Chapter 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 $(X, \mathscr{X})$ 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 $(X^k, \mathscr{X}^k)$ of the space $(X, \mathscr{X})$, introduce the k-th power of the normalized empirical measure $\sqrt{n}(\mu_n - \mu)$ on $(X^k, \mathscr{X}^k)$ 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 $\mathscr{F}$ of k variables on the product space $(X^k, \mathscr{X}^k)$, 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.

One may ask 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 I 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 problem is hard in its original form. But by applying an appropriate Taylor expansion of the function that appears in this equation and throwing out its higher order terms we get an approximation whose behaviour can be well understood. So to describe the limit behaviour of the maximum likelihood estimate it suffices to show that this approximation causes only a negligible error.

P. Major, *On the Estimation of Multiple Random Integrals and U-Statistics*,
Lecture Notes in Mathematics 2079, DOI 10.1007/978-3-642-37617-7_1,

One would try to apply a similar method in the study of more difficult questions. 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 show that the simplifying approximation causes only a negligible error. In this case 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, which is a fairly popular subject in probability theory. The investigation of multiple random integrals and U-statistics are closely related, and it turned out to be 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 analogous problems about multiple Gaussian integrals. The behaviour of multiple Gaussian integrals, called Wiener–Itô integrals in the literature, is fairly well understood, and it suggests that the tail distribution of a k-fold random integral with respect to a normalized empirical measure should satisfy such estimates as the tail distribution of the k-th power of a Gaussian random variable with expectation zero and appropriate variance. Beside this, if we consider 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, then we expect that under not too restrictive conditions this supremum is not much larger than the "worst" random integral with the largest variance taking part in this supremum. 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 presented heuristic considerations supply a fairly good description of the situation, but they do 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 hardest problems discussed in this work are related to this phenomenon.

The precise formulation of the results will be given in the main part of the work. Beside their proofs 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 are discussed in detail together with some of their 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, about so-called decoupled U-statistics and their relation to ordinary 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.

As all the topics treated in the individual chapters relate to each other it seemed natural to me to tell the history of how the various results were reached in one last chapter. This last chapter, Chap. 18, just before the Appendix, also contains the complete reference list. I tried to give satisfactory referencing to all essential problems discussed, concentrate on explaining the main ideas behind the proofs and indicate where they were published. I did not attempt to provide an exhaustive literature list for fear that more would be less. As a consequence the reference list reflects my subjective preferences, my way of thinking.

http://www.springer.com/978-3-642-37616-0

On the Estimation of Multiple Random Integrals and U-Statistics

Major, P.

2013, XIII, 288 p. 11 illus., Softcover
ISBN: 978-3-642-37616-0