j)}$ to get random sequences indexed by $l = 1, \dots, n-u$.) With such a choice

$$E\left(\bar{I}^V_{n,k}(f, l_{u+1}, \dots, l_k)^2 | \mathcal{B}\right) = \left(\frac{u!}{k!}\right)^2 \int I_{n,u}(\bar{f}, y)^2 \rho(\, dy) = \left(\frac{u!}{k!}\right)^2 H_{n-u,u}(\bar{f}) \quad (16.19)$$

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 $y = (x_{u+1}, \dots, x_k)$ and the function $H(\cdot)$ defined in (15.6). The Corollary of Proposition 15.4 yields that

$$P\left(\sup_{\bar{f} \in \bar{\mathcal{F}}} (n-u)^{-u} H_{n-u,u}(\bar{f}) \ge A^2 (n-u)^u \sigma^{2u+2}\right) \le e^{-A^{1/(2u+1)}(n-u)\sigma^2} \quad \text{for } A > A_0(u). \quad (16.20)$$

It is not difficult to derive formula (16.18) from relations (16.19) and (16.20). It is enough to check that the level $\frac{A^2 n^{k+u} \sigma^{2k+2}}{2^{(3k+3)}}$ in the probability at the left-hand side of (16.18) can be replaced by $A^2 \left(\frac{k!}{u!}\right)^2 (n-u)^{2u} \sigma^{2u+2}$. This statement holds, since $A^2 \left(\frac{k!}{u!}\right)^2 (n-u)^{2u} \sigma^{2u+2} < A^2 \left(\frac{k!}{u!}\right)^2 n^{2u} \sigma^{2u+2} \le \frac{A^2 n^{k+u} \sigma^{2k+2}}{2^{(3k+3)}}$ if the constant K in the condition $n\sigma^2 > K \log n$ of Proposition 15.3 is chosen sufficiently large.

Relations (16.17) and (16.18) imply that

$$P\left(\sup_{f \in \mathcal{F}} E\left(\bar{I}^V_{n,k}(f)^2 \big| \mathcal{B}\right)(\omega) > 2^{-(3k+3)} A^2 n^{2k} \sigma^{2k+2}\right) \le n^{k-u} e^{-A^{1/(2u+1)}(n-u)\sigma^2}.$$

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 (16.15) 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 15.3, and we can get relation (16.15) with the set V by applying it for the set of function $\mathcal{F}_V$ and set $\bar{V}$.

Now we prove Lemma 16.1A. It will be proved with the help of Lemma 16.2A, the generalized symmetrization lemma 15.2 and Lemma 16.3A.

Proof of Lemma 16.1A. First we show with the help of the generalized symmetrization lemma, i.e. of Lemma 15.2 and Lemma 16.3A that

$$P\left(\sup_{f\in\mathcal{F}} n^{-k/2}\left|\bar{I}_{n,k}(f)\right| > An^{k/2}\sigma^{k+1}\right) < 2P\left(\sup_{f\in\mathcal{F}}|S(f)| > \frac{A}{2}n^k\sigma^{k+1}\right) + 2^k n^{k-1}e^{-A^{1/(2k-1)}n\sigma^2/k} \tag{16.21}$$

with the function $S(f)$ defined in (16.13). To prove relation (16.21) introduce the random variables $Z(f) = I_{n,k}^{\{1,\dots,k\}}(f)$ and $\bar{Z}(f) = -\sum_{V\subset\{1,\dots,k\},\, V\neq\{1,\dots,k\}} (-1)^{|V|}\bar{I}_{n,k}^{V}(f)$ for all $f\in\mathcal{F}$, the σ-algebra $\mathcal{B}$ considered in Lemma 16.3A and the set

$$B = \bigcap_{\substack{V\subset\{1,\dots,k\}\\ V\neq\{1,\dots,k\}}} \left\{\omega\colon \sup_{f\in\mathcal{F}} E\left(\bar{I}_{n,k}^{V}(f)^2\middle|\,\mathcal{B}\right)(\omega) \le 2^{-(3k+3)}A^2n^{2k}\sigma^{2k+2}\right\}.$$

Observe that $S(f) = Z(f) - \bar{Z}(f)$, $f\in\mathcal{F}$, $B\in\mathcal{B}$, and by Lemma 16.3A the inequality $1 - P(B) \le 2^k n^{k-1}e^{-A^{1/(2k-1)}n\sigma^2/k}$ holds. To prove relation (16.21) apply Lemma 15.2 with the above introduced random variables $Z(f)$ and $\bar{Z}(f)$, $f\in\mathcal{F}$, (both here and in the subsequent proof of Lemma 16.1B we work with random variables $Z(\dot{})$ and $\bar{Z}(\dot{})$ indexed by functions $f\in\mathcal{F}$, hence these functions play the role of the parameter p when Lemma 15.2 is applied) random set B and $\alpha = \frac{A}{2}n^k\sigma^{k+1}$, $u = \frac{A}{2}n^k\sigma^{k+1}$. It is enough to show that

$$P\left(|\bar{Z}(f)| > \frac{A}{2}n^k\sigma^{k+1}|\mathcal{B}\right)(\omega) \le \frac{1}{2} \quad \text{for all } f\in\mathcal{F} \quad \text{if } \omega\in\mathcal{B}. \tag{16.22}$$

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 (16.22) and (16.21) hold.

Lemma 16.1A follows from relation (16.21), Lemma 16.2A and the observation that the random variables $\bar{I}_{n,k}^{(V,\varepsilon)}(f)$, $f\in\mathcal{F}$, defined in (16.12′) have the same distribution for all $V\subset\{1,\dots,k\}$ as the random variables $\bar{I}_{n,k}^{\varepsilon}(f)$, defined in formula (14.11). Hence the definition (16.13) of the random variables $S(f)$, $f\in\mathcal{F}$, implies the inequality

$$P\left(\sup_{f\in\mathcal{F}}|S(f)| > \frac{A}{2}n^k\sigma^{k+1}\right) \le 2^k P\left(\sup_{f\in\mathcal{F}}\left|\bar{I}_{n,k}^{\varepsilon}(f)\right| > 2^{-(k+1)}An^k\sigma^{k+1}\right).$$

Lemma 16.1A is proved.

Lemma 16.1B will be proved with the help of the following Lemma 16.3B, which is a version of Lemma 16.3A.

Lemma 16.3B. *Let us consider a class of functions $\mathcal{F}$ satisfying the conditions of Proposition 15.4 and the random variables $\bar{I}_{n,k}^V(f,y)$, $f \in \mathcal{F}$, $V \subset \{1,\dots,k\}$, defined in formula (16.2). 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 those random variables which appear in the definition of the random variables $\bar{I}_{n,k}^V(f,y)$ and $H_{n,k}^V(f)$ introduced in formulas (16.2) and (16.3), and have second argument 1 in their upper index.*

a) Then for all $V \subset \{1,\dots,k\}$, $V \neq \{1,\dots,k\}$, there exists a number $A_0 = A_0(k) > 0$ such that the inequality

$$P\left(\sup_{f\in\mathcal{F}} E(H_{n,k}^V(f)|\mathcal{B}) > 2^{-(4k+4)}A^{(2k-1)/k}n^{2k}\sigma^{2k+2}\right) < n^{k-1}e^{-A^{1/2k}n\sigma^2/k} \tag{16.23}$$

holds for all $A \ge A_0$.

b) Given two subsets $V_1, V_2 \subset \{1,\dots,k\}$ of the set $\{1,\dots,k\}$ define the integrals (of random kernel functions)

$$H_{n,k}^{(V_1,V_2)}(f) = \int |\bar{I}_{n,k}^{V_1}(f,y)\bar{I}_{n,k}^{V_2}(f,y)|\rho(\,dy), \quad f \in \mathcal{F}, \tag{16.24}$$

with the help of the functions $\bar{I}_{n,k}^V(f,y)$ defined in (16.2). 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

$$P\left(\sup_{f\in\mathcal{F}} E(H_{n,k}^{(V_1,V_2)}(f)|\mathcal{B}) > 2^{-(2k+2)}A^2n^{2k}\sigma^{2k+2}\right) < 2n^{k-1}e^{-A^{1/2k}n\sigma^2/k} \tag{16.25}$$

holds for all $A \ge T$.

Proof of Lemma 16.3B. Part a) of Lemma 16.3B can be proved in almost the same way as Lemma 16.3A. Hence I only briefly explain the main step of the proof. In the case $V = \emptyset$ the identity $E(H_{n,k}^V(f)|\mathcal{B}) = E(H_{n,k}^V(f))$ holds, hence it is enough to show that $E(H_{n,k}^V(f)) \le \frac{n^k\sigma^2}{k!} \le \frac{n^{2k}\sigma^{2k+2}}{k!}$ for all $f \in \mathcal{F}$ under the conditions of Proposition 15.4. (These relations hold, because the functions of the class $\mathcal{F}$ are canonical.) The case of a general set V, $V \neq \emptyset$ and $V \neq \{1,\dots,k\}$, 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

$$\bar{I}_{n,k}^V(f,l_{u+1},\dots,l_k,y) = \frac{1}{k!} \sum_{\substack{(l_1,\dots,l_u):\\ 1\le l_j\le n,\ j=1,\dots,u,\\ l_j\neq l_{j'} \text{ if } j\neq j'}} f\left(\xi_{l_1}^{(1,1)},\dots,\xi_{l_u}^{(u,1)},\xi_{l_{u+1}}^{(u+1,-1)},\dots,\xi_{l_k}^{(k,-1)},y\right)$$

for all $f \in \mathcal{F}$. It can be shown that because of the canonical property of the functions $f \in \mathcal{F}$

$$E\left(\bar{H}_{n,k}^{V}(f)^2 \middle| \mathcal{B}\right) = \sum_{\substack{(l_{u+1},\dots,l_k):\\ 1\le l_j\le n,\ j=u+1,\dots,k,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} \int E\left(\bar{I}_{n,k}^{V}(f, l_{u+1},\dots,l_k,y)^2 \middle| \mathcal{B}\right) \rho(\,dy),$$

and the proof of part a) of Lemma 16.3B can be reduced to the inequality

$$P\left(\sup_{f\in\mathcal{F}} E\left(\int \bar{I}_{n,k}^{V}(f, l_{u+1},\dots,l_k,y)^2 \rho(\,dy)\middle| \mathcal{B}\right) > \frac{A^{(2k-1)/k} n^{k+u}\sigma^{2k+2}}{2^{(4k+4)}}\right)$$
$$\le e^{-A^{(2k-1)/2(2u+1)k}(n-u)\sigma^2}.$$

This inequality can be proved, similarly to relation (16.18) in the proof of Lemma 16.3A with the help of the Corollary of Proposition 15.4. 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 16.3B will be proved with the help of Part a) and the inequality

$$\sup_{f\in\mathcal{F}} E(H_{n,k}^{(V_1,V_2)}(f)|\mathcal{B}) \le \left(\sup_{f\in\mathcal{F}} E(H_{n,k}^{V_1}(f)|\mathcal{B})\right)^{1/2} \left(\sup_{f\in\mathcal{F}} E(H_{n,k}^{V_2}(f)|\mathcal{B})\right)^{1/2}$$

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

$$P\left(\sup_{f\in\mathcal{F}} E(H_{n,k}^{(V_1,V_2)}(f)|\mathcal{B}) > 2^{-(2k+2)} A^2 n^{2k}\sigma^{2k+2}\right)$$
$$\le P\left(\sup_{f\in\mathcal{F}} E(H_{n,k}^{V_1}(f)|\mathcal{B}) > 2^{-(4k+4)} A^{(2k-1)/k} n^{2k}\sigma^{2k+2}\right)$$
$$+ P\left(\sup_{f\in\mathcal{F}} E(H_{n,k}^{V_2}(f)|\mathcal{B}) > A^{(2k+1)/k} n^{2k}\sigma^{2k+2}\right)$$

Hence the estimate (16.23) together with the inequality

$$P\left(\sup_{f\in\mathcal{F}} E(H_{n,k}^{V_2}(f)|\mathcal{B}) > A^{(2k+1)/k} n^{2k}\sigma^{2k+2}\right) \le n^{k-1} e^{-A^{1/2k} n\sigma^2} \tag{16.26}$$

imply relation (16.25). Relation 16.26 follows from Part a) of Lemma 16.3B if $V_2 \ne \{1,\dots,k\}$ and $A \ge 1$ since 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 (16.26). In the case $V_2 = \{1, \dots, k\}$ it follows from the conditions of Part b) of Lemma 16.3B. Indeed, since $A^{(2k+1)/2k} \ge T^{(2k+1)/2k}$, by the conditions of Proposition 15.4 the estimate (15.7) holds if the number A is replaced in it by $A^{(2k+1)/2k}$ (at both side of the inequality), and this relation implies inequality (16.26) in this case.

Now we turn to the proof of Lemma 16.1B.

Proof of Lemma 16.1B. By Lemma 16.2B it is enough to prove that relation (16.8) holds if the random variables $\bar{W}(f)$ are replaced in it by the random variables $W(f)$ defined in formula (16.14). We shall prove this by applying the generalized form of the symmetrization lemma, Lemma 15.2, with the choice of $Z(f) = H_{n,k}^{(\bar{V},\bar{V})}(f)$, $\bar{V} = \{1, \dots, k\}$, $\bar{Z}(f) = Z(f) - W(f)$, $f \in \mathcal{F}$, $\mathcal{B} = \mathcal{B}(\xi_1^{(j,1)}, \dots, \xi_n^{(j,1)};\ 1 \le j \le k)$, $\alpha = \frac{A^2}{2}n^{2k}\sigma^{2k+2}$, $u = \frac{A^2}{2}n^{2k}\sigma^{2k+2}$ and the set

$$B = \bigcap_{\substack{(V_1,V_2)\colon V_j \in \{1,\dots,k\},\ j=1,2,\\ V_1 \neq \{1,\dots,k\} \text{ or } V_2 \neq \{1,\dots,k\}}} \left\{ \omega\colon \sup_{f \in \mathcal{F}} E(H_{n,k}^{(V_1,V_2)}(f)|\mathcal{B})(\omega) \le 2^{-(2k+2)}A^2 n^{2k}\sigma^{2k+2} \right\}.$$

By Lemma 16.3B the inequality $1 - P(B) \le 2^{2k+1}n^{k-1}e^{-A^{1/2k}n\sigma^2/k}$ holds. Observe that $Z(f) = H_{n,k}^{(\bar{V},\bar{V})}(f) = H_{n,k}(f)$ for all $f \in \mathcal{F}$. Hence to prove Lemma 16.1B with the help of Lemma 15.2 it is enough to show that

$$P\left(|\bar{Z}(f)| > \frac{A^2}{2}n^{2k}\sigma^{2(k+1)} \middle| \mathcal{B} \right)(\omega) \le \frac{1}{2} \quad \text{for all } f \in \mathcal{F} \text{ if } \omega \in B. \tag{16.27}$$

To prove this relation observe that because of the definition of the set B

$$E(|\bar{Z}(f)||\mathcal{B})(\omega) \le \sum_{\substack{(V_1,V_2)\colon V_j \in \{1,\dots,k\},\ j=1,2,\\ V_1 \neq \{1,\dots,k\} \text{ or } V_2 \neq \{1,\dots,k\}}} E(H_{n,k}^{(V_1,V_2)}(f)|\mathcal{B})(\omega) \le \frac{A^2}{4}n^{2k}\sigma^{2k+2}$$

if $\omega \in B$ for all $f \in \mathcal{F}$. Hence the 'conditional Markov inequality' implies inequality (16.27). Lemma 16.1B is proved.

17. The proof of the main result.

This section contains the proof of Proposition 15.3 together with Proposition 15.4. They complete the proof of the main result of this work, of Theorem 8.4.

A.) The proof of Proposition 15.3.

The proof of Proposition 15.3 is similar to that of Proposition 6.2. It applies an induction procedure with respect to the parameter k. In the proof of Proposition 15.3 for parameter k we may assume that Propositions 15.3 and 15.4 hold for $u < k$. In the proof we want to give a good estimate on the expression

$$P\left(\sup_{f\in\mathcal{F}}\left|\bar{I}^{\varepsilon}_{n,k}(f)\right| > 2^{-(k+1)}An^k\sigma^{k+1}\right)$$

appearing in the estimate (16.1) of Lemma 16.1A. To estimate this probability we introduce (using the notation of Proposition 15.3) the functions

$$S^2_{n,k}(f)(x_l^{(j)},\, 1\le l\le n,\, 1\le j\le k) = \frac{1}{k!}\sum_{\substack{(l_1,\dots,l_k):\\ 1\le l_j\le n,\ j=1,\dots,k,\\ l_j\ne l_{j'}\ \text{if}\ j\ne j'}} f^2\left(x_{l_1}^{(1)},\dots,x_{l_k}^{(k)}\right),\quad f\in\mathcal{F}, \tag{17.1}$$

with $x_l^{(j)}\in X$, $1\le l\le n$, $1\le j\le k$. 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 $H\subset X^{kn}$

$$H = H(A) = \Big\{(x_l^{(j)},\, 1\le l\le n,\, 1\le j\le k):\ \sup_{f\in\mathcal{F}} S^2_{n,k}(f)(x_l^{(j)},\, 1\le l\le n,\, 1\le j\le k) > 2^kA^{4/3}n^k\sigma^2\Big\}. \tag{17.2}$$

We want to show that

$$P(\{\omega\colon\ (\xi_l^{(j)}(\omega),\, 1\le j\le n,\, 1\le j\le k)\in H\}) \le 2^ke^{-A^{2/3k}n\sigma^2}\quad \text{if } A\ge T. \tag{17.3}$$

Relation (17.3) 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)

$$f^2(x_1,\dots,x_k) = \sum_{V\subset\{1,\dots,k\}} f_V(x_j, j\in V) \tag{17.4}$$

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 (17.4) 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 $f \in \mathcal{F}$ with the expression $\bar{I}_{n,k}(\cdot)$ defined in (14.10). By applying the Hoeffding decomposition (17.4) for each term $f^2(\xi_{l_1}^{(1)} \dots, \xi_{l_k}^{(k)})$ in the expression $S^2_{n,k}(f)$ we get that

$$\begin{aligned} P\left(\sup_{f \in \mathcal{F}} S^2_{n,k}(f)(\xi_l^{(j)},\, 1 \le l \le n,\, 1 \le j \le k) > 2^k A^{4/3} n^k \sigma^2 \right) \\ \le \sum_{V \subset \{1,\dots,k\}} P\left(\frac{|V|!}{k!} \sup_{f \in \mathcal{F}} n^{k-|V|} |\bar{I}_{n,|V|}(f_V)| > A^{4/3} n^k \sigma^2 \right) \end{aligned} \tag{17.5}$$

with the functions f_V in (17.4). We want to give a good estimate for all terms in the sum at the right-hand side in (17.5). For this goal first we show that the classes of functions $\{f_V\colon\ f \in \mathcal{F}\}$ in the expansion (17.4) satisfy the conditions of Proposition 15.4 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 15.3 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 $\{2^k f_1^2, \dots, 2^k f_m^2\}$ is an ε-dense subset of the class of functions $\mathcal{F}' = \{2^k f^2\colon\ f \in \mathcal{F}\}$, 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 (15.7) equals zero if $A_0 \ge 1$ under the conditions of Proposition 15.3. I claim that the terms corresponding to sets V, $1 \le |V| \le k$, in these sums satisfy the inequality

$$P\left(\frac{|V|!}{k!} n^{k-|V|} \sup_{f \in \mathcal{F}} |\bar{I}_{n,|V|}(f_V)| > A^{4/3} n^k \sigma^2 \right)$$

$$\le P\left(\sup_{f\in\mathcal{F}}|\bar{I}_{n,|V|}(f_V)|>A^{4/3}\frac{k!}{|V|!}n^{|V|}\sigma^{|V|+1}\right)\le e^{-A^{2/3k}n\sigma^2}\quad\text{if } 1\le|V|\le k. \tag{17.6}$$

The first inequality in (17.6) holds, since $\sigma^{|V|+1}\le\sigma^2$ for $|V|\ge 1$. 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 15.3 is sufficiently large. 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 15.3 have a good tail behaviour at level $T^{4/3}$. Relations (17.5) and (17.6) together with the estimate in the case $V=\emptyset$ imply formula (17.3).

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 13.3) that

$$\begin{aligned}
&P\left(\left|\bar{I}^{\varepsilon}_{n,k}(f)\right|>2^{-(k+2)}An^k\sigma^{k+1}\,\middle|\,\xi_l^{(j)}(\omega)=x_l^{(j)},1\le l\le n,1\le j\le k\right)\\
&\quad\le C\exp\left\{-\frac{1}{2}\left(\frac{A^2n^{2k}\sigma^{2(k+1)}}{2^{2k+4}S^2_{n,k}(x_l^{(j)},1\le l\le n,\,1\le j\le k)/k!}\right)^{1/k}\right\} \qquad (17.7)\\
&\quad\le Ce^{-2^{-4-4/k}A^{2/3k}(k!)^{1/k}n\sigma^2}\quad\text{for all } f\in\mathcal{F}\quad\text{if } (x_l^{(j)},\,1\le l\le n,\,1\le j\le k)\notin H
\end{aligned}$$

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

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 $\rho_j(x_l^{(j)})=\frac{1}{n}$, $1\le l\le n$. 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. Introduce the numbers

$$\bar{\delta}=An^{-k/2}2^{-(k+2)}\sigma^{k+1}\quad\text{and}\quad\delta=2^{-(k+2)}n^{-k-1/2}\le\bar{\delta}.$$

(The inequalities $\delta\le 1$ and $\delta\le\bar{\delta}$ hold, since $A\ge A_0\ge 1$, and $\sigma\ge n^{-1/2}$.) Choose a δ-dense set $\mathcal{F}_\delta=\{f_1,\dots,f_m\}$ in the space $L_2(X^k,\mathcal{X}^k,\rho)$ with $m\le D\delta^{-L}\le 2^{(k+2)L}n^{\beta+(k+1)L/2}$ elements with $\beta=\max\left(\frac{logD}{\log n},0\right)$. Then for all $f\in\mathcal{F}$ there exists some $f_s\in\mathcal{F}_\delta$ such that $\int(f-f_s)^2\,d\rho\le\delta^2$. Hence

$$\begin{aligned}
&P\left(\left|\bar{I}^{\varepsilon}_{n,k}(f-f_s)\right|>A2^{-(k+2)}n^k\sigma^{k+1}\,\middle|\,\xi_l^{(j)}=x_l^{(j)},\,1\le l\le n,\,1\le j\le k\right)\\
&\quad=P\left(\left|\bar{I}^{\varepsilon}_{n,k}(f-f_s)\right|>\bar{\delta}n^{3k/2}\,\middle|\,\xi_l^{(j)}=x_l^{(j)},\,1\le l\le n,\,1\le j\le k\right)\\
&\quad\le P\left(\left|\bar{I}^{\varepsilon}_{n,k}(f-f_s)\right|>\delta n^{3k/2}\,\middle|\,\xi_l^{(j)}=x_l^{(j)},\,1\le l\le n,\,1\le j\le k\right)=0
\end{aligned}$$

The above relations and formula (17.7) imply that

$$P\left(\sup_{f\in\mathcal{F}}\left|\bar{I}^{\varepsilon}_{n,k}(f)\right|>2^{-(k+1)}An^k\sigma^{k+1}\middle|\,\xi_l^{(j)}(\omega)=x_l^{(j)},1\le l\le n,1\le j\le k\right)$$
$$\le\sum_{f_s\in\mathcal{F}_\delta}P\left(\left|\bar{I}^{\varepsilon}_{n,k}(f_s)\right|>2^{-(k+2)}An^k\sigma^{k+1}\middle|\,\xi_l^{(j)}(\omega)=x_l^{(j)},1\le l\le n,1\le j\le k\right)$$
$$\le C2^{(k+2)L}n^{\beta+(k+1)L/2}e^{-2^{-4-4/k}A^{2/3k}(k!)^{1/k}n\sigma^2} \tag{17.8}$$
$$\text{if } \{x_l^{(j)},\ 1\le l\le n,\ 1\le j\le k\}\notin H.$$

Relations (17.3) and (17.8) imply that

$$P\left(\sup_{f\in\mathcal{F}}\left|\bar{I}^{\varepsilon}_{n,k}(f)\right|>2^{-(k+1)}An^k\sigma^{k+1}\right) \tag{17.9}$$
$$\le C2^{(k+2)L}n^{\beta+(k+1)L/2}e^{-2^{-4-4/k}A^{2/3k}n\sigma^2}+2^ke^{-A^{2/3k}k!n\sigma^2}\quad\text{if } A\ge T.$$

Proposition 15.3 follows from the estimates (16.1) and (17.9) 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 (15.5).

Now we turn to the proof of Proposition 15.4.

B.) The proof of Proposition 15.4.

Because of formula (16.11) in the Corollary of Lemma 16.1B to prove Proposition 15.4 i.e. inequality (15.7) it is enough to choose a sufficiently large parameter A_0 and to show that under the conditions of Proposition 15.4 sufficiently large and to show that

$$P\left(\sup_{f\in\mathcal{F}}|H_{n,k}(f|G,V_1,V_2)|>\frac{A^2}{2^{4k+1}k!}n^{2k}\sigma^{2(k+1)}\right)\le 2^{k+1}e^{-A^{1/2k}n\sigma^2} \tag{17.10}$$
$$\text{for all } G\in\mathcal{G}\quad\text{and}\quad V_1,V_2\in\{1,\dots,k\}\quad\text{if } A>T\ge A_0$$

with the random variables $H_{n,k}(f|G,V_1,V_2)$ defined in formula (16.9).

Let us first prove formula (17.10) 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 'symmetrizing term' ε_j. In this case we apply a special argument to prove relation (17.10).

If G is such a diagram for which $|e(G)| = k$, then the Schwarz inequality yields that

$$|H_{n,k}(f|G, V_1, V_2)| \le \frac{1}{k!} \left(\sum_{\substack{(l_1,\dots,l_k):\\ 1\le l_j\le n,\ 1\le j\le k,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} \int f^2(\xi_{l_1}^{(1,\delta_1(V_1))}, \dots, \xi_{l_k}^{(k,\delta_k(V_1))}, y)\rho(\,dy) \right)^{1/2}$$
$$\frac{1}{k!} \left(\sum_{\substack{(l_1,\dots,l_k):\\ 1\le l_j\le n,\ 1\le j\le k,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} \int f^2(\xi_{l_1}^{(1,\bar\delta_1(V_2))}, \dots, \xi_{l_k}^{(k,\bar\delta_k(V_2))}, y)\rho(\,dy) \right)^{1/2} \tag{17.11}$$

with $\delta_j(V_1) = 1$ if $j \in V_1$, $\delta_j(V-1) = -1$ if $j \notin V_1$, and $\bar\delta_j(V_2) = 1$ if $j \in V_2$, $\bar\delta_j(V_2) = -1$ if $j \notin V_2$.

Relation (17.11) can be proved for instance by bounding first each integral in formula (16.9) by means of the Schwarz inequality, and then by bounding the sum appearing in such a way by means of the inequality $\sum |a_j b_j| \le \left(\sum a_j^2\right)^{1/2} \left(\sum b_j^2\right)^{1/2}$. (Observe during this calculation that the sets of indices $\{l_1, \dots, l_k\}$ and $\{l'_1, \dots, l'_k\}$ agree if $|e(G)| = k$, only the elements of these sets are listed in different order.)

By formula (17.11)

$$\left\{\omega\colon \sup_{f\in\mathcal{F}} |H_{n,k}(f|G, V_1, V_2)(\omega)| > \frac{A^2}{2^{4k+1}k!} n^{2k}\sigma^{2(k+1)}\right\}$$
$$\subset \left\{\omega\colon \sup_{f\in\mathcal{F}} \sum_{\substack{(l_1,\dots,l_k):\\ 1\le l_j\le n,\ 1\le j\le k,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} \int f^2(\xi_{l_1}^{(1,\delta_1(V_1))}(\omega), \dots, \xi_{l_k}^{(k,\delta_k(V_1))}(\omega), y)\rho(\,dy) > \frac{A^2 n^{2k}\sigma^{2(k+1)}k!}{2^{4k+1}}\right\}$$
$$\cup \left\{\omega\colon \sup_{f\in\mathcal{F}} \sum_{\substack{(l_1,\dots,l_k):\\ 1\le l_j\le n,\ 1\le j\le k,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} \int f^2(\xi_{l_1}^{(1,\bar\delta_1(V_2))}(\omega), \dots, \xi_{l_k}^{(k,\bar\delta_k(V_2))}(\omega), y)\rho(\,dy) > \frac{A^2 n^{2k}\sigma^{2(k+1)}k!}{2^{4k+1}}\right\},$$

hence

$$P\left(\sup_{f\in\mathcal{F}} |H_{n,k}(f|G, V_1, V_2)| > \frac{A^2}{2^{4k+1}k!} n^{2k}\sigma^{2(k+1)}\right) \tag{17.12}$$

$$\le 2P\left(\sup_{f\in\mathcal{F}}\frac{1}{k!}\sum_{\substack{(l_1,\dots,l_k)\colon 1\le l_j\le n,\ 1\le j\le k,\\ l_j\ne l_{j'}\text{ if } j\ne j'}} h_f(\xi_{l_1}^{(1,1)},\dots,\xi_{l_k}^{(k,1)})>\frac{A^2n^{2k}\sigma^{2(k+1)}}{2^{4k+1}}\right)$$

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 $\delta_j(V-1)$ and $\bar{\delta}_j(V_2)$, 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 them.)

I claim that

$$P\left(\sup_{f\in\mathcal{F}}|\bar{I}_{n,k}(h_f)|\ge 2^kAn^k\sigma^2\right)\le 2^ke^{-A^{1/2k}n\sigma^2}\quad\text{for } A\ge A_0 \tag{17.13}$$

if the constants A_0 and K are chosen sufficiently large in Proposition 15.4. Relation (17.13) together with the relation $A^2\frac{n^{2k}\sigma^{2(k+1)}}{2^{4k+1}}\ge 2^kAn^k\sigma^2$ (if the parameter K is sufficiently large) imply that the probability at the right-hand side of (17.12) can be bounded by $2^{k+1}e^{-A^{1/2k}n\sigma^2}$, and the estimate (17.10) holds in the case $|e(G)|=k$.

Relation (17.13) is similar to relation (17.3) (together with the definition of the random set H in formula (17.2)), and a modification of the proof of the latter estimate yields the proof also in this case. Indeed, it follows from the conditions of Proposition 15.4 that $0\le\int h_f(x_1,\dots,x_k)\mu(\,dx_1)\dots\mu(\,dx_k)\le\sigma^2$ for all $f\in\mathcal{F}$, and it is not difficult to check that $\sup|h_f(x_1,\dots,x_k)|\le 2^{-2(k+1)}$, and the class of functions $\mathcal{H}=\{2^kh_f,\ f\in\mathcal{F}\}$ is an L_2-dense class with exponent L and parameter D. Hence by applying the Hoeffding decomposition of the functions h_f, $f\in\mathcal{F}$, similarly to formula (17.4) we get for all $V\subset\{1,\dots,k\}$ such sets of functions $(h_f)_V$, $f\in\mathcal{F}$, which satisfy the conditions of Proposition 15.3. Hence a natural adaptation of the estimate given for the expression at the right-hand side of (17.5) yields the proof of formula (17.13). 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 by our inductive hypothesis the result of Proposition 15.3 holds also for k.

In the case $e(G)<k$ formula (17.10) will be proved with the help of the multivariate version of Hoeffding's inequality, Theorem 13.3. In the proof of this case an expression, analogous to $S_{n,k}^2(f)$ defined in formula (17.1) 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 $J_0(G)$ contains those sequences $(l_1,\dots,l_k,l_1',\dots,l_k')$ which appear as indices in the summation in formula (16.9) for a fixed diagram G. We also introduce an appropriate partition of it.

For this aim let us first define the sets $M_1(G) = \{j(1), \dots, j(k - |e(G)|)\} = \{1, \dots, k\} \setminus v_1(G)$, $j(1) < \dots < j(k - |e(G)|)$, and $M_2(G) = \{\bar\jmath(1), \dots, \bar\jmath(k - |e(G|)\} = \{1, \dots, k\} \setminus v_2(G)$, $\bar\jmath(1) < \dots < \bar\jmath(k - |e(G|)$, the sets of those vertices of the first and second row of the diagram G in increasing order from which no edge starts. 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 those vectors which can appear 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 $M_1(G) \cup M_2(G)$. The elements of $V(G)$ are such vectors whose coordinates are indexed by the set $M_1(G) \cup M_2(G)$, and they take different integer values between 1 and n. Given a vector $v \in V(G)$ put $v = (v^{(1)}, v^{(2)})$ with $v^{(1)} = \{v(r),\ 1 \le r \le k - |e(G)|\}$, and $v^{(2)} = \{\bar v(r),\ 1 \le r \le k - |e(G)|\}$, where $v^{(1)}$ and $v^{(2)}$ denote the set of coordinates of v indexed by the elements of the set $M_1(G)$ and $M_2(G)$ 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 $\{j(1), \dots, j(k - |e(G)|)\} = M_1(G)$, $\{\bar\jmath(1), \dots, \bar\jmath(k - |e(G)|)\} = M_2(G)$, $v = (v^{(1)}, v^{(2)})$ 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.

Given a vector $v \in V(G)$, $v = (v^{(1)}, v^{(2)})$, the set $E_G(v)$ consists of the following vectors $\ell = (l_1, \dots, l_k, l'_1, \dots, l'_k) \in J_0(G)$: For $j \in M_1(G)$ the coordinate l_j agrees with the corresponding element of $v^{(1)}$, for $\bar\jmath \in M_2(G)$ 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 $(j_s, \bar\jmath_{s'})$, $1 \le s, s' \le |e(G)|$ in such a way that $(j_s, \bar\jmath_{s'}) \in e(G)$. The identity $l_{j_s} = l'_{\bar\jmath_{s'}}$ holds for the elements of these pairs, and the coordinates lj_s and $l]_{\bar\jmath_{s'}}$ are all different if their coordinates are not elements of one of these pairs. 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 the random variables $H_{n,k}(f|G, V_1, V_2)$ defined in (16.9) can be written with their help in the following way:

$$\begin{aligned} H_{n,k}(f|G, V_1, V_2) = & \sum_{v=(v^{(1)},v^{(2)})\in V(G)} \prod_{s=1}^{k-|e(G)|} \varepsilon_{l_{j(s)}} \prod_{s=1}^{k-|e(G)|} \varepsilon_{l'_{\bar\jmath(s)}} \\ & \sum_{(l_1,\dots,l_k,l'_1\dots,l'_k)\in E_G(v)} \frac{1}{k!^2} \int f(\xi_{l_1}^{(1,\delta_1(V_1))}, \dots, \xi_{l_k}^{(k,\delta_k(V_1))}, y) \\ & \qquad f(\xi_{l'_1}^{(1,\bar\delta_1(V_2))}, \dots, \xi_{l'_k}^{(k,\bar\delta_k(V_2))}, y)\rho(\,dy), \end{aligned} \tag{17.14}$$

where $\delta_j(V_1)=1$ if $j\in V_1$, $\delta_j(V_1)=-1$ if $j\notin V_1$, and $\bar\delta_j(V_2)=1$ if $j\in V_2$, $\bar\delta_j(V_2)=-1$ if $j\notin V_2$.

The inequality

$$P\left(S^2(\mathcal{F}|G,V_1,V_2)>2^{2k}A^{8/3}n^{2k}\sigma^4\right)\le 2^{k+1}e^{-A^{2/3k}n\sigma^2}\quad \text{if } A\ge A_0 \text{ and } e(G)<k \tag{17.15}$$

will be proved for the random variable

$$S^2(\mathcal{F}|G,V_1,V_2)=\sup_{f\in\mathcal{F}}\frac{1}{k!^2}\sum_{v\in V(G)}\Bigg(\sum_{(l_1,\dots,l_k,l'_1,\dots,l'_k)\in E_G(v)}\int f(\xi_{l_1}^{(1,\delta_1(V_1))},\dots,\xi_{l_k}^{(k,\delta_k(V_1))},y)$$
$$f(\xi_{l'_1}^{(1,\bar\delta_1(V_2))},\dots,\xi_{l'_k}^{(k,\bar\delta_k(V_k))},y)\rho(\,dy)\Bigg)^2, \tag{17.16}$$

where $\delta_j(V_1)=1$ if $j\in V_1$, $\delta_j(V_1)=-1$ if $j\notin V_1$, and $\bar\delta_j(V_2)=1$ if $j\in V_2$, $\bar\delta_j(V_2)=-1$ if $j\notin V_2$. The random variable $S^2(\mathcal{F}|G,V_1,V_2)$ defined in (17.16) plays a similar role in the proof of Proposition 15.4 as the random variable $\sup\limits_{f\in\mathcal{F}}S^2_{n,k}(f)$ with $S^2_{n,k}(f)$ defined in formula (17.1) played in the proof of Proposition 15.3.

To prove formula (17.15) let us first fix some $v\in V(G)$ and let us observe that, similarly to the proof of relation (17.11), the Schwarz inequality implies the relation

$$\Bigg(\sum_{(l_1,\dots,l_k,l'_1,\dots,l'_k)\in E_G(v)}\int f(\xi_{l_1}^{(1,\delta_1(V_1))},\dots,\xi_{l_k}^{(k,\delta_k(V_1))},y)$$
$$f(\xi_{l'_1}^{(1,\bar\delta_1(V_2))},\dots,\xi_{l'_k}^{(k,\bar\delta_k(V_2))},y)\rho(\,dy)\Bigg)^2$$
$$\le\left(\sum_{(l_1,\dots,l_k,l'_1,\dots,l'_k)\in E_G(v)}\int f^2(\xi_{l_1}^{(1,\delta_1(V_1))},\dots,\xi_{l_k}^{(k,\delta_k(V_1))},y)\rho(\,dy)\right)$$
$$\left(\sum_{(\bar l_1,\dots,\bar l_k,\bar l'_1,\dots,\bar l'_k)\in E_G(v)}\int f^2(\xi_{\bar l'_1}^{(1,\bar\delta_1(V_2))},\dots,\xi_{\bar l'_k}^{(k,\bar\delta_k(V_2))},y)\rho(\,dy)\right)$$

for all $v\in V(G)$. Summing up these inequalities for all $v\in V(G)$ we get that

$$S^2(\mathcal{F}|G,V_1,V_2)$$
$$\le\sup_{f\in\mathcal{F}}\sum_{v\in V(G)}\frac{1}{k!}\left(\sum_{(l_1,\dots,l_k,l'_1,\dots,l'_k)\in E_G(v)}\int f^2(\xi_{l_1}^{(1,\delta_1(V_1))},\dots,\xi_{l_k}^{(k,\delta_k(V_1))},y)\rho(\,dy)\right)$$
$$\frac{1}{k!}\left(\sum_{(\bar l_1,\dots,\bar l_k,\bar l'_1,\dots,\bar l'_k)\in E_G(v)}\int f^2(\xi_{\bar l'_1}^{(1,\bar\delta_1(V_2))},\dots,\xi_{\bar l'_k}^{(k,\bar\delta_k(V_2))},y)\rho(\,dy)\right) \tag{17.17}$$

$$\le \sup_{f\in\mathcal{F}} \frac{1}{k!} \left(\sum_{\substack{(l_1,\dots,l_k)\colon 1\le l_j\le n,\, 1\le j\le k,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} \int f^2(\xi_{l_1}^{(1,\delta_1(V_1))},\dots,\xi_{l_k}^{(k,\delta_k(V_1))},y)\rho(\,dy) \right)$$

$$\sup_{f\in\mathcal{F}} \frac{1}{k!} \left(\sum_{\substack{(\bar l_1,\dots,\bar l_k)\colon 1\le \bar l_j\le n,\, 1\le j\le k,\\ \bar l_j\ne \bar l_{j'} \text{ if } j\ne j'}} \int f^2(\xi_{\bar l_1}^{(1,\bar\delta_1(V_2))},\dots,\xi_{\bar l_k}^{(k,\bar\delta_k(V_2))},y)\rho(\,dy) \right).$$

To check the second inequality of formula (17.17) let us first observe that it can be reduced to the simpler relation, where the expression $\sup_{f\in\mathcal{F}}$ is omitted at each place. The simplified inequality obtained 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 (17.17). At both sides of the inequality a sum consisting of terms of the form

$$\frac{1}{k!^2} \int f^2(\xi_{l_1}^{(1,\delta_1(V_1))},\dots,\xi_{l_k}^{(k,\delta_k(V_1))},y)\rho(\,dy) \int f^2(\xi_{\bar l_1}^{(1,\bar\delta_1(V_1))},\dots,\xi_{\bar l_k}^{(k,\bar\delta_k(V_2))},y)\rho(\,dy), \tag{17.18}$$

appears. It is enough to check that if a term of this form appears in the middle term of the simplified version formula of (17.17), 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 (17.18) which appears in the sum we get by carrying out the multiplications in middle term of (17.17) determines uniquely the index $v=(v^{(1)},v^{(2)})\in V(G)$ in the outer sum in the original form of this expression for which the product of the inner sums yields this term. Indeed, that vector $v=(v^{(1)},v^{(2)})\in V(G)$ (with coordinates with indices in $M_1(G)\cup M_2(G)$) must be taken for which $v^{(1)}$ agrees with the restriction of the vector $l=(l_1,\dots,l_k)$ to coordinates with indices in $M_1(G)$ and $v^{(2)}$ agrees with the restriction of the vector $(\bar l_1,\dots,\bar l_k)$ to coordinates with indices in $M_2(G)$. Beside this, if the multiplication is carried out at the right-hand side of (17.17) then the sum contains all such terms of the form (17.18) which appeared in the previous sum.

Relation (17.17) implies that

$$P(S^2(\mathcal{F}|G,V_1,V_2)) > 2^{2k}A^{8/3}n^{2k}\sigma^4) \le 2P\left(\sup_{f\in\mathcal{F}} \bar I_{n,k}(h_f) > 2^kA^{4/3}n^k\sigma^2\right)$$

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 (17.15) it is enough to show that

$$2P\left(\sup_{f\in\mathcal{F}} \bar I_{n,k}(h_f) > 2^kA^{4/3}n^k\sigma^2\right) \le 2^{k+1}e^{-A^{2/3k}n\sigma^2} \quad \text{if } A\ge A_0. \tag{17.19}$$

Actually formula (17.19) follows from the already proven formula (17.13), only the parameter A has to be replaced by $A^{4/3}$ in it.

With the help of relation (17.15) the proof of Proposition 15.4 can be completed similarly to Proposition 15.3. It follows from the multivariate version of Hoeffding's inequality, Theorem 13.3 and the representation of the random variable $H_{n,k}(f|G,V_1,V_2)$ in the form (17.14) that

$$P\left(|H_{n,k}(f|G,V_1,V_2)| > \frac{A^2}{2^{4k+2}k!}n^{2k}\sigma^{2(k+1)}\middle|\, \xi_l^{j,\pm 1},\ 1\le l\le n,\ 1\le j\le k\right)(\omega)$$
$$\le Ce^{-2^{-(6+2/k)}A^{2/3k}n\sigma^2} \quad \text{if} \quad S^2(\mathcal{F}|G,V_1,V_2)(\omega)\le 2^{2k}A^{8/3}n^{2k}\sigma^4 \text{ and } A\ge A_0 \tag{17.20}$$

with an appropriate constant $C=C(k)>0$ for all $f\in\mathcal{F}$ and $G\in\mathcal{G}$ such that $|e(G)|<k$ and $V_1,V_2\subset\{1,\dots,k\}$. (Observe that the conditional probability estimated in (17.20) can be represented in the following way. In a point $\omega\in\Omega$ fix the values of $\xi_l^{(j,\pm 1)}(\omega)$ for all indices $1\le l\le n$ and $1\le j\le k$ in the random variable $H_{n,k}(f|G,V_1,V_k)$, and the conditional probability in this point ω equals the probability that the random variable, (depending on the random variables ε_l, $1\le l\le n$), obtained in such a way is greater than $\frac{A^2}{2^{4k+2}k!}n^{2k}\sigma^{2(k+1)}$.)

Indeed, in this case the conditional probability considered in (17.20) can be bounded because of the multivariate version of the Hoeffding inequality (Theorem 13.3) by $C\exp\left\{-\frac{1}{2}\left(\frac{A^4n^{4k}\sigma^{4(k+1)}}{2^{8k+4}(k!)^2S^2(\mathcal{F}|G,V_1,V_2)/(k!)^2}\right)^{1/2j}\right\}\le C\exp\left\{-\frac{1}{2}\left(\frac{A^{4/3}n^{2k}\sigma^{4k}}{2^{10k+4}}\right)^{1/2j}\right\}$ with an appropriate $C=C(k)>0$, where $2j=2k-2|e(G)|$, and $0\le|e(G)|\le k-1$. Since $j\le k$, $n\sigma^2\ge 1$, and also $\frac{A^{4/3}}{2^{10k+4}}\ge 1$ if A_0 is chosen sufficiently large the above calculation implies relation (17.20).

Let us show that also $\sup\limits_{f\in\mathcal{F}}H_{n,k}(f|G,V_1,V_2)$ can be estimated in the following form:

$$P\left(\sup_{f\in\mathcal{F}}|H_{n,k}(f|G,V_1,V_2)| > \frac{A^2}{2^{4k+1}k!}n^{2k}\sigma^{2(k+1)}\middle|\, \xi_l^{j,\pm 1},\ 1\le l\le n,\ 1\le j\le k\right)(\omega)$$
$$\le Cn^{(3k+2)L/2+\beta}e^{-2^{-(6+2/k)}A^{2/3k}n\sigma^2}$$
$$\text{if } S^2(\mathcal{F}|G,V_1,V_2))(\omega)\le 2^{2k}A^{8/3}n^{2k}\sigma^4 \text{ and } A\ge A_0 \tag{17.21}$$

for all $G\in\mathcal{G}$ such that $|e(G)|<k$ and $V_1,V_2\subset\{1,\dots,k\}$.

To prove formula (17.21) let us consider two sets $V_1,V_2\subset\{1,\dots,k\}$ and a diagram G such that $|e(G)|<k$, and take some points $x_l^{(j,\pm 1)}$, $1\le l\le n$, $1\le j\le k$, in the space $(X,\mathcal{X})$ such that if $\xi_l^{(j,\pm 1)}(\omega)=x_l^{(j,\pm 1)}$ for all $1\le l\le n$ and $1\le j\le k$, then the inequality $S^2(\mathcal{F}|G,V_1,V_2)(\omega)\le 2^{2k}A^{8/3}n^{2k}\sigma^4$ holds.

Introduce with the help of these points 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 $x_l^{(j,\delta_j(V_1))}$, $1\le l\le n$, and $\nu_j^{(2)}$ which are uniformly distributed on the points

$x_l^{(j,\bar\delta_j(V_2))}$, $1 \le l \le n$, i.e. let $\nu_j^{(1)}\left(\{x_l^{(j,\delta_j(V_1))}\}\right) = \frac{1}{n}$ and $\nu_j^{(2)}\left(\{x_l^{(j,\bar\delta_j(V_2))}\}\right) = \frac{1}{n}$, $1 \le l \le n$, $1 \le j \le k$, where $\delta_j(V_1) = 1$ if $j \in V_1$, $\delta_j(V_1) = -1$ if $j \notin V_1$, and similarly $\bar\delta_j(V_2) = 1$ if $j \in V_2$ and $\bar\delta_j(V_2) = -1$ if $j \notin V_2$. Let us consider the product measures $\alpha_1 = \nu_1^{(1)} \times \cdots \times \nu_k^{(1)} \times \rho$ and $\alpha_2 = \nu_1^{(2)} \times \cdots \times \nu_k^{(2)} \times \rho$ 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 15.4, and define the measure $\alpha = \frac{\alpha_1+\alpha_2}{2}$.

If $f \in \mathcal{F}$ and $g \in \mathcal{F}$ are two functions such that $\int (f-g)^2\, d\alpha \le \delta^2$ with some $\delta > 0$, then we give an upper bound for $|H_{n,k}(f|G, V_1, V_2)(\omega) - H_{n,k}(g|G, V_1, V_2)(\omega)|$. (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 $\int (f-g)^2\, d\alpha_j \le 2\delta^2$, and

$$\int |f(x_{l_1}^{(1,\delta_1(V_1))}, \dots, x_{l_k}^{(k,\delta_k(V_1))}, y) - g(x_{l_1}^{(1,\delta_1(V_1))}, \dots, x_{l_k}^{(k,\delta_k(V_1))}, y)|^2 \rho(\,dy) \le 2\delta^2 n^k,$$

$$\int |f(x_{l_1}^{(1,\delta_1(V_1))}, \dots, x_{l_k}^{(k,\delta_k(V_1))}, y) - g(x_{l_1}^{(1,\delta_1(V_1))}, \dots, x_{l_k}^{(k,\delta_k(V_1))}, y)| \rho(\,dy) \le \sqrt{2}\delta n^{k/2}$$

for all $1 \le l \le k$, and $1 \le l_j \le n$, and the same result holds if all $\delta_j(V_1)$ are replaced by $\delta_j(V_2)$, $1 \le j \le k$. Since $|f| \le 1$, $|g| \le 1$ if $f, g \in \mathcal{F}$, the condition $\int (f-g)^2\, d\alpha \le \delta^2$ implies that

$$\begin{aligned}\int |f(\xi_{l_1}^{(1,\delta_1(V_1))}(\omega), \dots, \xi_{l_k}^{(k,\delta_k(V_1))}(\omega), y) f(\xi_{l'_1}^{(1,\bar\delta_1(V_2))}(\omega), \dots, \xi_{l'_k}^{(k,\bar\delta_k(V_2))}(\omega), y) \rho(\,dy) \\ - g(\xi_{l_1}^{(1,\delta_1)}(\omega), \dots, \xi_{l_k}^{(k,\delta_k)}(\omega), y) g(\xi_{l'_1}^{(1,\bar\delta_1(V_2))}(\omega), \dots, \xi_{l'_k}^{(k,\bar\delta_k(V_2))}(\omega), y) \rho(\,dy)| \\ \le 2\sqrt{2}\delta n^{k/2}\end{aligned}$$

for all vectors $(l_1, \dots, l_k, l'_1, \dots, l'_k)$ in that summation which appears in the definition of $H_{n,k}(\cdot|G, V_1, V_2)$ in formula (16.9). Hence

$$|H_{n,k}(f|G, V_1, V_2)(\omega) - H_{n,k}(g|G, V_1, V_2)(\omega)| \le 2\sqrt{2}\delta n^{5k/2}$$

if $f, g \in \mathcal{F}$, $\int (f-g)^2\, d\alpha < \delta^2$ and such an $\omega \in \Omega$ is considered for which $\xi_l^{(j,\pm1)}(\omega) = x_l^{(j,\pm1)}$, $1 \le l \le n$, $1 \le j \le k$.

Put $\bar\delta = \frac{A^2 n^{-k/2} \sigma^{2(k+1)}}{2^{(4k+7/2)} k!}$, and $\delta = n^{-(3k+2)/2} \le \bar\delta$ (the inequality $\delta \le \bar\delta$ holds, since $\sigma \ge \frac{1}{\sqrt{n}}$ and we may assume that $A \ge A_0$ is sufficiently large), choose a δ-dense subset $\mathcal{F}_\delta = \{f_1, \dots, f_m\} \subset \mathcal{F}$ in the $L_2(X^k \times Y, \mathcal{X}^k \times \mathcal{Y}, \alpha)$ space with $m \le D\delta^{-L} \le n^{(3k+2)L/2+\beta}$ elements. Then an argument similar to that at the end of the proof of Proposition 15.3 shows that if such an $\omega \in \Omega$ is taken for which $\xi_l^{(j,\pm1)}(\omega) = x_l^{(j,\pm1)}$ for all $1 \le l \le n$ and $1 \le j \le k$, then for all $f \in \mathcal{F}$ such a function $f_j \in \mathcal{F}_\delta$ can be chosen for which $\int (f - f_j)^2\, d\alpha \le \delta^2$, and this inequality implies that $|H_{n,k}(f|G, V_1, V_2) - H_{n,k}(f_j|G, V_1, V_2)| \le 2\sqrt{2}\bar\delta n^{k/2} = \frac{A^2}{2^{4k+2} k!} n^{2k} \sigma^{2(k+1)}$. This inequality together with relation (17.20) and the bound on the cardinality of the set $\mathcal{F}_\delta$ yield inequality (17.21).

It follows from relations (17.15) and (17.21) that

$$P\left(\sup_{f\in\mathcal{F}}|H_{n,k}(f|G,V_1,V_2)|>\frac{A^2}{2^{4k+1}k!}n^{2k}\sigma^{2(k+1)}\right)\le 2^{k+1}e^{-A^{2/3k}n\sigma^2}$$
$$+Cn^{(3k+2)L/2+\beta}e^{-2^{-(6+2/k)}A^{2/3k}n\sigma^2}\quad\text{if } A\ge A_0$$

for all $V_1, V_2 \subset \{1,\dots,k\}$ also in the case $|e(G)| \le k-1$. This inequality implies that relation (17.10) holds also in this case if the constants A_0 and K are chosen sufficiently large in Proposition 15.4. Proposition 15.4 is proved.

18. 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 some references. I also write about the background of these problems which may explain the motivation for their study.

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. In the estimation of the empirical distribution function with the help of censored data we cannot apply the classical maximum likelihood method, since in this case we have to choose our estimate from a too large class of distribution functions. The main problem is that there is no dominating measure with respect to which all candidates which can appear as our estimate have a density function. 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 observed 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 the above principle in the following way: Let us estimate the distribution function $F(x)$ of the censored data simultaneously together 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, (see [25]), 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 have to show that the linearization procedure gives a small error. The estimation of this error led to the problem about a good estimate on the tail distribution of the integral of an appropriate function of two variables with respect to the product of a normalized empirical measure with itself. Moreover, as a more detailed investigation showed, we actually need the solution of a more general problem where we have to bound the tail distribution of the supremum of a class of such integrals. The main subject of this work is to solve the above problems in a more general setting, to estimate not only two-fold, but also k-fold random integrals and the supremum of such integrals for an approriate class of kernel functions with respect to a normalized empirical distribution for all $k \geq 1$.

The proof of the the limit theorem for the Kaplan–Meyer estimate explained in this work applied the explicit form of this estimate. It would be interesting to find such a modification of this proof which only 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.

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 [37]. Paper [38] 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 [29]. These results are completely satisfactory for the study in [37], 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. Beside this, 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 [29] did not allow a real generalization of these 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* [7]. 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. The main difference between them is that integration with respect to $\mu_n - \mu$ instead of the empirical distribution μ_n means some sort of normalization, while this normalization is missing in the definition of U-statistics. I return to this question later.

The main part of this work starts at Section 3. A general overview of the results without the hard technical details can be found in [32].

First the estimation of sums of independent random variables or one-fold random integrals with respect to a normalized empirical distribution and the supremum of such expressions is investigated in Sections 3 and 4. This question has a fairly big literature. I would mention first of all the books *A course on empirical processes* [11], *Real Analysis and Probability* [12] and *Uniform Central Limit Theorems* [13] of 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.

Section 3 deals with the tail behaviour of sums of independent and bounded random variables with expectation zero. The proof of two already classical results, Bernstein's and Bennett's inequalities is given there. (Their proofs can be found e.g. in Theorem 1.3.2 of [13] and [5]). We are also interested in the question when they give an estimate suggested by the central limit theorem. Actually, as it is explained in Section 3, Bennett's inequality gives a bound suggested by a Poissonian approximation of partial sums of independent random variables. Bernstein's inequality provides an estimate suggested by the central limit theorem if the variance of the sum is not too small. (The results in Section 3 explain this statement more explicitly.) If the variance of the sum is too small, then Bennett's inequality provides a slight improvement. Moreover, as Example 3.2 shows, Bennett's inequality is essentially sharp in this case.

The estimate on the tail distribution of a sum of independent random variables is weak if this sum has a small variance. This means that in this case the probability that the sum is larger than a given value may be much larger than the (rather small) value suggested by the central limit theorem. Such a behaviour may occur, because the contribution of some unpleasant irregularities to this probability may be non-negligible in the case of a small variance.

In the study of the supremum of sums of independent random variables a good control is needed on the tail distribution of the (supremum of) sums of independent random variables even if they have small variance. (A natural multivariate analog of this problem appears in the general case.) The solution of this problem turned out to be the hardest part of this work. The results based on the similar behaviour of partial sums and their Gaussian counterpart is not sufficient in this case, some new ideas have to be applied. In the proof of sharp estimates in this case we also use some kind of symmetrization arguments. The last result of Section 3, Hoeffding's inequality presented in Theorem 3.4 is an important ingredient of these symmetrization arguments. It is also a classical result whose proof can be found for instance in [23].

Section 4 contains 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 defined with the help of a sequence of independent and identically distributed random variables. 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. To handle such cases 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 supremum of countably approximable classes is taken. 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 is 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 measurable also with respect to the smaller σ-algebra generated by the random variable S_t, $t \in T'$. Beside this, if the finite dimensional distributions of the random variables S_t, $t \in T$, are given, then by the results of classical measure theory the probability of the events measurable with respect to the σ-algebra generated by these random variables S_t, $t \in T$, is also determined. But we cannot get the probability of all events we are interested in such a way. In particular, if T is a non-countable set, then the events $\left\{\omega\colon \sup\limits_{t\in T} S_t(\omega) > u\right\}$ are non-measurable with respect to the above σ-algebra, and 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 [13].

I restricted my attention to such cases 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 $\left\{\omega\colon \sup\limits_{t\in T} S_t(\omega) > u\right\}$ is contained in a set measurable with respect to the σ-algebra generated by the random variables S_t, $t \in T$, which has probability zero. In such a case

the probability $P\left(\sup_{t\in T} S_t > u\right)$ can be defined as $P(A_u)$. This approach led me to the definition of countable approximable classes of random variables. If this property holds, then we can 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 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 on a set of probability zero, although the finite dimensional distributions of the random variables we consider remain the same after such an operation. Hence, if we are interested in the probability of the supremum of a non-countable set of random variables with described finite dimensional distributions 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 processes 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 this property is proved 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 it demands 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′. It 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 standard definition of the ε-entropy. The term of L_2-dense classes seemed to be the appropriate notion to introduce in this work. To apply the results related to L_2-dense classes we also need

some knowledge about how to check this property 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 [13], [43], [52] and the references in them) deal with this subject. An important result in this field is Sauer's lemma, (Lemma 5.1) which together with some other results, like Lemma 5.3 imply that several interesting classes of sets or functions are Vapnik–Červonenkis classes.

I put the proof of these results to the Appendix, partly because they can be found in the literature, partly because in this work Vapnik–Červonenkis classes play a different and less important role than at other places. Here 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, and as a consequence, 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 would deserve 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 contains 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. This means the application of an inductive procedure, in which an appropriate sequence of finite subsets of the original set of random variables is introduced, and a good estimate is given on the supremum of the random variables in these subsets by means of an inductive procedure. The subsets became denser subsets of the original set of the random variables at each step of this procedure. This chaining argument is a popular method in certain investigation. It is hard to say with whom to attach it. Its introduction may be connected to some works of R. M. Dudley. It is worth mentioning that Talagrand [51] worked out a sharpened version of it which yields in the investigation of certain problems a sharper estimate. But it seems to me that in the study of the problems of this work it does not provide a real improvement.

Theorem 4.2 can be proved in such a way, but this method is not strong enough to supply a proof of Theorem 4.1. The cause of this weakness is that there is no 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 much weaker estimate than the result we want to prove under the conditions of Theorem 4.1. Lemma 6.1 contains the result the chaining argument yields under these conditions. In Section 6 still another result, Lemma 6.2 is formulated. It can be considered as a special case of Theorem 4.1 where only the supremum of partial sums with small variances is estimated. 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. It has to be checked 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 [2], and although its presentation is different from 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, (see e.g. [49] or [50]), and his research was also continued by some other authors. I would mention the works of M. Ledoux [27] and P. Massart [40]. Concentration inequalities give a bound about the difference between the supremum of a set of appropriately defined random variables and the expected value of this supremum; they express how strongly this supremum is concentrated 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 [50].

Theorem 18.1. (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 $Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(\xi_i)$ and $v = E(\sup_{f \in \mathcal{F}} \sum_{i=1}^n f^2(\xi_i))$. 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 18.1 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 18.1 contain the quantity $EZ = E\left(\sup_{f \in \mathcal{F}} \sum_{i=1}^n f(\xi_i)\right)$. 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. Talagrand's work [51] deals very much with such problems, and it contains many interesting results in this direction. But it seems to solve problems of different sort, and the results of [51] do not help in the proof of Theorem 4.1. There is a paper [16] which provides a useful estimate about the expected value of the supremum of random sums under the conditions of Theorem 4.1. But I preferred a direct proof of this result. 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 magnitude of the expected value EZ of the above supremum.

The L_2-dense property of the class $\mathcal{F}$ implies that the expected value of the supremum of the normalized random sums considered in Theorem 4.1 is bounded by const. $\sigma \log^{1/2} \frac{2}{\sigma}$, and Theorem 4.1 provides a good estimate on the supremum only above this level.

The main results of this work are presented in Section 8. A weaker version of Theorem 8.3 about an estimate of the distribution of a degenerate U-statistic was first proved in a paper of Arcones and Giné in [3]. The result of Theorem 8.3 in the present form is proved in my paper [35]. Its version about multiple integrals with respect to a normalized empirical measure formulated in Theorem 8.1 is proved in [31]. This paper contains a direct proof. On the other hand, Theorem 8.1 can be derived from Theorem 8.3 by means of Theorem 9.4 of this paper. Theorem 8.5 is the natural Gaussian counterpart of Theorem 8.3. The limit theorem about degenerate U-statistics, Theorem 10.4 (and its version about limit theorems for multiple integrals with respect to normalized empirical measures, Theorem 10.4′ in Appendix C) was discussed in this work to explain better the relation between degenerate U-statistics (or multiple integrals with respect to normalized empirical measures) and multiple Wiener–Itô integrals. A proof of this result based on similar ideas as that discussed here can be found in [14]. Theorem 6.6 of my lecture note [28] contains such a weakened version of Theorem 8.5 which does not take into account the variance of the random integral.

Example 8.7 is a natural supplement of Theorem 8.5 which shows that the estimate of Theorem 8.5 is sharp if only the variance of a Wiener–Itô integral is known. In this Lecture Note I mentioned the results of papers [1] and [26] without proof. I discussed mainly the content of [26] and explained its relation to the other work of this paper. The proof of these papers apply a method different of those of this work. It would be interesting to prove them with the methods discussed here. These papers contain such a refinement of Theorems 8.5 and 8.3 respectively whose estimates depend on some other rather complicated quantities. In some cases they supply a better estimate. On the other hand in the problems discussed here they have a restricted importance because their conditions are hard to check.

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 [33]. Earlier Arcones and Giné proved a weaker form of this result in paper [4], but their work did not help in the proof of the results of this note. They were based on an adaptation of Alexander's method to the multivariate case. Theorem 8.6 contains the natural Gaussian counterpart of Theorems 8.2 and 8.4.

Example 8.8 in Section 8 shows that the condition $u \le$ const. $n\sigma^3$ imposed in Theorem 8.3 in the case $k = 2$ cannot be dropped. The paper of Arcones and Giné [3] contains another example explained by Talagrand to the authors which also has a similar consequence. But that example does not provide such an explicit comparison of the upper and lower bound on the probability investigated in Theorem 8.3 as Example 8.8.

Similar examples could be constructed for all $k \geq 1$.

Example 8.8 shows that at high levels only a very weak (and from practical point of view not really important) improvement of the estimation on the tail distribution of degenerate U-statistics is possible. But probably there exists a multivariate version of Bennett's inequality Theorem 3.3 which provides such an estimate. Moreover, there is some hope to get a similar strengthened form of Theorems 8.2 and 8.4 (or of Theorem 4.2 in the one-dimensional case). This question is not investigated in the present work.

Section 9 deals with the properties of U-statistics. Its first result, Theorem 9.1, is a rather classical result. It is the so-called Hoeffding decomposition of U-statistics to the sum of degenerate statistics. Its proof first appeared in the paper [22], but it can be found at many places. The explanation of this work contains some ideas similar to [48]. I tried to explain that Hoeffding's decomposition is the natural multivariate version of the (trivial) decompositon of sums of independent random variables to sums of independent random variables *with expectation zero* plus the sum of the expectations of the original random variables. Moreover, Hoeffding's decomposition shows some similarity to this simple decomposition.

Theorem 9.2 and Proposition 9.3 can be considered as a continuation of the investigation of the Hoeffding's decomposition in Theorem 9.1. They tell how the properties of the kernel function of the original U-statistic are inherited in the properties of the kernel functions of the degenerate U-statistics taking part in its Hoeffding decomposition. In several applications of Hoeffding's decomposition such results are also needed.

The last result of Section 9, Theorem 9.4, enables us to reduce the estimation of multiple random integrals with respect to normalized empirical measures to the estimation of degenerate U-statistics. This result is a version of Hoeffding's decomposition, where multiple integrals with respect to a normalized empirical distribution are decomposed to the sum of degenerate U-statistics. Multiple random integrals with respect to a normalized empirical measure can be simply written as sums of U-statistics, and by applying the Hoeffding decomposition for each term of these sums we get the desired decomposition. Theorem 9.4 yields the result we get in such a way. This formula is very similar to the original Hoeffding decomposition. The main difference between them is that the coefficients of the degenerate U-statistics in the decomposition of Theorem 9.4 are relatively small. The cancellation effect caused by integration with respect to a *normalized* empirical measure is reflected in the appearance of small coefficients in the decomposition given in Theorem 9.4. Theorem 9.4 was proved in [33]. The same proof is given in this note, but some calculations are worked out in more detail.

Theorem 8.1 can be derived from Theorem 8.3 and Theorem 8.2 from Theorem 8.4 by means of Theorem 9.4. The proof of the latter results is simpler. The results of Sections 10–12 contain the results needed in the proof of Theorem 8.3 and its Gaussian counterpart Theorems 8.5 and 8.7. The proof of these results is based on good estimates of high moments of degenerate U-statistics and multiple Wiener–Itô integrals. The classical proof of the one-variate counterparts of these results is based on a good estimate of the moment generating function. This method was replaced by the estimate of the moments, because the moment generating function of a k-fold Wiener–Itô integral is divergent for $k \geq 3$, and this property is also reflected in the behaviour of degenerate U-

statistics. On the other hand, good estimates on high moments can replace the estimate of the moment generating function. A good estimate can be given for all moments of a Wiener–Itô integral, while we have a good estimate only on not too high moments of degenerate U-statistics. This is related to the fact that there is a good estimate on the tail distribution of degenerate U-statistic only for not too large values.

I know of two deep methods to study high moments of multiple Wiener–Itô integrals. Both of them can be adapted to the study of the moments of degenerate U-statistics. They deserve a more detailed discussion.

The first one of them is called Nelson's inequality named after Edward Nelson who published it in his paper [42]. This inequality simply implies Theorem 8.5 about multiple Wiener–Itô integrals, although with worse constants. Later Leonhard Gross discovered a deep and very useful generalization of this result which he published in the work *Logarithmic Sobolev inequalities* [19]. In that paper Gross compared two Markov processes with the same infinitesimal operator but with possibly different initial distribution, where the second Markov process had stationary distribution. He could give a sharp bound on the Radon–Nikodym derivative of the distribution of the first Markov process at a time T with respect to the (stationary) distribution of the second Markov process at time T on the basis of the properties of the infinitesimal operator of the Markov processes. This result made possible to generalize Nelson's inequality to more general cases. In particular, such a result may help to prove (a weaker version of) Theorem 8.3 (with worse universal constants). In a preliminary version of this Lecture Note [36] that can be found on my homepage I worked out a detailed proof of Theorem 8.3 on the basis of Gross' paper. Here I do not go into the details. Let me also remark that Gross' method works not only in the study of these problems, but in several hard problems of the probability theory. (See e.g [20] or [27]). Nevertheless, in the present note I followed a different method, because this seemed to be better applicable here.

I applied a method related to the names of Kyoshi Itô and Roland L'vovich Dobrushin. Here the theory of multiple Wiener–Itô integrals with respect to a white noise is applied. The notion of this integral was introduced in paper [24]. It is useful, because every random variable measurable with respect to the σ-algebra generated by the Gaussian random variables of the underlying white noise with finite second moment can be written as the sum of Wiener–Itô integrals of different order. Moreover, if only Wiener–Itô integrals of symmetric kernel functions are taken, then this representation is unique. An important result, the so-called diagram formula, formulated in Theorem 10.2, expresses products of Wiener–Itô integrals as a sum of such integrals. This result which shows some similarity to the Feynman diagrams applied in the statistical physics was proved in [9]. Actually this paper discussed a modified version of Wiener–Itô integrals which is more appropriate to study the action of shift operators for non-linear functionals of a stationary Gaussian field. But these modified Wiener–Itô integrals can be investigated in almost the same way as the original ones. The diagram formula has a simple consequence formulated in the form of a corollary of this note which enables us to calculate the expectation of products of Wiener–Itô integrals, in particular the moments of a Wiener–Itô integral. This result was useful in the proof of Theorem 8.5, in

the estimation of the tail-distribution of Wiener–Itô integrals. Itô's formula for multiple Wiener–Itô integrals (Theorem 10.3) was proved in [24].

The diagram formula has a natural and useful analog both for degenerate U-statistics and multiple integrals with respect to a normalized empirical measure. They enables us to express the product of degenerate U-statistics and multiple integrals as the sum of such expressions. These results make possible to adapt several useful methods in the study of non-linear functionals of a Gaussian random field to the study of non-linear functionals of normalized empirical measures, and this may be useful in many cases. The diagram formula was proved for degenerate U-statistics in [35] and for multiple random integrals with respect to a normalized empirical measures in [31]. Let me remark that the diagram formula for degenerate U-statistics was formulated in [35] in a different from than in the present note. In that paper I wanted to formulate the diagram formula with the help of such diagrams which appear in the diagram formula for Wiener–Itô integrals. I could do this only in a somewhat artificial way. The formulation of this result with the help of diagrams containing chains as it is done here seems to be more natural. Let me also remark that the study of results similar to the diagram for did not get such an attention in the literature that it would deserve in my opinion. I know only of one work where such questions were investigated. It is the paper of Surgailis [45], where a version of the diagram formula is proved for Poissonian integrals. The Corollary of Theorem 11.2 is of special interest for us, because it enables us to prove such moment estimates which are useful in the proof of Theorem 8.3.

It is worth mentioning that the problems about Wiener–Itô integrals are closely related to the study of Hermite polynomials or to their multivariate version, to the so-called Wick polynomials. (See e.g. [28] or [39] for the notion of Wick polynomials.) Appendix C contains the most important properties of Hermite polynomials needed in the study of Wiener–Itô integrals. In particular, it contains the proof of Proposition C2 which states that the set of all Hermite polynomials is a complete orthogonal system in the Hilbert space of the functions square integrable with respect to the standard normal measure. This result can be found for instance in Theorem 5.2.7 of [47]. In the present proof I wanted to show that this result is closely related to the so-called moment problem, i.e. to the question when a distribution is determined by its moments uniquely. This method, with some refinement, can be applied to prove some generalizations of Proposition C2 about the completeness of orthogonal polynomials with respect to more general weight functions.

Itô's formula makes a relation between Wiener–Itô integrals and Hermite polynomials. The results about multiple Wiener–Itô integrals have their analogs for Wick polynomials. Thus for instance there is a diagram formula for the product of Wick polynomials which also has some interesting generalizations. Such questions are studied both in probability theory and statistical physics, see [39] and [44]. The relation between Wiener–Itô integrals and Hermite polynomials also has a natural counterpart in the study of other multiple random integrals. The so-called Appell polynomials, (see [46]), appeared in such a way.

Theorems 8.3, 8.5 and 8.7 were proved on the basis of the results in Sections 10–12 in Section 13. This section also contains the proof of the multivariate version of Hoeffding's

inequality, formulated in Theorem 13.3. This result is needed in the symmetrization argument applied in the proof of Theorem 8.4. A weaker version of it (an estimate with a worse constant in the exponent) which would be satisfactory for our purposes would simply follow from a classical result, called Borell's inequality. But since this result is not discussed in this note, and I was interested in a proof which yields the best estimate in the exponent of this estimate I have chosen another proof, given in [34] which is based on the results of Sections 10–12. Later I have learned that this estimate is contained in an implicit form also in the paper [6] of A. Bonami.

Sections 14–17 are devoted to the proof of Theorems 8.4 and 8.6. They are based on a similar argument as their one-variate counterparts, Theorems 4.1 and 4.2. The proof of Theorem 8.6 about the supremum of Wiener–Itô integrals is based, similarly to the proof of Theorem 4.2 on the chaining argument. This is a rather general method which I cannot connect to a definite name. In the proof of Theorem 8.4, the chaining argument only yields a weaker result formulated in Proposition 14.1 which helps to reduce Theorem 8.4 to the proof of Proposition 14.2. In the one-variate case a similar approach was applied, where the proof of Theorem 4.1 was reduced to that of Proposition 6.2 by means of Proposition 6.1. The next step in the proof of Theorem 8.6 has no one-variate counterpart. The notion of so-called decoupled U-statistics was introduced, and Proposition 14.2 was reduced to a similar result about degenerate U-statistics formulated in Proposition 14.2′.

The adjective 'decoupled' in the expression decoupled U-statistic refers to the fact that it is such version of a U-statistic where independent copies of a sequence of independent and identically distributed random variables are put into different coordinates of the kernel function of the U-statistic. Their study is a popular subject of some authors. In particular, the main subject of the book [7] is a comparison of the properties of U-statistics and decoupled U-statistics. A result of de la Peña and Montgomery–Smith [8] formulated in Theorem 14.3 helps to reduce some problems about U-statistics to a similar problem about decoupled U-statistics. In this lecture note the proof of Theorem 14.3 is given in Appendix D. It follows the argument of the original proof, but several steps are worked out in detail where the authors gave only a very short explanation. Paper [8] also contains some kind of converse result of Theorem 14.3, but as it is not needed in the present work, I omitted its discussion.

Decoupled U-statistics behave similarly to the original U-statistics. Beside this, the decoupled property makes possible the application of a symmetrization arguments, and this may be useful in some investigations. For example Proposition 14.2′ can be proved in such a way. Its counterpart about usual U-statistics, Proposition 14.2, cannot be proved by means of a simple adaptation of this method, but it can be simply derived from Proposition 14.2′ with the help of Theorem 14.3.

The above example shows why the application of the symmetrization method and the introduction of degenerate U-statistics is useful in the study of certain problems. But the application of the symmetrization method has its price. Generally, if a (usual or decoupled) U-statistic is estimated with its help, then this U-statistic is replaced during the estimation by a larger U-statistic, (by a U-statistic with larger variance), and as a consequence we cannot get sharp estimates. In particular, the decoupling technique

and the symmetrization argument made with its help can provide only the proof of a weakened version of Theorem 8.3 with worse universal constants. This is the reason why a different proof of this result was given in this work. (The earlier version of this Lecture Note [36] contains the proof of a weakened version of Theorem 8.3 by means of a symmetrization argument.)

The proof of Theorem 8.4 was reduced to that of Proposition 14.2′ in Section 14. Sections 15–17 deal with the proof of this result. It was proved in my paper [33]. The proof is similar to that of its one-variate version Proposition 6.2, but some additional difficulties have to be overcome. 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 Proposition 6.2 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 14.2′ a more sophisticated construction has to be applied.

In this case Lemma 7.1 is not sufficient for us in its original form. We need a generalization of this result, and this is done in Lemma 15.2. The proof of Lemma 15.2 is not hard. The real difficulty arises when we want to apply it in our case. In an application of this lemma formula (15.3) has to be checked, and this means the estimation of some non-trivial conditional probabilities. The hardest part of the proof of Proposition 14.2′ is related to checking the validity of formula (15.3) when we want to apply Lemma 15.2. In the analogous relation, in formula (7.1) of Lemma 7.1 it was enough to bound a usual probability, and this was simple.

Proposition 14.2′ was proved by means of an inductive procedure formulated in Proposition 15.3, which is the multivariate analog of Proposition 6.2. A basic ingredient of both proofs was a symmetrization argument. But while this symmetrization argument could be simply applied in the one-variate case, it meant a most serious problem in the proof of Theorem 15.3. To overcome this difficulty another statement was introduced in Proposition 15.4. Propositions 15.3 and 15.4 were proved simultaneously by means of an appropriate inductive procedure. Their proof was based on a refinement of the arguments in the proof of Proposition 6.2. We also had 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). This result has several different proofs. Here I write down a relatively simple proof of P. Frankl and J. Pach which appeared in [15]. It is based on some linear algebraic arguments.

The following equivalent reformulation of Sauer's lemma will be proved. Let us take a set $S = S(n)$ consisting of n elements and a class $\mathcal{E}$ of subsets of S consisting of m elements $E_1, \dots, E_m \subset S$. Assume that $m \ge m_0 + 1$ with $m_0 = m_0(n,k) = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{k-1}$. Then there exists a set $F \subset S$ of cardinality k which the class of sets $\mathcal{E}$ shatters. Actually, it is enough to show that there exists a set F of cardinality greater than or equal to k which the class of sets $\mathcal{E}$ shatters, because if a set has this property, then all of its subsets have it. This latter statement will be proved.

To prove this statement let us first list the subsets $X_0, \dots, X_{m_0}$ of the set S of cardinality less than or equal to $k-1$, and correspond to all sets $E_i \in \mathcal{E}$ the vector $e_i = (e_{i,1}, \dots, e_{i,m_0})$, $1 \le i \le m$, with elements

$$e_{i,j} = \begin{cases} 1 & \text{if } X_j \subseteq E_i \\ 0 & \text{if } X_j \not\subseteq E_i \end{cases} \qquad 1 \le i \le m, \text{ and } 1 \le j \le m_0.$$

Since $m > m_0$, the vectors $e_1, \dots, e_m$ are linearly dependent. Because of the definition of the vectors e_i, $1 \le i \le m$, this can be expressed in the following way: There is a non-zero vector $(f(E_1), \dots, f(E_m))$ such that

$$\sum_{E_i:\ E_i \supseteq X_j} f(E_i) = 0 \quad \text{for all } 1 \le j \le m_0. \tag{A1}$$

Let F be a *minimal* set with the property

$$\sum_{E_i:\ E_i \supseteq F} f(E_i) = \alpha \ne 0. \tag{A2}$$

Such a set F really exists, since every maximal element of the family $\{E_i\colon\ 1 \le i \le m,\ f(E_i) \ne 0\}$ satisfies relation (A2). The requirement that F should be a minimal set means that if F is replaced by some $H \subset F$, $H \ne F$, at the left-hand side of (A2), then this expression equals zero. The inequality $|F| \ge k$ holds because of relation (A1) and the definition of the sets X_j.

Introduce the quantities

$$Z_F(H) = \sum_{E_i:\ E_i \cap F = H} f(E_i)$$

for all $H \subseteq F$.

Then $Z_F(F) = \alpha$, and for any set of the form $H = F \setminus \{x\}$, $x \in F$,

$$Z_F(H) = \sum_{E_i\colon E_i \cap F = H} f(E_i) = \sum_{E_i\colon E_i \supseteq H} f(E_i) - \sum_{E_i\colon E_i \supseteq F} f(E_i) = 0 - \alpha = -\alpha$$

because of the minimality property of the set F.

Moreover, the identity

$$Z_F(H) = (-1)^p \alpha \quad \text{for all } H \subseteq F \text{ such that } |H| = |F| - p,\ 0 \le p \le |F|. \tag{A3}$$

holds. To show relation (A3) observe that

$$Z_F(H) = \sum_{E_i\colon E_i \cap F = H} f(E_i) = \sum_{j=0}^{p} (-1)^j \sum_{G\colon H \subset G \subset F,\ |G| = |H| + j} \ \sum_{E_i\colon E_i \supseteq G} f(E_i) \tag{A4}$$

for all sets $H \subset F$ with cardinality $|H| = |F| - p$. Identity (A4) holds since the term $f(E_i)$ is counted at the right-hand side of (A4) $\sum\limits_{j=0}^{l} (-1)^j \binom{l}{j} = (1-1)^l = 0$ times if $E_i \cap F = G$ with some $H \subset G \subseteq F$ with $|G| = |H| + l$ elements, $1 \le l \le p$, while in the case $E_i \cap F = H$ it is counted once. Relation (A4) together with (A2) and the minimality property of the set F imply relation (A3).

It follows from relation (A3) and the definition of the function $Z_F(H)$ that for all sets $H \subseteq F$ there exists some set E_i such that $H = E_i \cap F$, i.e. F is shattered by $\mathcal{E}$. Since $|F| \ge k$, this implies Theorem 5.1.

Proof of Theorem 5.3. Let us fix an arbitrary set $F = \{x_1, \dots, x_{k+1}\}$ of the set X, and consider the set of vectors $\mathcal{G}_k(F) = \{(g(x_1), \dots, g(x_{k+1}))\colon\ g \in \mathcal{G}_k\}$ of the $k+1$-dimensional space R^{k+1}. By the conditions of Theorem 5.3 $\mathcal{G}_k(F)$ 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 $\sum\limits_{j=1}^{k+1} a_j g(x_j) = 0$ for all $g \in \mathcal{G}_k$. We may assume that the set $A = A(a) = \{j\colon\ a_j < 0,\ 1 \le j \le k+1\}$ is non-empty, by multiplying the vector a by -1 if it is necessary.

Thus the identity

$$\sum_{j \in A} a_j g(x_j) = \sum_{j \in \{1,\dots,k+1\} \setminus A} (-a_j) g(x_j), \qquad \text{for all } g \in \mathcal{G}_k \tag{A5}$$

holds. Put $B = \{x_j\colon\ j \in A\}$. Then $B \subset F$, and $F \setminus B \ne \{x\colon\ g(x) \ge 0\} \cap F$ for all $g \in \mathcal{G}_k$. Indeed, if there were some $g \in \mathcal{G}_k$ such that $F \setminus B = \{x\colon\ g(x) \ge 0\} \cap F$, then the left-hand side of the equation (A5) would be strictly positive, its right-hand side would be non-positive for this $g \in \mathcal{G}_k$, and this is a contradiction.

The above proved property means that $\mathcal{D}$ shatters no set $F \subset X$ of cardinality $k+1$. Hence Theorem 5.1 implies that $\mathcal{D}$ is a Vapnik–Červonenkis class.

Appendix B. The proof of the diagram formula for Wiener–Itô integrals.

The proof of Theorem (10.2A) (the diagram formula for the product of two Wiener–Itô integrals) will be started with the proof of inequality (10.11). To show that this relation holds let us observe that the Cauchy inequality yields the following bound on the function F_γ defined in (10.10) (with the notation introduced there):

$$\begin{aligned} &F_\gamma^2(x_{(1,j)}, x_{(2,j')},\ (1,j) \in V_1(\gamma),\ (2,j') \in V_2(\gamma)) \\ &\quad \le \int f^2(x_{\alpha_\gamma(1,1)}, \dots, x_{\alpha_\gamma(1,k)}) \prod_{(2,j) \in \{(2,1),\dots,(2,l)\} \setminus V_2(\gamma)} \mu(\, dx_{(2,j)}) \\ &\qquad \int g^2(x_{(2,1)}, \dots, x_{(2,l)}) \prod_{(2,j) \in \{(2,1),\dots,(2,l)\} \setminus V_2(\gamma)} \mu(\, dx_{(2,j)}). \end{aligned} \tag{B1}$$

The expression at the right-hand side of inequality (B1) is the product of two functions with different arguments. The first function has arguments $x_{(1,j)}$ with $(1,j) \in V_1(\gamma)$ and the second one $x_{(2,j')}$ with $(2,j') \in V_2(\gamma)$. Integration of both sides in inequality (B1) by these arguments yields inequality (10.11).

Relation (10.12) will be proved first for the product of the Wiener–Itô integrals of two elementary functions. Let us consider two (elementary) functions $f(x_1, \dots, x_k)$ and $g(x_1, \dots, x_l)$ given in the following form: Let some disjoint sets $A_1, \dots, A_M$, $\mu(A_s) < \infty$, $1 \le s \le M$, be given together with some real numbers $c(s_1, \dots, s_k)$ indexed with k-tuples $(s_1, \dots, s_k)$, $1 \le s_j \le M$, $1 \le j \le k$, such that the numbers $s_1, \dots, s_k$ in a k-tuple are all different. Put $f(x_1, \dots, x_k) = c(s_1, \dots, s_k)$ on the rectangles $A_{s_1} \times \dots \times A_{s_k}$ with edges A_s, indexed with the above k-tuples, and let $f(x_1, \dots, x_k) = 0$ outside of these rectangles. Take similarly some disjoint sets $B_1, \dots, B_{M'}$, $\mu(B_t) < \infty$, $1 \le t \le M'$, and some real numbers $d(t_1, \dots, t_l)$, indexed with l-tuples $(t_1, \dots, t_l)$, $1 \le t_{j'} \le M'$, $1 \le j' \le l$, such that the numbers $t_1, \dots, t_l$ in an l-tuple are different. Put $g(x_1, \dots, x_l) = d(t_1, \dots, t_l)$ on the rectangles $B_{t_1} \times \dots \times B_{t_l}$ with edges indexed with the above introduced l-tuples, and let $g(x_1, \dots, x_l) = 0$ outside of these rectangles.

Let us take some small number $\varepsilon > 0$ and rewrite the above introduced functions $f(x_1, \dots, x_k)$ and $g(x_1, \dots, x_l)$ with the help of this number $\varepsilon > 0$ in the following way. Divide the sets $A_1, \dots, A_M$ to smaller sets $A_1^\varepsilon, \dots, A_{M(\varepsilon)}^\varepsilon$, $\bigcup_{s=1}^{M(\varepsilon)} A_s^\varepsilon = \bigcup_{s=1}^{M} A_s$, in such a way that all sets $A_1^\varepsilon, \dots, A_{M(\varepsilon)}^\varepsilon$ are disjoint, and $\mu(A_s^\varepsilon) \le \varepsilon$, $1 \le s \le M(\varepsilon)$. Similarly, take sets $B_1^\varepsilon, \dots, B_{M'(\varepsilon)}^\varepsilon$, $\bigcup_{t=1}^{M'(\varepsilon)} B_t^\varepsilon = \bigcup_{t=1}^{M'} B_t$, in such a way that all sets $B_1^\varepsilon, \dots, B_{M'(\varepsilon)}^\varepsilon$ are disjoint, and $\mu(B_t^\varepsilon) \le \varepsilon$, $1 \le t \le M'(\varepsilon)$. Beside this, let us also demand that two sets A_s^ε and B_t^ε, $1 \le s \le M(\varepsilon)$, $1 \le t \le M'(\varepsilon)$, are either disjoint or they agree. Such a partition exists for a non-atomic measure μ. (See the footnote about non-atomic measures before formula (10.8).) The above defined functions $f(x_1, \dots, x_k)$ and $g(x_1, \dots, x_l)$ can be rewritten by means of these new sets A_s^ε and B_t^ε. Namely, let $f(x_1, \dots, x_k) = c^\varepsilon(s_1, \dots, s_k)$ on the rectangles $A_{s_1}^\varepsilon \times \dots \times A_{s_k}^\varepsilon$ with $1 \le s_j \le M(\varepsilon)$, $1 \le j \le k$, with different indices $s_1, \dots, s_k$, where $c^\varepsilon(s_1, \dots, s_k) = c(p_1, \dots, p_k)$ with

those indices $(p_1, \dots, p_k)$ for which $A^\varepsilon_{s_1} \times \dots \times A^\varepsilon_{s_k} \subset A_{p_1} \times \dots \times A_{p_k}$. The function f disappears outside of these rectangles. The function $g(x_1, \dots, x_l)$ can be written similarly in the form $g(x_1, \dots, x_l) = d^\varepsilon(t_1, \dots, t_l)$ on the rectangles $B^\varepsilon_{t_1} \times \dots \times B^\varepsilon_{t_l}$ with $1 \le t_{j'} \le M'(\varepsilon)$, $1 \le j' \le l$, and different indices, $t_1, \dots, t_l$. Beside this, the function g disappears outside of these rectangles.

The above representation of the functions f and g through a parameter ε is useful, since it enables us to give a good asymptotic formula for the product $k!Z_{\mu,k}(f)l!Z_{\mu,l}(g)$ which yields the diagram formula for the product of Wiener–Itô integrals of elementary functions with the help of a limiting procedure $\varepsilon \to 0$.

Fix a small number $\varepsilon > 0$, take the representation of the functions f and g with its help, and write

$$k!Z_{\mu,k}(f)l!Z_{\mu,l}(g) = \sum_{\gamma \in \Gamma(k,l)} Z_\gamma(\varepsilon) \tag{B2}$$

with

$$Z_\gamma(\varepsilon) = \sum\nolimits^{\gamma} c^\varepsilon(s_1, \dots, s_k) d^\varepsilon(t_1, \dots, t_l) \mu_W(A^\varepsilon_{s_1}) \dots \mu_W(A^\varepsilon_{s_k}) \mu_W(B^\varepsilon_{t_1}) \dots \mu_W(B^\varepsilon_{t_l}), \tag{B3}$$

where $\Gamma(k,l)$ denotes the class of diagrams introduced before the formulation of Theorem 10.2A, and $\sum^\gamma$ denotes summation for such $k+l$-tuples $(s_1, \dots, s_k, t_1, \dots, t_l)$, $1 \le s_j \le M(\varepsilon)$, $1 \le j \le k$, and $1 \le t_{j'} \le M'(\varepsilon)$, $1 \le j' \le l$, for which $A^\varepsilon_{s_j} = B^\varepsilon_{t_{j'}}$ if $((1,j),(2,j')) \in E(\gamma)$, i.e. if it is an edge of γ, and otherwise all sets $A^\varepsilon_{s_j}$ and $B^\varepsilon_{t_{j'}}$ are disjoint. (This sum also depends on ε.) In the case of an empty sum $Z_\gamma(\varepsilon)$ equals zero.

For all $\gamma \in \Gamma(k,l)$ the expression Z_γ will be written in the form

$$Z_\gamma(\varepsilon) = Z^{(1)}_\gamma(\varepsilon) + Z^{(2)}_\gamma(\varepsilon), \quad \gamma \in \Gamma(k,l), \tag{B4}$$

with

$$\begin{aligned} Z^{(1)}_\gamma(\varepsilon) = \sum\nolimits^{\gamma} & c^\varepsilon(s_1, \dots, s_k) d^\varepsilon(t_1, \dots, t_l) \\ & \prod_{j:\,(1,j) \in V_1(\gamma)} \mu_W(A^\varepsilon_{s_j}) \prod_{j':\,(2,j') \in V_2(\gamma)} \mu_W(B^\varepsilon_{t_{j'}}) \\ & \prod_{j:\,(1,j) \in \{(1,1),\dots,(1,k)\} \setminus V_1(\gamma)} \mu(A^\varepsilon_{s_j}) \end{aligned} \tag{B5}$$

and

$$\begin{aligned} Z^{(2)}_\gamma(\varepsilon) = \sum\nolimits^{\gamma} & c^\varepsilon(s_1, \dots, s_k) d^\varepsilon(t_1, \dots, t_l) \\ & \prod_{j:\,(1,j) \in V_1(\gamma)} \mu_W(A^\varepsilon_{s_j}) \prod_{j':\,(2,j') \in V_2(\gamma)} \mu_W(B^\varepsilon_{t_{j'}}) \\ & \Bigg[\prod_{j:\,(1,j) \in \{(1,1),\dots,(1,k)\} \setminus V_1(\gamma)} \mu_W(A^\varepsilon_{s_j}) \prod_{j':\,(2,j') \in \{(2,1),\dots,(2,l)\} \setminus \in V_2(\gamma)} \mu_W(B^\varepsilon_{t_{j'}}) \end{aligned}$$

$$- \prod_{j\colon (1,j)\in\{(1,1),\dots,(1,k)\}\setminus V_1(\gamma)} \mu(A^\varepsilon_{s_j})\Big], \tag{B6}$$

where $V_1(\gamma)$ and $V_2(\gamma)$ (introduced before formula (10.9) during the preparation to the formulation of Theorem 10.2A) are the sets of vertices in the first and second row of the diagram γ from which no edge starts.

I claim that there is some constant $C > 0$ not depending on ε such that

$$E\left(|\gamma|!Z_{\mu,|\gamma|}(F_\gamma) - Z^{(1)}_\gamma(\varepsilon)\right)^2 \le C\varepsilon \quad \text{for all } \gamma \in \Gamma(k,l) \tag{B7}$$

with the Wiener–Itô integral with the kernel function F_γ defined in (10.9), (10.9a) and (10.10), and

$$E\left(Z^{(2)}_\gamma(\varepsilon)\right)^2 \le C\varepsilon \quad \text{for all } \gamma \in \Gamma(k,l). \tag{B8}$$

Relations (B7) and (B8) imply relation (10.12) if f and g are elementary functions. Indeed, they imply that

$$\lim_{\varepsilon\to 0} \big\| \, |\gamma|!Z_{\mu,|\gamma|}(F_\gamma) - Z_\gamma(\varepsilon)\big\|_2 \to 0 \quad \text{for all } \gamma \in \Gamma(k,l),$$

and this relation together with (B2) yield relation (10.12) with the help of a limiting procedure $\varepsilon \to 0$.

To prove relation (B7) let us introduce the function

$$\begin{aligned} &F^\varepsilon_\gamma(x_{(1,j)}, x_{(2,j')},\ (1,j)\in V_1(\gamma),\ (2,j')\in V_2(\gamma)) \\ &\quad = F_\gamma(x_{(1,j)}, x_{(2,j')},\ (1,j)\in V_1(\gamma),\ (2,j')\in V_2(\gamma)) \\ &\qquad \text{for } x_{(1,j)} \in A^\varepsilon_{s_j},\ \text{for all } (1,j)\in V_1(\gamma), \\ &\qquad \text{and } x_{(2,j')} \in B^\varepsilon_{t_{j'}},\ \text{for all } (2,j')\in V_2(\gamma)), \end{aligned}$$

if all sets $A^\varepsilon_{s_j}$, $(1,j)\in V_1(\gamma)$, and $B^\varepsilon_{t_{j'}}$, $(2,j')\in V_2(\gamma)$ in the above formula are different. (As a consequence, they are disjoint.) Put

$$F^\varepsilon_\gamma(x_{(1,j)}, x_{(2,j')},\ (1,j)\in V_1(\gamma),\ (2,j')\in V_2(\gamma)) = 0 \quad \text{otherwise.}$$

The function F^ε_γ is elementary, and a comparison of its definition with relation (B5) and the definition of the function F_γ yield that

$$Z^{(1)}_\gamma(\varepsilon) = |\gamma|!Z_{\mu,|\gamma|}(F^\varepsilon_\gamma). \tag{B9}$$

The function F^ε_γ slightly differs from F_γ, since the function F_γ need not disappear in all such points $(x_{(1,j)}, x_{(2,j')},\ (1,j)\in V_1(\gamma), (2,j')\in V_2(\gamma))$ for which there is some pair (j,j') such that the relations $x_{(1,j)}\in A^\varepsilon_{s_j}$ and $x_{(2,j')}\in B^\varepsilon_{t_{j'}}$ hold with such sets for which $A^\varepsilon_{s_j} = B^\varepsilon_{t_{j'}}$, while F^ε_γ must be zero in such points. On the other hand, in the

case $|\gamma| = \max(k,l) - \min(k,l)$, i.e. if one of the sets $V_1(\gamma)$ or $V_2(\gamma)$ is empty, $F_\gamma = F_\gamma^\varepsilon$, $Z_\gamma^{(1)} = |\gamma|! Z_{\mu,|\gamma|}(F_\gamma)$, and relation (B7) clearly holds for such diagrams γ.

It will be shown that in the case $|\gamma| = \max(k,l) - \min(k,l) > 0$ the set where $F_\gamma \neq F_\gamma^\varepsilon$ is small, and this implies relation (B7).

Let us define the sets $A = \bigcup_{s=1}^{M(\varepsilon)} A_s^\varepsilon$ and $B = \bigcup_{t=1}^{M'(\varepsilon)} B_t^\varepsilon$. These sets A and B do not depend on the parameter ε. Beside this $\mu(A) < \infty$, and $\mu(B) < \infty$. Define for all pairs (j_0, j_0') such that $(1, j_0) \in V_1(\gamma)$, $(2, j_0') \in V_2(\gamma)$ the set

$$\begin{aligned} D(j_0, j_0') = \{&(x_{(1,j)}, x_{(2,j')},\ (1,j) \in V_1(\gamma),\ (2,j') \in V_2(\gamma)): \\ &x_{(1,j_0)} \in A_{s_{j_0}}^\varepsilon,\ x_{(1,j_0')} \in B_{t_{j_0'}}^\varepsilon \quad \text{for some } s_{j_0} \text{ and } t_{j_0'} \text{ such that } A_{s_{j_0}}^\varepsilon = B_{t_{j_0'}}^\varepsilon \\ &x_{(1,j)} \in A \text{ for all } (1,j) \in V_1(\gamma), \text{ and } x_{(2,j')} \in B \text{ for all } (2,j') \in V_2(\gamma)\}. \end{aligned}$$

Introduce the notation $x^\gamma = (x_{(1,j)}, x_{(2,j')}),\ (1,j) \in V_1(\gamma),\ (2,j') \in V_2(\gamma))$ and put $D_\gamma = \{x^\gamma\colon\ F_\gamma^\varepsilon(x^\gamma) \neq F_\gamma(x^\gamma)\}$. The relation $D_\gamma \subset \bigcup_{j=1}^{k} \bigcup_{j'=1}^{l} D(j_0, j_0')$ holds, since if $F_\gamma^\varepsilon(x^\gamma) \neq F_\gamma(x^\gamma)$ for some vector x^γ, then it has some coordinates $(1, j_0) \in V_1(\gamma)$ and $(2, j_0') \in V_2(\gamma)$ such that $x_{(1,j_0)} \in A_{s_{j_0}}^\varepsilon$ and $x_{(1,j_0')} \in B_{t_{j_0'}}^\varepsilon$ with some sets $A_{s_{j_0}}^\varepsilon = B_{t_{j_0'}}^\varepsilon$, and the relation in the last line of the definition of $D(j_0, j_0')$ must also hold for this vector x^γ, since otherwise $F_\gamma(x_\gamma) = 0 = F_\gamma^\varepsilon(x_\gamma)$. I claim that there is some constant C_1 such that

$$\mu^{|V_1(\gamma)|+|V_2(\gamma)|}(D(j_0, j_0')) \le C_1 \varepsilon \quad \text{for all sets } D(j_0, j_0'),$$

where $\mu^{|V_1(\gamma)|+|V_2(\gamma)|}$ denotes the direct product of the measure μ on some copies of the original space $(X, \mathcal{X})$ indexed by $(1,j) \in V_1(\gamma)$ and $(2,j') \in V_2(\gamma)$. To see this relation one has to observe that $\sum_{A_{s_{j_0}}^\varepsilon = B_{t_{j_0'}}^\varepsilon} \mu(A_{s_{j_0}}^\varepsilon)\mu(B_{t_{j_0'}}^\varepsilon) \le \sum \varepsilon \mu(A_{s_{j_0}}^\varepsilon) = \varepsilon \mu(A)$. Thus the set $D(j_0, j_0')$ can be covered by the direct product of a set whose μ measure is not greater than $\varepsilon\mu(A)$ and of a rectangle whose edges are either the set A or the set B.

The above relations imply that

$$\mu^{|V_1(\gamma)|+|V_2(\gamma)|}(D_\gamma) \le C_2 \varepsilon \tag{B10}$$

with some constant $C_2 > 0$.

Relation (B9), estimate (B10), the property c) formulated in Theorem 10.1 for Wiener–Itô integrals and the observation that the function $F_\gamma = F_\gamma(f,g)$ is bounded in supremum norm if f and g are elementary functions imply the inequality

$$\begin{aligned} E\left(|\gamma|! Z_{\mu,|\gamma|}(F_\gamma) - Z_\gamma^{(1)}(\varepsilon)\right)^2 &= |\gamma!|^2 E\left(Z_{\mu,|\gamma|}(F_\gamma - F_\gamma^\varepsilon)\right)^2 \le |\gamma|! \|F_\gamma - F_\gamma^\varepsilon\|_2^2 \\ &\le K \mu^{|V_1(\gamma)|+|V_2(\gamma)|}(D_\gamma) \le C\varepsilon. \end{aligned}$$

This means that relation (B7) holds.

To prove relation (B8) write $E\left(Z_\gamma^{(2)}(\varepsilon)\right)^2$ in the following form:

$$E\left(Z_\gamma^{(2)}(\varepsilon)\right)^2 = \sum\nolimits^{\gamma}\sum\nolimits^{\gamma} c^{\varepsilon}(s_1,\dots,s_k)d^{\varepsilon}(t_1,\dots,t_l)c^{\varepsilon}(\bar{s}_1,\dots,\bar{s}_k)d^{\varepsilon}(\bar{t}_1,\dots,\bar{t}_l) \\ EU(s_1,\dots,s_k,t_1,\dots,t_l,\bar{s}_1,\dots,\bar{s}_k,\bar{t}_1,\dots,\bar{t}_l) \tag{B11}$$

with

$$\begin{aligned} &U(s_1,\dots,s_k,t_1,\dots,t_l,\bar{s}_1,\dots,\bar{s}_k,\bar{t}_1,\dots,\bar{t}_l) \\ &\quad= \prod_{j:\,(1,j)\in V_1(\gamma)} \mu_W(A^{\varepsilon}_{s_j}) \prod_{j':\,(2,j')\in V_2(\gamma)} \mu_W(B^{\varepsilon}_{t_{j'}}) \\ &\qquad \prod_{\bar{\jmath}:\,(1,\bar{\jmath})\in V_1(\gamma)} \mu_W(A^{\varepsilon}_{\bar{s}_{\bar{\jmath}}}) \prod_{\bar{\jmath}':\,(2,\bar{\jmath}')\in V_2(\gamma)} \mu_W(B^{\varepsilon}_{\bar{t}_{\bar{\jmath}'}}) \\ &\qquad \Bigg[\prod_{j:\,(1,j)\in\{(1,1),\dots,(1,k)\}\setminus V_1(\gamma)} \mu_W(A^{\varepsilon}_{s_j}) \prod_{j':\,(2,j')\in\{(2,1),\dots,(2,l)\}\setminus\in V_2(\gamma)} \mu_W(B^{\varepsilon}_{t_{j'}}) \\ &\qquad\quad - \prod_{j:\,(1,j)\in\{(1,1),\dots,(1,k)\}\setminus V_1(\gamma)} \mu(A^{\varepsilon}_{s_j})\Bigg] \\ &\qquad \Bigg[\prod_{\bar{\jmath}:\,(1,\bar{\jmath})\in\{(1,1),\dots,(1,k)\}\setminus V_1(\gamma)} \mu_W(A^{\varepsilon}_{\bar{s}_{\bar{\jmath}}}) \prod_{\bar{\jmath}':\,(2,\bar{\jmath}')\in\{(2,1),\dots,(2,l)\}\setminus\in V_2(\gamma)} \mu_W(B^{\varepsilon}_{\bar{t}_{\bar{\jmath}'}}) \\ &\qquad\quad - \prod_{\bar{\jmath}:\,(1,j)\in\{(1,1),\dots,(1,k)\}\setminus V_1(\gamma)} \mu(A^{\varepsilon}_{\bar{s}_{\bar{\jmath}}})\Bigg]. \end{aligned} \tag{B12}$$

The double sum $\sum^{\gamma}\sum^{\gamma}$ in (B11) has to be understood in the following way. The first summation is taken for vectors $(s_1,\dots,s_k,t_1,\dots,t_l)$, and these vectors take such values which were defined in $\sum^{\gamma}$ in formula (B3). The second summation is taken for vectors $(\bar{s}_1,\dots,\bar{s}_k,\bar{t}_1,\dots,\bar{t}_l)$, and again with values defined in the summation $\sum^{\gamma}$.

Relation (B8) will be proved by means of some estimates about the expectation of the above defined random variable $U(\cdot)$ which will be presented in the following Lemma B. Before their formulation I introduce the following Properties A and B.

Property A. *A sequence $s_1,\dots,s_k,t_1,\dots,t_l,\bar{s}_1,\dots,\bar{s}_k,\bar{t}_1,\dots,\bar{t}_l$, with elements $1\le s_j,\bar{s}_{\bar{\jmath}}\le M(\varepsilon)$, for $1\le j,\bar{\jmath}\le k$, and $1\le t_j,\bar{t}_{\bar{\jmath}'}\le M'(\varepsilon)$ for $1\le j',\bar{\jmath}'\le l$, satisfies Property A (depending on a fixed diagram γ and number $\varepsilon>0$) if the sequences of sets $\{A^{\varepsilon}_{s_j},B^{\varepsilon}_{t_{j'}},(1,j)\in V_1(\gamma),(2,j')\in V_2(\gamma)\}$ and $\{A^{\varepsilon}_{\bar{s}_{\bar{\jmath}}},B^{\varepsilon}_{\bar{t}_{\bar{\jmath}'}},(1,\bar{\jmath})\in V_1(\gamma),(2,\bar{\jmath}')\in V_2(\gamma)\}$ agree. (Here we say that two sequences agree if they contain the same elements in a possibly different order.)*

Property B. *A sequence $s_1,\dots,s_k,t_1,\dots,t_l,\bar{s}_1,\dots,\bar{s}_k,\bar{t}_1,\dots,\bar{t}_l$, with elements $1\le s_j,\bar{s}_{\bar{\jmath}}\le M(\varepsilon)$, for $1\le j,\bar{\jmath}\le k$, and $1\le t_j,\bar{t}_{\bar{\jmath}'}\le M'(\varepsilon)$ for $1\le j',\bar{\jmath}'\le l$, satisfies Property B (depending on a fixed diagram γ and number $\varepsilon>0$) if the sequences of sets*

$$\{A^{\varepsilon}_{s_j},B^{\varepsilon}_{t_{j'}},(1,j)\in\{(1,1),\dots,(1,k)\}\setminus V_1(\gamma),\ (2,j')\in\{(2,1),\dots,(2,l)\}\setminus V_2(\gamma)\}$$

and

$$\{A^{\varepsilon}_{\bar{s}_{\bar{\jmath}}}, B^{\varepsilon}_{\bar{t}_{\bar{\jmath}'}}, (1,\bar{\jmath}) \in \{(1,1),\dots,(1,k)\} \setminus V_1(\gamma),\ (2,\bar{\jmath}') \in \{(2,1),\dots,(2,l)\} \setminus V_2(\gamma)\}$$

have at least one common element.

(In the above definitions two sets A^{ε}_s and B^{ε}_t are identified if $A^{\varepsilon}_s = B^{\varepsilon}_t$.)

Now I formulate the following

Lemma B. *Let us consider the function $U(\cdot)$ introduced in formula (B12). Assume that its arguments $s_1,\dots,s_k,t_1,\dots,t_l,\bar{s}_1,\dots,\bar{s}_k,\bar{t}_1,\dots,\bar{t}_l$ are chosen in such a way that the function $U(\cdot)$ with these arguments appears in the double sum $\sum^{\gamma}\sum^{\gamma}$ in formula (B11), i.e. $A^{\varepsilon}_{s_j} = B^{\varepsilon}t_{j'}$ if $((1,j),(2,j')) \in E(\gamma)$, otherwise all sets $A^{\varepsilon}_{s_j}$ and $B^{\varepsilon}_{t_{j'}}$ are disjoint, and an analogous statement holds if the coordinates $s_1,\dots,s_k,t_1,\dots,t_l$ are replaced by $\bar{s}_1,\dots,\bar{s}_k,\bar{t}_1,\dots,\bar{t}_l$. Then*

$$EU(s_1,\dots,s_k,t_1,\dots,t_l,\bar{s}_1,\dots,\bar{s}_k,\bar{t}_1,\dots,\bar{t}_l) = 0 \tag{B13}$$

if the sequence of the arguments in $U(\cdot)$ does not satisfies either Property A or Property B.

If the sequence of the arguments in $U(\cdot)$ satisfies both Property A and Property B, then

$$\begin{aligned} &|EU(s_1,\dots,s_k,t_1,\dots,t_l,\bar{s}_1,\dots,\bar{s}_k,\bar{t}_1,\dots,\bar{t}_l)| \\ &\quad \le C\varepsilon \prod{}' \mu(A^{\varepsilon}_{s_j})\mu(A^{\varepsilon}_{\bar{s}_{\bar{\jmath}}})\mu(B^{\varepsilon}_{t_{j'}})\mu(B^{\varepsilon}_{\bar{t}_{\bar{\jmath}'}}) \end{aligned} \tag{B14}$$

with some appropriate constant $C = C(k,l) > 0$ depending only on the number of variables k and l of the functions f and g. The prime in the product $\prod'$ at the right-hand side of (B14) means that in this product the measure μ of those sets $A^{\varepsilon}_{s_j}$, $A^{\varepsilon}_{\bar{s}_{\bar{\jmath}}}$, $B^{\varepsilon}_{t_{j'}}$ and $B^{\varepsilon}_{\bar{t}_{\bar{\jmath}'}}$ are considered, whose indices are listed among the arguments $s_j, \bar{s}_{\bar{\jmath}}, t_{j'}$ or $\bar{t}_{\bar{\jmath}'}$ of $U(\cdot)$, and the measure μ of each such set appears exactly once. (This means e.g. that if $A^{\varepsilon}_{s_j} = B^{\varepsilon}_{t_{j'}}$ or $A^{\varepsilon}_{s_j} = B^{\varepsilon}_{\bar{t}_{\bar{\jmath}'}}$ for some indices j and j' or $\bar{\jmath}'$, then one of the terms between $\mu(A^{\varepsilon}_{s_j})$ and $\mu(B^{\varepsilon}_{t_{j'}})$ or $\mu(B^{\varepsilon}_{\bar{t}_{\bar{\jmath}'}})$ is omitted from the product. For the sake of definitiveness let us preserve the set $\mu(A^{\varepsilon}_{j_s})$ in such a case.)

The proof of Lemma B. Let us prove first relation (B13). It will be exploited that for disjoint sets the random variables $\mu_W(A_s)$ and $\mu_W(B_t)$ are independent, and this provides a good factorization of the expectation of certain products. Let us carry out the multiplications in the definition of $U(\cdot)$ in formula (B12), and show that each product obtained in such a way has zero expectation. If Property A does not hold for the arguments of $U(\cdot)$, and beside this the arguments $s_1,\dots,s_k,t_1,\dots,t_l,\bar{s}_1,\dots,\bar{s}_k,\bar{t}_1,\dots,\bar{t}_l$ satisfy the remaining conditions of Lemma B, then each product we consider contains a factor $\mu_W(A^{\varepsilon}_{s_{j_0}})$, $(1,j_0) \in V_1(\gamma)$, which is independent of all those terms in this product which are in the following list: $\mu_W(A^{\varepsilon}_{s_j})$ with some $j \ne j_0$, $1 \le j \le k$, or $\mu_W(B^{\varepsilon}_{t_{j'}})$, $1 \le j \le l$, or $\mu_W(A^{\varepsilon}_{\bar{s}_{\bar{\jmath}}})$ with $(1,\bar{\jmath}) \in V_1(\gamma)$, or $\mu_W(B^{\varepsilon}_{\bar{t}_{\bar{\jmath}'}})$ with $(2,\bar{\jmath}') \in V_2(\gamma)$. We will show with the

help of this property that the expectation of each term has a factorization with a factor either of the form $E\mu_W(A^\varepsilon_{s_{j_0}}) = 0$ or $E\mu_W(A^\varepsilon_{s_{j_0}})^3 = 0$, hence it equals zero. Indeed, although the above properties do not exclude the appearance of such a pair of arguments $A^\varepsilon_{t_{\bar{j}'}}$, $(1,\bar{j}') \in \{(1,1),\dots,(1,k)\} \setminus V_1(\gamma)$ and $B^\varepsilon_{t_{\bar{j}'}}$, $(2,\bar{j}') \in \{(2,1),\dots,(2,l)\} \setminus V_2(\gamma)$ in the product for which $A^\varepsilon_{t_{\bar{j}}} = B^\varepsilon_{t_{\bar{j}'}} = A^\varepsilon_{s_{j_0}}$, and in such a case a term of the form $E\mu_W(A^\varepsilon_{s_{j_0}})$ will not appear in the product, but if this happens, then the product contains a factor of the form $E\mu_W(A^\varepsilon_{s_{j_0}})^3 = 0$. Hence an appropriate factorization of each term of $EU(\cdot)$ contains either a factor of the form $E\mu_W(A^\varepsilon_{s_{j_0}}) = 0$ or $E\mu_W(A^\varepsilon_{s_{j_0}})^3 = 0$ if $U(\cdot)$ does not satisfy Property A.

To finish the proof of relation (B13) it is enough consider the case when the arguments of $U(\cdot)$ satisfy Property A, but they do not satisfy Property B. The validity of Property A implies that the sets $\{A^\varepsilon_{s_j},\, j \in V_1\} \cup \{B^\varepsilon_{t_{j'}},\, j' \in V_2\}$ and $\{A^\varepsilon_{\bar{s}_j},\, j \in V_1\} \cup \{B^\varepsilon_{t_{j'}},\, j' \in V_2\}$ agree. The conditions of Lemma B also imply that the elements of these sets are such sets which are disjoint of the sets $A^\varepsilon_{s_j}$, $B^\varepsilon_{t_{j'}}$, $A^\varepsilon_{\bar{s}_{\bar{j}}}$ and $B^\varepsilon_{t_{\bar{j}'}}$ with indices $(1,j),(1,\bar{j}) \in \{(1,1),\dots,(1,k)\} \setminus V_1(\gamma)$ and $(2,j'),(2,\bar{j}') \in \{(2,1),\dots,(2,l)\} \setminus V_2(\gamma)$. If Property B does not hold, then the latter class of sets can be divided into two subclasses in such a way that the elements in different subclasses are disjoint. The first subclass consists of the sets $A^\varepsilon_{s_j}$ and $B^\varepsilon_{t_{j'}}$, and the second one of the sets $A^\varepsilon_{\bar{s}_{\bar{j}}}$ and $B^\varepsilon_{t_{\bar{j}'}}$ with indices such that $(1,j),(1,\bar{j}) \in \{(1,1),\dots,(1,k)\} \setminus V_1(\gamma)$ and $(2,j'),(2,\bar{j}') \in \{(2,1),\dots,(2,l)\} \setminus V_2(\gamma)$. These facts imply that $EU(\cdot)$ has a factorization, which contains the term

$$E\Bigg[\prod_{j\colon (1,j)\in\{(1,1),\dots,(1,k)\}\setminus V_1(\gamma)} \mu_W(A^\varepsilon_{s_j}) \prod_{j'\colon (2,j')\in\{(2,1),\dots,(2,l)\}\setminus\in V_2(\gamma)} \mu_W(B^\varepsilon_{t_{j'}}) - \prod_{j\colon (1,j)\in\{(1,1),\dots,(1,k)\}\setminus V_1(\gamma)} \mu(A^\varepsilon_{s_j})\Bigg] = 0,$$

hence relation (B13) holds also in this case. The last expression has zero expectation, since if we take such pairs $A^\varepsilon_{s_j}, B^\varepsilon_{t'_j}$ for the sets appearing in it for which that $((1,j),(2,j')) \in E(\gamma)$, i.e. these vertices are connected with an edge of γ, then $A^\varepsilon_{s_j} = B^\varepsilon_{t'_j}$ in a pair, and elements in different pairs are disjoint. This observation allows a factorization in the product whose expectation is taken, and then the identity $E\mu_W(A^\varepsilon_{s_j})\mu_W(B^\varepsilon_{t_{j'}}) = \mu(A^\varepsilon_{s_j})$ implies the desired identity.

To prove relation (B14) if the arguments of the function $U(\cdot)$ satisfy both Properties A and B consider the expression (B12) which defines $U(\cdot)$, carry out the term by term multiplication between the two differences at the end of this formula, take expectation for each term of the sum obtained in such a way, and factorize them. Since $E\mu_W(A)^2 = \mu(A)$, $E\mu_W(A)^4 = 3\mu(A)^2$ for all sets $A \in \mathcal{X}$, $\mu(A) < \infty$, some calculation shows that each term can be expressed as constant times a product whose elements are those probabilities $\mu(A^\varepsilon_s)$ and $\mu(B^\varepsilon_t)$ or their square which appear at the right-hand side of (B14). Moreover, since the arguments of $U(\cdot)$ satisfy Property B, there will be

at least one term of the form $\mu(A_s^\varepsilon)^2$ in this product. Since $\mu(A_s^\varepsilon)^2 \le \varepsilon\mu(A_s^\varepsilon)$, these calculations provide formula (B14). Lemma B is proved.

Relation (B11) implies that

$$E\left(Z_\gamma^{(2)}(\varepsilon)\right)^2 \le K \sum{}^{\gamma} \sum{}^{\gamma} |EU(s_1,\dots,s_k,t_1,\dots,t_l,\bar s_1,\dots,\bar s_k,\bar t_1,\dots,\bar t_l)| \tag{B15}$$

with some appropriate $K > 0$. By Lemma B it is enough to sum up only for such terms $U(\cdot)$ in (B15) whose arguments satisfy both Properties A and B. Moreover, each such term can be bounded by means of inequality (B14). Let us list the sets $A_{s_j}^\varepsilon, A_{\bar s_{\bar\jmath}}^\varepsilon, B_{t_{j'}}^\varepsilon, B_{\bar t_{\bar\jmath'}}^\varepsilon$ appearing in the upper bound at the right-hand side of (B14) for all functions $U(\cdot)$ taking part in the sum at the right-hand side of (B15). Since all fixed sequences of the sets A_s^ε and B_t^ε appear less than $C(k,l)$ times with an appropriate constant $C(k,l)$ depending only on the order k and l of the integrals we are considering, and $\sum\limits_{s=1}^{M(\varepsilon)} \mu(A_s^\varepsilon) + \sum\limits_{t=1}^{M'(\varepsilon)} \mu(B_t^\varepsilon) = \mu(A) + \mu(B) < \infty$, the above relations imply that

$$E\left(Z_\gamma^{(2)}(\varepsilon)\right)^2 \le C_1\varepsilon \sum_{j=1}^{k+l} (\mu(A)+\mu(B))^j \le C\varepsilon.$$

Hence relation (B8) holds.

To prove Theorem 10.2A in the general case take for all pairs of functions $f \in \mathcal{H}_{\mu,k}$ and $g \in \mathcal{H}_{\mu,l}$ two sequences of elementary functions $f_n \in \bar{\mathcal{H}}_{\mu,k}$ and $g_n \in \bar{\mathcal{H}}_{\mu,l}$, $n = 1,2,\dots$, such that $\|f_n - f\|_2 \to 0$ and $\|g_n - g\|_2 \to 0$ as $n \to \infty$. Let us introduce the notation $F_\gamma(f,g) = F_\gamma$ if the function F_γ is defined with the help of the functions f and g. It is enough to show that

$$E|k!Z_{\mu,k}(f)l!Z_{\mu,l}(g) - k!Z_{\mu,k}(f_n)l!Z_{\mu,l}(g_n)| \to 0 \quad \text{as } n \to \infty, \tag{B16}$$

and

$$|\gamma|!E\left|Z_{\mu,|\gamma|}(F_\gamma(f,g)) - Z_{\mu,|\gamma|}(F_\gamma(f_n,g_n))\right| \to 0 \quad \text{as } n \to \infty \quad \text{for all } \gamma \in \Gamma(k,l), \tag{B17}$$

since then a simple limiting procedure $n \to \infty$, and the already proved part of the theorem for Wiener–Itô integrals of elementary functions imply Theorem 10.2A.

To prove relation (B16) write

$$\begin{aligned}
&E|k!Z,\mu,k(f)l!Z_{\mu,l}(g) - k!Z_{\mu,k}(f_n)l!Z_{\mu,l}(g_n)| \\
&\quad \le k!l!\left(E|Z_{\mu,k}(f)Z_{\mu,l}(g-g_n)| + E|Z_{\mu,k}(f-f_n)Z_{\mu,l}(g_n))\right| \\
&\quad \le k!l!\left(\left(EZ_{\mu,k}^2(f)\right)^{1/2}\left(EZ_{\mu,l}^2(g-g_n)\right)^{1/2} + \left(EZ_{\mu,k}^2(f-f_n)\right)^{1/2}\left(EZ_{\mu,l}^2(g_n)\right)^{1/2}\right) \\
&\quad \le (k!l!)^{1/2}\left(\|f\|_2\|g-g_n\|_2 + \|f-f_n\|_2\|g_n\|_2\right).
\end{aligned}$$

Relation (B16) follows from this inequality with a limiting procedure $n \to \infty$.

To prove relation (B17) write

$$\begin{aligned}
&|\gamma|!E\left|Z_{\mu,|\gamma|}(F_\gamma(f,g)) - Z_{\mu,|\gamma|}(F_\gamma(f_n,g_n))\right| \\
&\quad \le |\gamma|!E\left|Z_{\mu,|\gamma|}(F_\gamma(f,g-g_n))\right| + |\gamma|!E\left|Z_{\mu,|\gamma|}(F_\gamma(f-f_n,g_n))\right| \\
&\quad \le |\gamma|!\left(EZ^2_{\mu,|\gamma|}(F_\gamma(f,g-g_n))\right)^{1/2} + |\gamma|!\left(EZ^2_{\mu,|\gamma|}(F_\gamma(f-f_n,g_n))\right)^{1/2} \\
&\quad \le (|\gamma|!)^{1/2}\left(\|F_\gamma(f,g-g_n)\|_2 + \|F_\gamma(f-f_n,g_n)\|_2\right),
\end{aligned}$$

and observe that by relation (10.11) $\|F_\gamma(f,g-g_n)\|_2 \le \|f\|_2\|g-g_n\|_2$, and $\|F_\gamma(f-f_n,g_n)\|_2 \le \|f-f_n\|_2\|g_n\|_2$. Hence

$$\begin{aligned}
&|\gamma|!E\left|Z_{\mu,|\gamma|}(F_\gamma(f,g)) - Z_{\mu,|\gamma|}(F_\gamma(f_n,g_n))\right| \\
&\qquad \le (|\gamma|!)^{1/2}\left(\|f\|_2\|g-g_n\|_2 + \|f-f_n\|_2\|g_n\|_2\right).
\end{aligned}$$

The last inequality implies relation (B17) with a limiting procedure $n \to \infty$. Theorem 10.2A is proved.

Appendix C. The proof of some results about Wiener–Itô integrals.

First I prove Itô's formula about multiple Wiener–Itô integrals (Theorem 10.3). The proof is based on the diagram formula for Wiener–Itô integrals and a recursive formula about Hermite polynomials proved in Proposition C. In Proposition C2 I present the proof of another important property of Hermite polynomials. This result states that the class of all Hermite polynomials is a *complete* orthogonal system in an appropriate Hilbert space. It is needed in the proof of Theorem 10.5 about the isomorphism of Fock spaces to the Hilbert space generated by Wiener–Itô integrals. At the end of Appendix C the proof of Theorem 10.4, a limit theorem about degenerated U-statistics is given.

Proposition C about some properties of Hermite polynomials. *The functions*

$$H_k(x) = (-1)^k e^{x^2/2}\frac{d^k}{dx^k}e^{-x^2/2}, \quad k = 0,1,2,\dots \tag{C1}$$

are the Hermite polynomials with leading coefficient 1, i.e. $H_k(x)$ is a polynomial of order k with leading coefficient 1 such that

$$\int_{-\infty}^{\infty} H_k(x)H_l(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx = 0 \quad \textit{if } k \ne l, \tag{C2}$$

and

$$\int_{-\infty}^{\infty} H_k^2(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx = k! \quad \textit{for all } k = 0,1,2\dots. \tag{C2$'$}$$

The recursive relation

$$H_k(x) = xH_{k-1}(x) - (k-1)H_{k-2}(x) \tag{C3}$$

holds for all $k = 1, 2, \dots$.

Remark. It is more convenient to consider relation (C3) valid also in the case $k = 1$. In this case $H_1(x) = x$, $H_0(x) = 1$, and relation holds with an arbitrary function $H_{-1}(x)$.

Proof of Proposition C. It is clear from formula (C1) that $H_k(x)$ is a polynomial of order k with leading coefficient 1. Take $l \ge k$, and write by means of integration by parts

$$\begin{aligned}\int_{-\infty}^{\infty} H_k(x)H_l(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}H_k(x)(-1)^l\frac{d^l}{dx^l}e^{-x^2/2}\,dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\frac{d}{dx}H_k(x)(-1)^{l-1}\frac{d^{l-1}}{dx^{l-1}}e^{-x^2/2}\,dx.\end{aligned}$$

Successive partial integration together with the identity $\frac{d^k}{dx^k}H_k(x) = k!$ yield that

$$\int_{-\infty}^{\infty} H_k(x)H_l(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx = k!\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}(-1)^{l-k}\frac{d^{l-k}}{dx^{l-k}}e^{-x^2/2}\,dx.$$

The last relation supplies formulas (C2) and (C2′).

To prove relation (C3) observe that $H_k(x) - xH_{k-1}(x)$ is a polynomial of order $k-2$. (The term x^{k-1} is missing from this expression. Indeed, if k is an even number, then the polynomial $H_k(x) - xH_{k-1}(x)$ is an even function, and it does not contain the term x^{k-1} with an odd exponent $k-1$. Similar argument holds if the number k is odd.) Beside this, it is orthogonal (with respect to the standard normal distribution) to all Hermite polynomials $H_l(x)$ with $0 \le l \le k-3$. Hence $H_k(x) - xH_{k-1}(x) = CH_{k-2}(x)$ with some constant C to be determined.

Multiply both sides of the last identity with $H_{k-2}(x)$ and integrate them with respect to the standard normal distribution. Apply the orthogonality of the polynomials $H_k(x)$ and $H_{k-2}(x)$, and observe that the identity

$$\int H_{k-1}(x)xH_{k-2}(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx = \int H_{k-1}^2(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx = (k-1)!$$

holds. (In this calculation we have exploited that $H_{k-1}(x)$ is orthogonal to $H_{k-1}(x) - xH_{k-2}(x)$, because the order of the latter polynomial is less than $k-1$.) In such a way we get the identity $-(k-1)! = C(k-2)!$ for the constant C in the last identity, and this implies relation (C3).

Proof of Itô's formula for multiple Wiener–Itô integrals. Let $K = \sum_{p=1}^{m} k_p$, the sum of the order of the Hermite polynomials, denote the order of the expression in relation (10.20).

Formula (10.20) clearly holds for expressions of order $K = 1$. It will be proved in the general case by means of induction with respect to the order K.

In the proof the functions $f(x_1) = \varphi_1(x_1)$ and

$$g(x_1, \dots, x_{K_m-1}) = \prod_{j=1}^{K_1-1} \varphi_1(x_j) \cdot \prod_{p=2}^{m} \prod_{j=K_{p-1}}^{K_p-1} \varphi_p(x_j),$$

will be introduced and the product $Z_{\mu,1}(f)(K_m - 1)!Z_{\mu,K_m-1}(g)$ will be calculated by means of the diagram formula. (The same notation is applied as in Theorem 10.3. In particular, $K = K_m$, and in the case $K_1 = 1$ the convention $\prod_{j=1}^{K_1-1} \varphi_1(x_j) = 1$ is applied.) In the application of the diagram formula diagrams with two rows appear. The first row of these diagrams contains the vertex $(1, 1)$ and the second row contains the vertices $(2, 1), \dots, (2, K_m - 1)$. It is useful to divide the diagrams to three disjoint classes. The first class contains only the diagram γ_0 without any edges. The second class Γ_1 consists of those diagrams which have an edge of the form $((1, 1), (2, j))$ with some $1 \le j \le k_1 - 1$, and the third class Γ_2 is the set of those diagrams which have an edge of the form $((1, 1), (2, j))$ with some $k_1 \le j \le K_m - 1$. Because of the orthogonality of the functions φ_s for different indices s $F_\gamma \equiv 0$ and $Z_{\mu,K_m-2}(F_\gamma) = 0$ for $\gamma \in \Gamma_2$. The class Γ_1 contains $k_1 - 1$ diagrams. Let us consider a diagram γ from this class with an edge $((1, 1), (2, j_0))$, $1 \le j \le k_1 - 1$. We have for such a diagram $F_\gamma = \prod_{j \in \{1,\dots,K_1-1\} \setminus \{j_0\}} \varphi_1(x_{(2,j)}) \prod_{p=2}^{m} \prod_{j=K_{p-1}}^{K_p-1} \varphi_p(x_{(2,j)})$, and by our inductive hypothesis $(K_m - 2)!Z_{\mu,K_m-2}(F_\gamma) = H_{k_1-2}(\eta_1) \prod_{p=2}^{m} H_{k_p}(\eta_p)$. Finally

$$K_m!Z_{\mu,K_m}(F_{\gamma_0}) = K_m!Z_{\mu,K_m}\left(\prod_{p=1}^{m}\left(\prod_{j=K_{p-1}+1}^{K_p} \varphi_p(x_j)\right)\right)$$

for the diagram γ_0 without any edge.

Our inductive hypothesis also imply that

$$Z_{\mu,1}(f)(K_m - 1)!Z_{\mu,K_m-1}(g) = \eta_1 H_{k_1-1}(\eta_1) \prod_{p=2}^{m} H_{k_p}(\eta_p).$$

Let us express $K_m!Z_{\mu,K_m}(F_{\gamma_0})$ by applying the diagram formula in the above example, and express each term appearing in this identity by means of the above written relations. This calculation together with the observation $|\Gamma_1| = k_1 - 1$ yield the identity

$$K_m!Z_{\mu,K_m}\left(\prod_{p=1}^{m}\left(\prod_{j=K_{p-1}+1}^{K_p} \varphi_p(x_j)\right)\right) = K_m!Z_{\mu,K_m}(F_{\gamma_0})$$

$$= Z_{\mu,1}(f)(K_m - 1)! Z_{\mu,K_m-1}(g) - \sum_{\gamma \in \Gamma_1} (K_m - 2)! Z_{\mu,K_m-2}(F_\gamma)$$

$$= \eta_1 H_{k_1-1}(\eta_1) \prod_{p=2}^{m} H_{k_p}(\eta_p) - (k_1 - 1) H_{k_1-2}(\eta_1) \prod_{p=2}^{m} H_{k_p}(\eta_p)$$

$$= [\eta_1 H_{k_1-1}(\eta_1) - (k_1 - 1) H_{k_1-2}(\eta_1)] \prod_{p=2}^{m} H_{k_p}(\eta_p). \tag{C4}$$

On the other hand, $\eta_1 H_{k_1-1}(\eta_1) - (k_1 - 1) H_{k_1-2}(\eta_1) = H_{k_1}(\eta_1)$ by formula (C3). These relations imply formula (10.20), i.e. Itô's formula.

I present the proof of another important property of the Hermite polynomials in the following Proposition C2.

Proposition C2 on the completeness of the orthogonal system of Hermite polynomials. *The Hermite polynomials $H_k(x)$, $k = 0, 1, 2, \dots$, defined in formula (C4) constitute a complete orthonormal system of in the L_2-space of the functions square integrable with respect to the Gaussian measure $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$ on the real line.*

Proof of Proposition C2. Let us consider the orthogonal complement of the subspace generated by the Hermite polynomials in the space of the square integrable functions with respect to the measure $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$. It is enough to prove that this orthogonal completition contains only the identically zero function. Since the orthogonality of a function to all polynomials of the form x^k, $k = 0, 1, 2, \dots$ is equivalent to the orthogonality of this function to all Hermite polynomials $H_k(x)$, $k = 0, 1, 2, \dots$, Proposition C2 can be reformulated in the following form:

If a function $g(x)$ on the real line is such that

$$\int_{-\infty}^{\infty} x^k g(x) \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 0 \quad \text{for all } k = 0, 1, 2, \dots \tag{C5}$$

and

$$\int_{-\infty}^{\infty} g^2(x) \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx < \infty, \tag{C6}$$

then $g(x) = 0$ for almost all x.

Given a function $g(x)$ satisfying (C6) define the (finite) measure ν_g,

$$\nu_g(A) = \int_A g(x) \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

on the measurable sets of the real line (because of relation (C6) the function g is also L_1-integrable with respect to the Gaussion measure, hence ν_g is a finite measure) together with its Fourier transform $\tilde{\nu}_g(t) = \int_{-\infty}^{\infty} e^{itx} \nu_g(\, dx)$. First I show that Proposition C2 can be reduced to the following statement: If a function g satisfies both (C5) and (C6) then $\tilde{\nu}_g(t) = 0$ for all $-\infty < t < \infty$.

Indeed, if there were a function g satisfying (C5) and (C6) which is not identically zero, then the non-negative functions $g^+(x) = \max(0, g(x))$ and $g^-(x) = -\min(0, g(x))$ would be different. Then also their Fourier transform $\tilde{\nu}_{g^+}(t)$ and $\tilde{\nu}_{g^-}(t)$ would be different, since a finite measure is uniquely determined by its Fourier transform. (This statement is equivalent to an important result in probability theory, by which a probability measure on the real line is determined by its characteristic function.) But this would mean that $\tilde{\nu}_g(t) = \tilde{\nu}_{g^+}(t) - \tilde{\nu}_{g^-}(t) \neq 0$ for some t. Hence Proposition C2 can be reduced to the above statement.

Since $\left| e^{itx} - 1 - (itx) - \cdots - \frac{(itx)^k}{k!} \right| \le \frac{|tx|^{(k+1)}}{(k+1)!}$ for all real numbers t, x and integer $k = 1, 2, \dots$ we may write because of relation (C5)

$$\begin{aligned} |\tilde{\nu}_g(t)| &= \left| \int_{-\infty}^{\infty} \left(e^{itx} - 1 - (itx) - \cdots - \frac{(itx)^k}{k!} \right) g(x) \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \right| \\ &\le \int_{-\infty}^{\infty} \frac{|t|^{(k+1)}}{(k+1)!} |x|^{k+1} |g(x)| \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \end{aligned}$$

for all $k = 1, 2, \dots$ and real number t if the function g satisfies relation (C5). If it satisfies both relation (C5) and (C6), then from the last relation and the Schwarz inequality

$$\begin{aligned} |\tilde{\nu}_g(t)|^2 &\le \text{const.} \frac{|t|^{2(k+1)}}{((k+1)!)^2} \int_{-\infty}^{\infty} |x|^{2(k+1)} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\ &= \text{const.} \frac{|t|^{2(k+1)}}{((k+1)!)^2} 1 \cdot 3 \cdot 5 \cdots (2k+1) \end{aligned}$$

for all real number t and integer $k = 1, 2, \dots$. Simple calculation shows that the right-hand side of the last estimate tends to zero as $k \to \infty$. This implies that $\tilde{\nu}_g(t) = 0$ for all t, and Proposition C2 holds.

I finish Appendix C with the proof of Theorem 10.4, a limit theorem about a sequence of normalized degenerate U-statistics. It is based on an appropriate representation of the U-statistics by means of multiple random integrals which makes possible to carry out an appropriate limiting procedure.

Proof of Theorem 10.4. For all $n = 1, 2, \dots$, the normalized degenerate U-statistics $n^{-k/2} I_{n,k}(f)$ can be written in the form

$$\begin{aligned} n^{-k/2} k! I_{n,k}(f) &= n^{k/2} \int' f(x_1, \dots, x_k) \mu_n(\, dx_1) \dots \mu_n(\, dx_k) \\ &= n^{k/2} \int' f(x_1, \dots, x_k) (\mu_n(\, dx_1) - \mu(\, dx_1)) \dots (\mu_n(\, dx_k) - \mu(\, dx_1)), \end{aligned} \tag{C7}$$

where μ_n is the empirical distribution function of the sequence $\xi_1, \dots, \xi_n$ defined in (4.5), and the prime in $\int'$ denotes that the diagonals, i.e. the points $x = (x_1, \dots, x_k)$ such that $x_j = x_{j'}$ for some pairs of indices $1 \le j, j' \le k$, $j \neq j'$, are omitted from the

domain of integration. The second identity in relation (C7) can be justified by means of the identity

$$\int' f(x_1,\dots,x_k)(\mu_n(\,dx_1)-\mu(\,dx_1))\dots(\mu_n(\,dx_k)-\mu(\,dx_1))-I_{n,k}(f)$$
$$=\sum_{V\colon V\in\{1,\dots,k\},\,|V|\ge 1}(-1)^{|V|}\int' f(x_1,\dots,x_k)\prod_{j\in V}\mu(\,dx_j)\prod_{j\in\{1,\dots,k\}\setminus V}\mu_n(\,dx_j))=0. \tag{C8}$$

This identity holds for a function f canonical with respect to a non-atomic measure μ, because each term in the sum at the right-hand side of (C8) equals zero. Indeed, the integral of a canonical function f with respect to $\mu(\,dx_j)$ with some index $j\in V$ equals zero for all fixed values $x_1,\dots,x_{j-1},x_{j+1},\dots,x_k$. The non-atomic property of the measure μ was needed to guarantee that this integral equals zero also in the case when the diagonals are omitted from the domain of integration.

We would like to derive Theorem 10.4 from relation (C7) by means of an appropriate limiting procedure which exploits the convergence of the random fields $n^{1/2}(\mu_n(A)-\mu(A))$, $A\in\mathcal{X}$, to a Gaussian field $\nu(A)$, $A\in\mathcal{X}$, as $n\to\infty$. But some problems arise if we want to carry out such a program, because the fields $n^{1/2}(\mu_n-\mu)$ converge to a non white noise type Gaussian field. The limit we get is similar to a Wiener bridge on the real line. Hence a relation between Wiener processes and Wiener bridges suggests to write the following version of formula (C7).

Let us take a standard Gaussian random variable η, independent of the random sequence $\xi_1,\xi_2,\dots$. For a canonical function f the following version of (C7) holds.

$$n^{-k/2}k!I_{n,k}(f)=J'_{n,k}(f) \tag{C9}$$

with

$$J'_{n,k}(f)=\int' f(x_1,\dots,x_k)\left[\sqrt{n}(\mu_n(\,dx_1)-\mu(\,dx_1))+\eta\mu(\,dx_1)\right]\dots\left[\sqrt{n}(\mu_n(\,dx_k)-\mu(\,dx_k))+\eta\mu(\,dx_k)\right]. \tag{C10}$$

This relation can be seen similarly to (C7).

The random measures $n^{1/2}(\mu_n-\mu)+\eta\mu$ converge to a white noise with reference measure μ. Hence Theorem 10.4 can be proved by means of formulas (C9) and (C10) with the help of an appropriate limiting procedure. More explicitly, I claim that the following slightly more general result holds. The expressions $J'_{n,k}(f)$ introduced in (C10) converge in distribution to the Wiener–Itô integral $k!Z_{\mu,k}(f)$ as $n\to\infty$ for all functions f square integrable with respect to the product measure μ^k. This result also holds for non-canonical functions f. This limit theorem together with relation (C9) imply Theorem 10.4.

The convergence of the random variables $J'_{n,k}(f)$ defined in (C10) to the Wiener–Itô integral $k!Z_{\mu,k}(f)$ can be easily checked for elementary functions $f\in\bar{\mathcal{H}}_{\mu,k}$. Indeed, if $A_1,\dots,A_M$ are disjoint sets with $\mu(A_s)<\infty$, then the multi-dimensional central limit theorem implies that the random vectors $\{\sqrt{n}((\mu_n(A_s)-\mu(A_s))+\eta\mu(A_s),\,1\le s\le M\}$

converge in distribution to the random vector $\{(\mu_W(A_s),\, 1 \le s \le M\}$, i.e. to a set of independent normal random variables ζ_s, $E\zeta_s = 0$, $1 \le s \le M$, with variance $E\zeta_s^2 = \mu(A_s)$ as $n \to \infty$. The definition of the elementary functions given in (10.2) shows that this central limit theorem implies the demanded convergence of the sequence $J'_{n,k}(f)$ to $k!Z_{\mu,k}(f)$ for elementary functions.

To show the convergence of the sequence $J'_{n,k}(f)$ to $k!Z_{\mu,k}(f)$ in the general case, take for any function $f \in \mathcal{H}_{\mu,k}$ a sequence of elementary functions $f_N \in \bar{\mathcal{H}}_{\mu,k}$ such that $\|f - f_N\|_2 \to 0$ as $N \to \infty$. Then $E(Z_{\mu,k}(f) - Z_{\mu,k}(f_N))^2 = E(Z_{\mu,k}(f - f_N))^2 \to 0$ as $N \to \infty$ by Property c) in Theorem 10.1. Hence the already proved part of the theorem implies that there exists some sequence of positive integers, $N(n)$, $n = 1, 2, \dots$, in such a way that $N(n) \to \infty$, and the sequence $J'_{n,k}(f_{N(n)})$ converges to $k!Z_{\mu,k}(f)$ in distribution as $n \to \infty$. Thus to complete the proof of Theorem 10.4 it is enough to show that $E(J'_{n,k}(f_{N(n)}) - J'_{n,k}(f))^2 = E(J'_{n,k}(f_{N(n)} - f))^2 \to 0$ as $n \to \infty$.

It is enough to show that

$$E(J'_{n,k}(f))^2 \le C\|f\|_2^2 \quad \text{for all } f \in \mathcal{H}_{\mu,k} \tag{C11}$$

with a constant $C = C_k$ depending only on the order k of the function f and to apply inequality (C11) for the functions $f_{N(n)} - f$. Relation (C11) is a relatively simple consequence of Corollary 1 of Theorem 9.4.

Indeed,

$$J'_{n,k}(f) = \sum_{V \subset \{1,\dots,k\}} \eta^{k-|V|}|V|!J_{n,|V|}(f_V)$$

with

$$f_V(x_j,\, j \in V) = \int f(x_1, \dots, x_k) \prod_{j' \in \{1,\dots,k\} \setminus V} \mu(dx_{j'})$$

and the random integral $J_{n,k}(\cdot)$ defined in (4.8), hence

$$E(J'_{n,k}(f))^2 \le 2^k \sum_{V \subset \{1,\dots,k\}} (|V|!)^2 E\eta^{2(k-|V|)} \cdot EJ^2_{n,|V|}(f_V). \tag{C12}$$

Inequality $\|f_V\|_2 \le \|f\|_2$ holds for all sets $V \subset \{1, \dots, k\}$, hence an application of Corollary 1 of Theorem (9.4) to all random integrals $J_{n,|V|}(f)$ supplies (C12).

The above proof also yields the following slight generalization of Theorem 10.4. Let us consider a finite sequence of functions $f_j \in \mathcal{H}_{\mu,j}$, $1 \le j \le k$, canonical with respect to a non-atomic probability measure μ. The vectors $\{n^{-j/2}I_{n,j}(f_j), 1 \le j \le k\}$, consisting of normalized degenerate U-statistics defined with the help of a sequence of independent μ-distributed random variables converge to the random vector $\{Z_{\mu,j}(f_j), 1 \le j \le k\}$ in distribution as $n \to \infty$. This result together with Theorem 9.4 imply the following limit theorem about multiple random integrals $J_{n,k}(f)$.

Theorem 10.4′. Limit theorem about multiple random integrals with respect to normalized empirical measures. *Let a sequence of independent and identically*

distributed random variables $\xi_1, \xi_2, \dots$ be given with some non-atomic distribution μ on a measurable space $(X, \mathcal{X})$ together with a function $f(x_1, \dots, x_k)$ on the k-fold product $(X^k, \mathcal{X}^k)$ of the space $(X, \mathcal{X})$ such that

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

Let us consider for all $n = 1, 2, \dots$ the random integrals $J_{n,k}(f)$ of order k defined in formulas (4.5) and (4.8) with the help of the empirical distribution μ_n of the sequence $\xi_1, \dots, \xi_n$ and the function f. These random integrals $J_{n,k}(f)$ converge in distribution, as $n \to \infty$, to the following sum $U(f)$ of multiple Wiener–Itô integrals:

$$\begin{aligned} U(f) &= \sum_{V \subset \{1,\dots,k\}} C(k, V) Z_{\mu,|V|}(f_V) \\ &= \sum_{V \subset \{1,\dots,k\}} \frac{C(k, V)}{|V|!} \int f_V(x_j,\, j \in V) \prod_{j \in V} \mu_W(dx_j), \end{aligned}$$

where the functions $f_V(x_j\, j \in V)$, $V \subset \{1, \dots, k\}$, are those functions defined in formula (9.2) which appear in the Hoeffding decomposition of the function $f(x_1, \dots, x_k)$, the constants $C(k, V)$ are the limits appearing in the limit relation $\lim_{n\to\infty} C(n, k, V) = C(k, V)$ satisfied by the coefficients $C(n, k, V)$ in formula (9.9), and μ_W is a white noise with reference measure μ.

An essential step of the proof of Theorem 10.4 was the reduction of the case of general kernel functions to the case of elementary kernel functions. Let me make some comments about it.

It would be simple to make such a reduction if we had a good approximation of a canonical function with such elementary functions which are alsso canonical. But it is very hard to find such an approximation. To overcome this difficulty we reduced the proof of Theoerem 10.4 to a modified version of this result, where instead of a limit theorem for degenerate U-statistics a limit theorem for the random variables $J'_{n,k}(f)$ introduced in formula (C10) has to be proved. In the proof of such a version we could apply the approximation of a general kernel function with not necessarily canonical elementary functions. Theorem 9.4 helped us to work with such an approximation. Another natural way to overcome the above difficulty is to apply a Poissonian approximation of the normalized empirical measure. Such an approach was applied in [14] and in [30], where some generalizations of Theorem 10.4 were proved.

Appendix D. The proof of Theorem 14.3.

A result about the comparison of U-statistics and decoupled U-statistics.

The proof of Theorem 14.3. It will be simpler to formulate and prove a generalized version of Theorem 14.3 where such generalized U-statistics are considered in which different kernel functions may appear in each term of the sum. More explicitly, let $\ell = \ell(n,k)$ denote the set of all such sequences $l = (l_1, \dots, l_k)$ of integers of length k for which $1 \le l_j \le n$, $1 \le j \le k$. To define generalized U-statistics let us fix a set 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, and have the property $f_{l_1,\dots,l_k}(x_1,\dots,x_k) \equiv 0$ if $l_j = l_{j'}$ for some indices $j \neq j'$. (The last condition corresponds to that property of U-statistics that the diagonals are omitted from the summation in their definition.) Let us denote this set 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

$$I_{n,k}(f(\ell)) = \frac{1}{k!} \sum_{(l_1,\dots,l_k)\colon\, 1 \le l_j \le n,\ j=1,\dots,k} f_{l_1,\dots,l_k}\left(\xi_{l_1},\dots,\xi_{l_k}\right) \tag{D1}$$

and

$$\bar{I}_{n,k}(f(\ell)) = \frac{1}{k!} \sum_{(l_1,\dots,l_k)\colon\, 1 \le l_j \le n,\ j=1,\dots,k} f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(1)},\dots,\xi_{l_k}^{(k)}\right) \tag{D2}$$

(with the same independent random variables ξ_l and $\xi_l^{(j)}$, $1 \le l \le n$, $1 \le j \le k$, as in the definition of the original U-statistics and decoupled U-statistics.)

The following generalization of relation (14.12) will be proved.

$$P\left(\|I_{n,k}(f(\ell))\| > u\right) \le A(k) P\left(\|\bar{I}_{n,k}(f(\ell))\| > \gamma(k)u\right) \tag{14.12d}$$

with some constants $A(k) > 0$ and $\gamma(k) > 0$ depending only on the order k of these generalized U-statistics.

We concentrate mainly on the proof of the generalization (14.12d) of relation (14.12). Formula (14.13) is a relatively simple consequence of it. Formula (14.12d) will be proved by means of an inductive procedure which works only in this more general setting. It will be derived from 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 their help the following version of decoupled U-statistics

$$I_{n,k,V}(f(\ell)) = \frac{1}{k!} \sum_{(l_1,\dots,l_k)\colon\, 1 \le l_j \le n,\ j=1,\dots,k} f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(\alpha_V(1))},\dots,\xi_{l_k}^{(\alpha_V(k))}\right)$$
$$\text{for all } V \subset \{1,\dots,k\}. \tag{D3}$$

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 $u > 0$

$$P\left(\|I_{n,k}(f(\ell))\| > u\right) \le \sum_{V \subset \{1,\dots,k\},\, 1 \le |V| \le k-1} C_k P\left(D_k \|I_{n,k,V}(f(\ell))\| > u\right). \tag{D4}$$

Here $|V|$ denotes the cardinality of the set V, and the condition $1 \le |V| \le k-1$ in the summation of formula (D4) means that the sets $V = \emptyset$ and $V = \{1, \dots, k\}$ are omitted from the summation, i.e. the terms where either $\alpha_V(j) = 1$ or $\alpha_V(j) = 2$ for all $1 \le j \le k$ are not considered. Formula (14.12d) can be derived from formula (D4) by means of an inductive argument. The hard part of the problem is to prove formula (D4). To do this first the following simple lemma will be proved.

Lemma D1. *Let ξ and η be two independent and identically distributed random variables taking values in 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 D1. Let ξ, η and ζ be 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 (D4) we introduce the random variable

$$T_{n,k}(f(\ell)) = \frac{1}{k!} \sum_{\substack{(l_1,\dots,l_k),\ (s_1,\dots,s_k):\\ 1 \le l_j \le n,\ s_j = 1 \text{ or } s_j = 2,\ j = 1,\dots,k,}} f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(s_1)}, \dots, \xi_{l_k}^{(s_k)}\right) = \sum_{V \subset \{1,\dots,k\}} I_{n,k,V}(f(\ell)). \tag{D5}$$

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 D1 yields that

$$\begin{aligned} P(\|I_{n,k}(f(\ell))\| > u) &\le 3P\left(\|I_{n,k,\emptyset}(f(\ell)) + I_{n,k,\{1,\dots,k\}}(f(\ell))\| > \frac{2}{3}u\right) \\ &= 3P\left(\left\|T_{n,k}(f(\ell)) - \sum_{V:\, V \subset \{1,\dots,k\},\, 1 \le |V| \le k-1} I_{n,k,|V|}(f(\ell))\right\| > \frac{2}{3}u\right) \end{aligned}$$

$$\le 3P(3\cdot 2^{k-1}\|T_{n,k}(f(\ell))\| > u) \qquad \text{(D6)}$$
$$+ \sum_{V\colon V\subset\{1,\dots,k\},\, 1\le|V|\le k-1} 3P(3\cdot 2^{k-1}\|I_{n,k,|V|}(f(\ell))\| > u).$$

To derive relation (D4) from relation (D6) a good estimate is needed on the probability $P(3\cdot 2^{k-1}\|T_{n,k}(f(\ell))\| > u)$. To get such an estimate the tail distribution of $\|T_{n,k}(f(\ell))\|$ will be compared with that of $\|I_{n,k,V}(f(\ell))\|$ for an arbitrary set $V \subset \{1,\dots,k\}$. This will be done with the help of Lemmas D2 and D4 formulated below.

In Lemma D2 such a random variable $|\bar I_{n,k,V}(f(\ell))\|$ will be constructed whose distribution agrees with that of $\|I_{n,k,V}(f(\ell))\|$. This expression $I_{n,k,v}f(\ell)$ will be defined in formulas (D7) and (D8). It is a random polynomial of some Rademacher functions $\varepsilon_1,\dots,\varepsilon_n$. The coefficients of this polynomial are random variables, independent of the Rademacher functions $\varepsilon_1,\dots,\varepsilon_n$. Beside this, the constant term of this polynomial equals $T_{n,k}(f(\ell)$. These properties of the polynomial $\|\bar I_{n,k,V}(f(\ell))\|$ together with Lemma D4 formulated below enable us prove such an estimate on the distribution of $\|T_{n,k}(f(\ell))\|$ that together with formula (D6) imply relation (D4). Let us formulate these lemmas.

Lemma D2. *Let us consider a sequence of independent random variables $\varepsilon_1,\dots,\varepsilon_n$, $P(\varepsilon_l = 1) = P(\varepsilon_l = -1) = \frac{1}{2}$, $1 \le l \le n$, 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 modified decoupled U-statistics $I_{n,k,V}(f(\ell))$ given in formula (D3). 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 defined by the formula*

$$(\eta_l^{(1)}(\varepsilon_l),\eta_l^{(2)}(\varepsilon_l)) = \left(\frac{1+\varepsilon_l}{2}\xi_l^{(1)} + \frac{1-\varepsilon_l}{2}\xi_l^{(2)}, \frac{1-\varepsilon_l}{2}\xi_l^{(1)} + \frac{1+\varepsilon_l}{2}\xi_l^{(2)}\right),$$

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 $\eta_1^{(2)},\dots,\eta_n^{(2)}$, which is also independent of the sequence $\varepsilon_1,\dots,\varepsilon_n$.

Let us fix some $V \subset \{1,\dots,k\}$, and introduce the random variable

$$\bar I_{n,k,V}(f(\ell)) = \frac{1}{k!}\sum_{(l_1,\dots,l_k)\colon\, 1\le l_j\le n,\ j=1,\dots,k} f_{l_1,\dots,l_k}\left(\eta_{l_1}^{(\alpha_V(1))},\dots,\eta_{l_k}^{(\alpha_V(k))}\right), \qquad \text{(D7)}$$

where similarly to formula (D3) $\alpha_V(j) = 1$ if $j \in V$, and $\alpha_V(j) = 2$ if $j \notin V$. Then the identity

$$2^k\bar I_{n,k,V}(f(\ell)) \qquad \text{(D8)}$$
$$= \frac{1}{k!}\sum_{\substack{(l_1,\dots,l_k),\ (s_1,\dots,s_k)\colon\\ 1\le l_j\le n,\ s_j=1 \text{ or } s_j=2,\ j=1,\dots,k,}} (1+\kappa_{s_1,V}^{(1)}\varepsilon_{l_1})\cdots(1+\kappa_{s_k,V}^{(k)}\varepsilon_{l_k}) f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(s_1)},\dots,\xi_{l_k}^{(s_k)}\right)$$

holds, where $\kappa_{1,V}^{(j)} = 1$ *and* $\kappa_{2,V}^{(j)} = -1$ *if* $j \in V$, *and* $\kappa_{1,V}^{(j)} = -1$ *and* $\kappa_{2,V}^{(j)} = 1$ *if* $j \notin V$, *i.e.* $\kappa_{1,V}^{(j)} = 3 - 2\alpha_V(j)$ *and* $\kappa_{2,V}^{(j)} = -\kappa_{1,V}^{(j)}$.

Before the formulation of Lemma D4 another Lemma D3 will be presented which will be applied in its proof.

Lemma D3. *Let* Z *be a random variable taking values in a separable Banach space* B *with expectation zero, i.e. let* $E\kappa(Z) = 0$ *for all* $\kappa \in B'$, *where* B' *denotes the (Banach) space of all (bounded) linear transformations of* B *to the real line. Then* $P(\|v + Z\| \ge \|v\|) \ge \inf_{\kappa \in B'} \frac{(E|\kappa(Z)|)^2}{4E\kappa(Z)^2}$ *for all* $v \in B$.

Lemma D4. *Let us consider a positive integer* n *and a sequence of independent random variables* $\varepsilon_1, \dots, \varepsilon_n$, $P(\varepsilon_l = 1) = P(\varepsilon_l = -1) = \frac{1}{2}$, $1 \le l \le n$. *Beside this, fix some positive integer* k, *take a separable Banach space* B *and choose some elements* $a(l_1, \dots, l_s)$ *of this Banach space* B, $1 \le s \le k$, $1 \le l_j \le n$, $l_j \ne l_{j'}$ *if* $j \ne j'$, $1 \le j, j' \le s$. *With the above notations the inequality*

$$P\left(\left\| v + \sum_{s=1}^{k} \sum_{\substack{(l_1,\dots,l_s):\ 1 \le l_j \le n,\ j=1,\dots,s,\\ l_j \ne l_{j'} \text{ if } j \ne j'}} a(l_1, \dots, l_s)\varepsilon_{l_1} \cdots \varepsilon_{l_s} \right\| \ge \|v\| \right) \ge c_k \tag{D9}$$

holds for all $v \in B$ *with some constant* $c_k > 0$ *which depends only on the parameter* k.

Proof of Lemma D2. Let us consider the conditional joint distribution of the sequences of random variables $\eta_1^{(1)}, \dots, \eta_n^{(1)}$ and $\eta_1^{(2)}, \dots, \eta_n^{(2)}$ 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 $\xi_1^{(1)}, \dots, \xi_n^{(1)}$ and $\xi_1^{(2)}, \dots, \xi_n^{(2)}$ for all possible conditions. This fact implies the statement about the joint distribution of the sequences $\eta_l^{(1)}, \eta_l^{(2)}$, $1 \le l \le n$ and their independence of the sequence $\varepsilon_1, \dots, \varepsilon_n$.

To prove identity (D8) 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 $l \notin M$. Put $\beta_{V,M}(j, l) = 1$ if $j \in V$ and $l \in M$ or $j \notin V$ and $l \notin M$, and let $\beta_{V,M}(j, l) = 2$ otherwise. Then we have for all $(l_1, \dots, l_k)$, $1 \le l_j \le n$, $1 \le j \le n$, and our fixed set V

$$\begin{aligned} \sum_{(s_1,\dots,s_k):\ s_j=1 \text{ or } s_j=2,\ j=1,\dots,k,} & (1 + \kappa_{s_1,V}^{(1)}\varepsilon_{l_1}) \cdots (1 + \kappa_{s_k,V}^{(k)}\varepsilon_{l_k}) f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(s_1)}, \dots, \xi_{l_k}^{(s_k)}\right) \\ &= 2^k f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(\beta_{V,M}(1,l_1))}, \dots, \xi_{l_k}^{(\beta_{V,M}(k,l_k))}\right), \end{aligned} \tag{D10}$$

since the product $(1 + \kappa_{s_1,V}^{(1)}\varepsilon_{l_1}) \cdots (1 + \kappa_{s_k,V}^{(k)}\varepsilon_{l_k})$ equals either zero or 2^k, and it equals 2^k for that sequence $(s_1, \dots, s_k)$ for which $\kappa_{s_j,V}^{(j)}\varepsilon_{l_j} = 1$ for all $1 \le j \le k$. This relation is equivalent to $\beta_{V,M}(j, l_j) = s_j$ for all $1 \le j \le k$. (In relation (D10) it is sufficient to

consider only such products for which $l_j \neq l_{j'}$ if $j \neq j'$ because of the properties of the functions $f_{l_1,\dots,l_k}$.)

Beside this, $\xi_l^{\beta_{V,M}(l,j)} = \eta_l^{\alpha_V(j)}$ for all $1 \le l \le n$ and $1 \le j \le k$, and as a consequence

$$f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(\beta_{V,M}(1,l_1))},\dots,\xi_{l_k}^{(\beta_{V,M}(k,l_k))}\right) = f_{l_1,\dots,l_k}\left(\eta_{l_1}^{(\alpha_V(1))},\dots,\eta_{l_k}^{(\alpha_V(k))}\right).$$

Summing up the identities (D10) for all $1 \le l_1,\dots,l_k \le n$ and applying the last identity we get relation (D8), since the identity obtained in such a way holds for all $M \subset \{1,\dots,n\}$.

Proof of Lemma D3. Let us first observe that if ξ is a real valued random variable with zero expectation, then $P(\xi \ge 0) \ge \frac{(E|\xi|)^2}{4E\xi^2}$ since $(E|\xi|)^2 = 4(E(\xi I(\{\xi \ge 0\}))^2 \le 4P(\xi \ge 0)E\xi^2$ by the Schwarz inequality, where $I(A)$ denotes the indicator function of the set A. (In the above calculation and in the subsequent proofs I apply the convention $\frac{0}{0} = 1$. We need this convention if $E\xi^2 = 0$. In this case we have, because of the condition $E\xi = 0$ $P(\xi = 0) = 1$, hence the above proved identity holds in this case, too.)

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 $\{\omega\colon \|v + Z(\omega)\| \ge \|v\|\} \supset \{\omega\colon \kappa(v + Z(\omega)) \ge \kappa(v)\} = \{\omega\colon \kappa(Z(\omega)) \ge 0\}$. Beside this $E\kappa(Z) = 0$. Hence we can apply the above proved inequality for $\xi = \kappa(Z)$, and it yields that $P(\|v + Z\| \ge \|v\|) \ge P(\kappa(Z) \ge 0) \ge \frac{(E|\kappa(Z)|)^2}{4E\kappa(Z)^2}$. Lemma D3 is proved.

Proof of Lemma D4. Take the class of random polynomials

$$Y = \sum_{s=1}^{k} \sum_{\substack{(l_1,\dots,l_s)\colon 1 \le l_j \le n,\ j=1,\dots,s,\\ l_j \neq l_{j'} \text{ if } j \neq j'}} b(l_1,\dots,l_s)\varepsilon_{l_1}\cdots\varepsilon_{l_s},$$

where ε_l, $1 \le l \le n$, are independent random variables with $P(\varepsilon_l = 1) = P(\varepsilon_l = -1) = \frac{1}{2}$, and the coefficients $b(l_1,\dots,l_s)$, $1 \le s \le k$, are arbitrary real numbers. The proof of Lemma D4 can be reduced to the statement that there exists a constant $c_k > 0$ depending only on the order k of these polynomials such that the inequality

$$(E|Y|)^2 \ge 4c_k EY^2. \tag{D11}$$

holds for all such polynomials Y. Indeed, consider the polynomial

$$Z = \sum_{s=1}^{k} \sum_{\substack{(l_1,\dots,l_s)\colon 1 \le l_j \le n,\ j=1,\dots,s,\\ l_j \neq l_{j'} \text{ if } j \neq j'}} a(l_1,\dots,l_s)\varepsilon_{l_1}\cdots\varepsilon_{l_s},$$

and observe that $E\kappa(Z) = 0$ for all linear functionals κ on the space B. Hence Lemma D3 implies that the left-hand side expression in (D9) is bounded from below by $\inf\limits_{\kappa \in B'} \frac{(E|\kappa(Z)|)^2}{4E\kappa(Z)^2}$. On the other hand, relation (D11) implies that $\inf\limits_{\kappa \in G'} \frac{(E|\kappa(Z)|)^2}{4E\kappa(Z)^2} \ge c_k$.

To prove relation (D11) first we compare the moments EY^2 and EY^4. Let us introduce the random variables

$$Y_s = \sum_{\substack{(l_1,\dots,l_s)\colon 1\le l_j\le n,\ j=1,\dots,s,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} b(l_1,\dots,l_s)\varepsilon_{l_1}\cdots\varepsilon_{l_s} \quad 1\le s\le k.$$

We shall show that the estimates of Section 13 imply that

$$EY_s^4 \le 2^{4s}\left(EY_s^2\right)^2 \tag{D12}$$

for these random variables Y_s.

Relation (D12) together with the uncorrelatedness of the random variables Y_s, $1\le s\le k$, imply that

$$EY^4 = E\left(\sum_{s=1}^k Y_s\right)^4 \le k^3\sum_{s=1}^k EY_s^4 \le k^3 2^{4k}\sum_{s=1}^k (EY_s^2)^2$$
$$\le k^3 2^{4k}\left(\sum_{s=1}^k EY_s^2\right)^2 = k^3 2^{4k}(EY^2)^2.$$

This estimate together with the Hölder inequality with $p=3$ and $q=\frac{3}{2}$ yield that $EY^2 = E|Y|^{4/3}|\cdot|Y|^{2/3} \le (EY^4)^{1/3}(E|Y|)^{2/3} \le k2^{4k/3}(EY^2)^{2/3}(E|Y|)^{2/3}$, i.e. $EY^2 \le k^3 2^{4k}(E|Y|)^2$, and relation (D11) holds with $4c_k = k^{-3}2^{-4k}$. Hence to complete the proof of Lemma D4 it is enough to check relation (D12).

In the proof of relation (D12) it can be assumed that the coefficients $b(l_1,\dots,l_s)$ of the random variable Y_s are symmetric functions of the arguments $l_1,\dots,l_s$, since a symmetrization of these coefficients does not change the value of Y. Put

$$B_s^2 = \sum_{\substack{(l_1,\dots,l_s)\colon 1\le l_j\le n,\ j=1,\dots,s,\\ l_j\ne l_{j'} \text{ if } j\ne j'}} b^2(l_1,\dots,l_s), \quad 1\le s\le k.$$

Then

$$EY_s^2 = s!B_s^2,$$

and

$$EY_s^4 \le 1\cdot 3\cdot 5\cdots(4s-1)B_s^4 = \frac{(4s)!}{2^{2s}(2s)!}B_s^4$$

by Lemmas 13.4 and 13.5. Inequality (D12) follows from the last two relations. Indeed, to prove formula (D12) by means of these relations it is enough to check that $\frac{(4s)!}{2^{2s}(2s)!(s!)^2} \le 2^{4s}$. But it is easy to check this inequality with induction with respect to s. (Actually, there is a well-known inequality in the literature, known under the name

Borell's inequality, which implies inequality (D12) with a better coefficient at the right hand side of this estimate.)

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 $V \subset \{1, \dots, k\}$. I show with the help of Lemma D4 and formula (D8) that there exists some constant $c_k > 0$ such that the random variables $T_{n,k}f(\ell))$ defined in formula (D5) and $\bar{I}_{n,k,V}(f(\ell))$ defined in formula (D7) satisfy the inequality

$$P\left(\|2^k \bar{I}_{n,k,V}(f(\ell))\| > \|T_{n,k}(f(\ell))\| | \mathcal{F}\right) \ge c_k \quad \text{with probability 1.} \tag{D13}$$

In the proof I shall exploit that in formula (D8) $2^n \bar{I}_{n,k,V}(f(\ell))$ is represented by a polynomial of the Rademacher functions $\varepsilon_1, \dots, \varepsilon_n$ whose constant term is $T_{n,k}(f(\ell))$. The coefficients of this polynomial are functions of the random variables $\xi_l^{(1)}$ and $\xi_l^{(2)}$, $1 \le l \le n$. The independence of these random variables from ε_l, $1 \le l \le n$, and the definition of the σ-algebra $\mathcal{F}$ yield that

$$\begin{aligned}
&P\left(\|2^k \bar{I}_{n,k,V}(f(\ell))\| > \|T_{n,k}(f(\ell))\| | \mathcal{F}\right) \\
&\quad = P_{\varepsilon_V}\Bigg(\Bigg\| \frac{1}{k!} \sum_{\substack{(l_1,\dots,l_k),\ (s_1,\dots,s_k):\\ 1 \le l_j \le n, s_j = 1 \text{ or } s_j = 2,\ j=1,\dots,k,}} (1 + \kappa_{s_1,V}^{(1)} \varepsilon_{l_1}) \cdots (1 + \kappa_{s_k,V}^{(k)} \varepsilon_{l_k}) f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(s_1)}, \dots, \xi_{l_k}^{(s_k)}\right) \Bigg\| \\
&\qquad\qquad > \|T_{n,k}(f(\ell))(\xi_l^{(j)},\ 1 \le l \le n,\ j = 1, 2)\| \Bigg),
\end{aligned} \tag{D14}$$

where P_{ε_V} means that the values of the random variables $\xi_l^{(1)}$, $\xi_l^{(2)}$, $1 \le l \le n$, are fixed, (their value depend on the atom of the σ-algebra $\mathcal{F}$ we are considering) and the probability is taken with respect to the remaining random variables ε_l, $1 \le l \le n$. At the right-hand side of (D14) the probability of such an event is considered that the norm of a polynomial of order k of the random variables $\varepsilon_1, \dots, \varepsilon_n$ is larger than $\|T_{n,k}(f(\ell))(\xi_l^{(j)},\ 1 \le l \le n,\ j = 1, 2)\|$. Beside this, the constant term of this polynomial equals $T_{n,k}(f(\ell))(\xi_l^{(j)},\ 1 \le l \le n,\ j = 1, 2)$. Hence this probability can be bounded by means of Lemma D4, and this result yields relation (D13).

As the distributions of $I_{n,k,V}(f(\ell))$ and $\bar{I}_{n,k,V}(f(\ell))$ agree, relation (D13) implies that

$$\begin{aligned}
&P\left(\|2^k I_{n,k,V}(f(\ell))\| \ge \frac{1}{3} \cdot 2^{1-k} u\right) = P\left(\|2^k \bar{I}_{n,k,V}(f(\ell))\| \ge \frac{1}{3} \cdot 2^{1-k} u\right) \\
&\quad \ge P\left(\|2^k \bar{I}_{n,k,V}(f(\ell))\| \ge \|T_{n,k}(f(\ell))\|,\ \|T_{n,k}(f(\ell))\| \ge \frac{1}{3} \cdot 2^{1-k} u\right) \\
&\quad = \int_{\{\omega:\ \|T_{n,k}(f(\ell))(\omega)\| \ge \frac{1}{3} \cdot 2^{1-k} u\}} P\left(\|2^k \bar{I}_{n,k,V}(f(\ell))\| > \|T_{n,k}(f(\ell))\| | \mathcal{F}\right) dP \\
&\quad \ge c_k P(3 \cdot 2^{k-1} \|T_{n,k}(f(\ell))\| \ge u).
\end{aligned}$$

The last inequality with the choice of any set $V \subset \{1, \dots, k\}$, $1 \le |V| \le k-1$, together with relation (D6) imply formula (D4).

Relation (14.12d) will be proved together with another inductive hypothesis with the help of relation (D4) by means of an induction procedure with respect to the order k of the U-statistic. To formulate the other inductive hypothesis some new quantities will be introduced. 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 $W = (U_1, \dots, U_s) \in \mathcal{W}(k)$ 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 value of the function $s_W(j)$, $1 \le j \le k$, in a point j is defined by the relation $j \in V_{s_W(j)}$. Let us introduce the expression

$$I_{n,k,W}(f(\ell)) = \frac{1}{k!} \sum_{(l_1,\dots,l_k)\colon\, 1 \le l_j \le n,\ j=1,\dots,k} f_{l_1,\dots,l_k}\left(\xi_{l_1}^{(s_W(1))}, \dots, \xi_{l_k}^{(s_W(k))}\right)$$
$$\text{for all } W \in \mathcal{W}(k).$$

An expression of the form $I_{n,k,W}(f(\ell))$, $W \in \mathcal{W}_k$, will be called a decoupled U-statistic with generalized decoupling. Given a partition $W = (U_1, \dots, U_s) \in \mathcal{W}_k$ let us call the number s of the elements of this partition the rank both of the partition W and of the decoupled U-statistic $I_{n,k,W}(f(\ell))$ with generalized decoupling.

Now I formulate the following hypothesis. For all $k \ge 2$ and $2 \le j \le k$ there exist some constants $C(k,j) > 0$ and $\delta(k,j) > 0$ such that for all $W \in \mathcal{W}_k$ a decoupled U-statistic $I_{n,k,W}(f(\ell))$ with generalized decoupling satisfies the inequality

$$P(\|I_{n,k,W}(f(\ell))\| > u) \le C(k,j) P\left(\|\bar{I}_{n,k}(f(\ell))\| > \delta(k,j)u\right)$$
$$\text{for all } 2 \le j \le k \text{ if the rank of } W \text{ equals } j. \tag{D15}$$

It will be proved by induction with respect to k that both relations (14.12d) and (D15) hold for U-statistics of order k. Let us observe that for $k = 2$ relation (14.12d) follows from (D4). Relation (D15) also holds for $k = 2$, since in this case we have to consider only the case $j = k = 2$, and relation (D15) clearly holds in this case with $C(2,2) = 1$ and $\delta(2,2) = 1$. Hence we can start our inductive proof with $k = 3$. First I prove relation (D15).

In relation (D15) the tail-distribution of decoupled U-statistics with generalized decoupling is compared with that of the decoupled U-statistic $\bar{I}_{n,k}(f(\ell))$ introduced in (D2). Given the order k of these U-statistics it will be proved by means of a backward induction with respect to the rank j of the decoupled U-statistics $I_{n,k,W}(f(\ell))$ with generalized decoupling.

Relation (D15) clearly holds for $j = k$ with $C(k,k) = 1$ and $\delta(k,k) = 1$. To prove it for decoupled U-statistics with generalized decoupling of rank $2 \le j < k$ first the following observation will be made. 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 $U_j = \{t, \dots, k\}$ with some $2 \le t \le k-1$. 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 arguments in the U-statistics if it is necessary. Let us consider the partition $\bar{W} = (U_1, \dots, U_{j-1}, \{t\}, \dots, \{k\})$ and the decoupled U-statistic $I_{n,k,\bar{W}}(f(\ell))$ with generalized decoupling corresponding to this partition $\bar{W}$. It will be shown that our inductive hypothesis implies the inequality

$$P(\|I_{n,k,W}(f(\ell))\| > u) \le \bar{A}(k) P\left(\|I_{n,k,\bar{W}}(f(\ell))\| > \bar{\gamma}(k)u\right) \tag{D16}$$

with $\bar{A}(k) = \sup_{2 \le p \le k-1} A(p)$, $\bar{\gamma}(k) = \inf_{2 \le p \le k-1} \gamma(p)$ if the rank j of W is such that $2 \le j \le k-1$, where the constants $A(p)$ and $\gamma(p)$ agree with the corresponding coefficients in formula (14.12d).

To prove relation (D16) (in the case when $U_j = \{t, \dots, k\}$) let us define the σ-algebra $\mathcal{F}$ generated by the random variables appearing in the first $t-1$ coordinates of these U-statistics, i.e. by the random variables $\xi_{l_j}^{s_W(j)}$, $1 \le j \le t-1$, and $1 \le l_j \le n$ for all $1 \le j \le t-1$. We have $2 \le t \le k-1$. By our inductive hypothesis relation (14.12d) holds for U-statistics of order $p = k-t+1 \le k-1$. I claim that this implies that

$$P(\|I_{n,k,W}(f(\ell))\| > u|\mathcal{F}) \le A(k-t+1) P\left(\|I_{n,k,\bar{W}}(f(\ell))\| > \gamma(k-t+1)u|\mathcal{F}\right) \tag{D17}$$

with probability 1. Indeed, by the independence properties of the random variables $\xi_l^{s_W(j)}$ (and $\xi_l^{s_{\bar{W}}(j)}$), $1 \le j \le k$, $1 \le l \le n$,

$$P(\|I_{n,k,W}(f(\ell))\| > u|\mathcal{F}) = P_{\xi_l^{s_W(j)}, 1 \le j \le t-1}(\|I_{n,k,W}(f(\ell)\| > u)$$

and

$$P\left(\|I_{n,k,\bar{W}}(f(\ell))\| > \gamma(k-t+1)u|\mathcal{F}\right) = P_{\xi_l^{s_W(j)}, 1 \le j \le t-1}(\|I_{n,k,\bar{W}} f(\ell)\| > \gamma(k-t+1)u),$$

where $P_{\xi_l^{s_W(j)}, 1 \le j \le t-1}$ denotes that the values of the random variables $\xi_l^{s_W(j)}(\omega)$, $1 \le j \le t-1$, $1 \le l \le n$, are fixed, and we consider the probability that the appropriate functions of these fixed values and of the remaining random variables $\xi^{s_W(j)}$ and $\xi^{s_{\bar{W}}(j)}$, $t \le j \le k$, satisfy the desired relation. These identities and the relation between the sets W and $\bar{W}$ imply that relation (D17) is equivalent to the identity (14.12d) for the generalized U-statistics of order $k-t+1 \le k-1$ with kernel functions

$$f_{l_t,\dots,l_k}(x_t, \dots, x_k) = \sum_{(l_1,\dots,l_{t-1}):\ 1 \le l_j \le n,\ 1 \le j \le t-1} f_{l_1,\dots,l_k}(\xi_{l_1}^{s_W(1)}(\omega), \dots, \xi_{l_{t-1}}^{s_W(t-1)}(\omega), x_t, \dots, x_k).$$

Relation (D16) follows from inequality (D17) if expectation is taken at both sides. As the rank of $\bar{W}$ is strictly greater than the rank of W, relation (D16) together with our backward inductive assumption imply relation (D15) for all $2 \le j \le k$.

Relation (D15) implies in particular (with the applications of partitions of order k and rank 2) that the terms in the sum at the right-hand side of (D4) satisfy the inequality $P\left(D_k\|I_{n,k,V}(f(\ell))\| > u\right) \le \bar{C}(k,j)P\left(\|\bar{I}_{n,k}(f(\ell))\| > \bar{D}_k u\right)$ with some appropriate $\bar{C}_k > 0$ and $\bar{D}_k > 0$ for all $V \subset \{1,\dots,k\}$, $1 \le |V| \le k-1$. This inequality together with relation (D4) imply that inequality (14.12d) also holds for the parameter k.

In such a way we get the proof of relation (14.12d). Let us prove formula (14.13) with its help first in the simpler case when the supremum of finitely many functions is taken. If $M < \infty$ functions $f_1,\dots,f_M$ are considered, then relation (14.13) for the supremum of the U-statistics and decoupled U-statistics with these kernel functions can be derived from formula (14.12) if it is applied for the function $f = (f_1,\dots,f_M)$ with values in the separable Banach space B_M which consists of the vectors $(v_1,\dots,v_M)$, $v_j \in B$, $1 \le j \le M$, and the norm $\|(v_1,\dots,v_M)\| = \sup\limits_{1\le j\le m}\|v_j\|$ is introduced in it. The application of formula (14.12) with this choice yields formula (14.13) for this supremum. 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 that of $\sup\limits_{1\le s<\infty}\|I_{n,k}(f_s)\|$, and the distribution of the random variables $\sup\limits_{1\le s\le M}\left\|\bar{I}_{n,k}(f_s)\right\|$ converge to that of $\sup\limits_{1\le s<\infty}\left\|\bar{I}_{n,k}(f_s)\right\|$ as $M\to\infty$, relation (14.13) in the general case follows from its already proved special case and a limiting procedure $M\to\infty$.

Remark. The above proved formula (14.12d) can be slightly generalized. It also holds if the expressions $I_{n,k)}f(\ell)$ and $\bar{I}_{n,k}(f(\ell)$ appearing in this inequality are defined in a more general way. Namely, they are the random functions introduced in formulas (D1) and (D2), but the sequences $\xi_1,\dots,\xi_n$ and their independent copies $\xi_1^{(j)},\dots,\xi_n^{(j)}$ in these formulas are independent random variables which may also be non-identically distributed. This generalization can be shown without any essential change in the original proof.

References:

1.) Adamczak, R. (2006) Moment inequalities for U-statistics. *Annals of Probability* **34**, 2288–2314

2.) Alexander, K. (1987) The central limit theorem for empirical processes over Vapnik–Červonenkis classes. *Annals of Probability* **15**, 178–203

3.) Arcones, M. A. and Giné, E. (1993) Limit theorems for U-processes. *Annals of Probability*, **21**, 1494–1542

4.) 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

5.) Bennett, G. (1962) Probability inequality for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57**, 33–45

6.) Bonami, A. (1970) Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Ann. Inst. Fourier (Grenoble)* **20** 335–402

7.) 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

8.) 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

9.) Dobrushin, R. L. (1979) Gaussian and their subordinated fields. *Annals of Probability* **7**, 1-28

10.) Dudley, R. M. (1978) Central limit theorems for empirical measures. *Annals of Probability* **6**, 899–929

11.) Dudley, R. M. (1984) A course on empirical processes. *Lecture Notes in Mathememtics* **1097**, 1–142 Springer Verlag, New York

12.) Dudley, R. M. (1989) *Real Analysis and Probability.* Wadsworth & Brooks, Pacific Grove, California

13.) Dudley, R. M. (1998) *Uniform Central Limit Theorems.* Cambridge University Press, Cambridge U.K.

14.) Dynkin, E. B. and Mandelbaum, A. (1983) Symmetric statistics, Poisson processes and multiple Wiener integrals. *Annals of Statistics* **11**, 739–745

15.) Frankl, P. and Pach J. (1983) On the number of sets in null-t-design. *European J. Combinatorics* **4** 21–23

16.) Giné, E. and Guillou, A. (2001) On consistency of kernel density estimators for randomly censored data: Rates holding uniformly over adaptive intervals. *Ann. Inst. Henri Poincaré PR* **37** 503–522

17.) Giné, E., Kwapień, S, Latała, R. and Zinn, J. (2001) The LIL for canonical U-statistics of order 2. Annals of Probability **29** 520–527

18.) Giné, E., Latała, R. and Zinn, J. (2000) Exponential and moment inequalities for U-statistics in *High dimensional probability II.* Progress in Probability 47. 13–38. Birkhäuser Boston, Boston, MA.

19.) Gross, L. (1975) Logarithmic Sobolev inequalities. Amer. J. Math. **97**, 1061–1083

20.) Guionnet, A. and Zegarlinski, B. (2003) Lectures on Logarithmic Sobolev inequalities. *Lecture Notes in Mathematics* **1801** 1–134 2. Springer Verlag, New York

21.) Hanson, D. L. and Wright, F. T. (1971) A bound on the tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.* **42** 52–61

22.) Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325

23.) Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Math. Society* **58**, 13–30

24.) Itô K. (1951) Multiple Wiener integral. *J. Math. Soc. Japan* **3**. 157–164

25.) Kaplan, E.L. and Meier P. (1958) Nonparametric estimation from incomplete data, *Journal of American Statistical Association*, **53**, 457–481.

26.) Latała, R. (2006) Estimates of moments and tails of Gaussian chaoses. *Annals of Probability* **34** 2315–2331

27.) Ledoux, M. (1996) On Talagrand deviation inequalities for product measures. *ESAIM: Probab. Statist.* **1.** 63–87. Available at http://www.emath./fr/ps/.

28.) Major, P. (1981) Multiple Wiener–Itô integrals. *Lecture Notes in Mathematics* **849**, Springer Verlag, Berlin, Heidelberg, New York,

29.) Major, P. (1988) On the tail behaviour of the distribution function of multiple stochastic integrals. *Probability Theory and Related Fields*, **78**, 419–435

30.) Major, P. (1994) Asymptotic distributions for weighted U-statistics. *The Annals of Probability*, **22** 1514–1535

31.) Major, P. (2005) An estimate about multiple stochastic integrals with respect to a normalized empirical measure. *Studia Scientarum Mathematicarum Hungarica.* 295–341

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

33.) Major, P. (2006) An estimate on the maximum of a nice class of stochastic integrals. *Probability Theory and Related Fields.* **134**, 489–537

34.) Major, P. (2006) A multivariate generalization of Hoeffding's inequality. *Electronic Communication in Probability* **2** (220–229)

35.) Major, P. (2007) On a multivariate version of Bernstein's inequality *Electronic Journal of Probability* **12** 966–988

36.) Major, P. (2005) On the tail behaviour of multiple random integrals and degenerate U-statistics. (First version of this lecture note) http://www.renyi.hu/˜major

37.) Major, P. and Rejtő, L. (1988) Strong embedding of the distribution function under random censorship. *Annals of Statistics* **16**, 1113–1132

38.) Major, P. and Rejtő, L. (1998) A note on nonparametric estimations. In the conference volume to the 65. birthday of Miklós Csörgő. 759–774

39.) Malyshev, V. A. and Minlos, R. A. (991) Gibbs Random Fields. Method of cluster expansion. Kluwer, Academic Publishers, Dordrecht

40.) Massart, P. (2000) About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability* **28**, 863–884

41.) Mc. Kean, H. P. (1973) Wiener's theory of non-linear noise. in *Stochastic Differential Equations* SIAM–AMS Proc. 6 197–209

42.) Nelson, E. (1973) The free Markov field. J. Functional Analysis **12**, 211–227

43.) Pollard, D. (1984) *Convergence of Stochastic Processes.* Springer Verlag, New York

44.) Rota, G.-C. and Wallstrom, C. (1997) Stochastic integrals: a combinatorial approach. *Annals of Probability* **25** (3) 1257–1283

45.) Surgailis, D. (1984) On multiple Poisson stochastic integrals and associated Markov semigroups. *Probab. Math. Statist.* 3. no. **2** 217-239

46.) Surgailis, D. (2000) Long-range dependence and Appell rank. *Annals of Probability* **28** 478–497

47.) Szegő, G. (1967) *Orthogonal Polynomials.* American Mathematical Society Colloquium Publications. Vol. 23

48.) Takemura, A. (1983) Tensor Analysis of ANOVA decomposition. *J. Amer. Statist. Assoc.* **78**, 894–900

49.) Talagrand, M. (1994) Sharper bounds for Gaussian and empirical processes. *Annals of Probability* **22**, 28–76

50.) Talagrand, M. (1996) New concentration inequalities in product spaces. *Invent. Math.* **126**, 505–563

51.) Talagrand, M. (2005) *The general chaining.* Springer Monographs in Mathematics. Springer Verlag, Berlin Heidelberg New York

52.) Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory.* Springer Verlag, New York

CONTENT