EMPIRICAL PROCESSES

AND STATISTICS OF EXTREME VALUES

John H.J. Einmahl[1]
Laurens de Haan[2]

AIO Course
Academic year 1999-2000
Utrecht

[1]Eindhoven University of Technology and EURANDOM, Eindhoven; email: j.h.j.einmahl@tue.nl
[2]Erasmus University, Rotterdam; email: ldehaan@few.eur.nl

CONTENTS OF THE 'EMPIRICAL PROCESSES' PART

1. INTRODUCTION AND ELEMENTARY PROPERTIES

The concept of random sampling from a probability distribution is of fundamental importance in statistics. As the relevant information contained in the sample may be summarized in the empirical probability distribution, it will be clear that the latter concept is of equal importance. Typically in statistics the sampled probability distribution is partly or entirely unknown and the empirical probability distribution enables us to learn about the unknown underlying distribution with increasing accuracy as the sample size increases. A proper formulation of this loose statement is sometimes called the fundamental theorem of mathematical statistics.

Since many important statistics can be conveniently written in terms of the empirical probability distribution, it is a fundamental concept behind various statistical procedures. In particular when the question of the asymptotic distribution of these statistics arises, certain properties of empirical probability distributions turn out to be extremely useful. Especially in non- and semiparametric statistics the empirical probability distribution has various interesting applications, e.g. in density estimation, nonparametric regression, random censoring, and extreme value theory.

It will become clear later, that apart from statistical relevance, the empirical probability distribution is also a very interesting object from a probabilistic point of view. Especially uniform laws of large numbers, central limit theorems and laws of the iterated logarithm for the associated empirical process and its modifications and extensions yield challenging problems of probabilistic nature. Therefore the topic of this series of lectures can be viewed as lying in the broad intersection of mathematical statistics and probability theory, with offshoots in both directions.

Let us now proceed to give the definition of the empirical distribution function which determines the empirical probability distribution, in the case that our random elements are real-valued. We will mainly, but not only, focus on this one dimensional situation in the sequel. Throughout all random variables will be defined on a sufficiently rich probability space $(\Omega, \mathcal{F}, I\!\!P)$. Let $(I\!\!R, \mathcal{B})$ be the set of real numbers equipped with the Borel σ-field. A random variable X is a mapping

$$X : \Omega \to I\!\!R$$

which is supposed to be $(\mathcal{F}, \mathcal{B})$-measurable. This induces a probability measure on $I\!\!R$ according to

$$P(B) = I\!\!P(\omega \in \Omega : X(\omega) \in B),\ B \in \mathcal{B}\ ,$$

which is also called the law $\mathcal{L}(X)$ of X. Let us next assume that $X_1, X_2, \ldots$ is a sequence of independent random variables, each with law P, i.e. $\mathcal{L}(X_i) = P$ for $i \in I\!\!N$, and with corresponding right-continuous distribution function F, defined by $F(t) = I\!\!P(X_1 \le t), t \in I\!\!R$. Now we call $X_1, \ldots, X_n$ a random sample of size n from P. Statistical model assumptions take the form $P \in \mathcal{P}$ for a given class $\mathcal{P}$ of probability measures on $I\!\!R$.

We define the empirical distribution function at stage $n \in I\!\!N$, the usual estimator of F for sample size n, to be

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(X_i) = \frac{1}{n}\#\ \{1 \le i \le n : X_i \le t\},\ t \in I\!\!R\ .$$

Of course, $F_n(t)$ does not depend only on t but also on ω. In fact we have for fixed $\omega \in \Omega$ that F_n is a distribution function pertaining to a discrete measure which assigns mass $1/n$ to each $X_i(\omega)$. On the other hand for fixed $t \in I\!R$, $F_n(t)$ is a random variable taking values in $[0, 1]$. To be more precise, observe that $nF_n(t)$ has a binomial distribution with parameters n and $F(t)$, which entails that

$$I\!E F_n(t) = F(t), \text{ Var } F_n(t) = \frac{1}{n} F(t)(1 - F(t)) \ .$$

Hence $F_n(t)$ is an unbiased estimator of $F(t)$. We also know from the strong law of large numbers that

$$(1.1) \qquad F_n(t) \to F(t) \text{ a.s. } (n \to \infty) \ ,$$

meaning that $I\!P(\lim_{n\to\infty} F_n(t) = F(t)) = 1$. Writing (1.1) as $|F_n(t) - F(t)| \to 0$ a.s., raises immediately the question whether this convergence is uniform in $t \in I\!R$ and, if so, what can be said about the rate of convergence. The first part of this question is indeed answered affirmatively by the theorem below; the second part will be addressed later.

The central limit theorem immediately gives that

$$n^{1/2}(F_n(t) - F(t)) \to^d N(0, F(t)(1 - F(t))) \ ,$$

yielding the obvious question whether this convergence also can be made uniform in t; this is the second main question. Answering both questions and some ramifications satisfactory will be our main goal in the first part of this course.

THEOREM 1.1 (Glivenko-Cantelli theorem; fundamental theorem of mathematical statistics). For any distribution function F we have

$$(1.2) \qquad \sup_{t\in I\!R} |F_n(t) - F(t)| \to 0 \text{ a.s. } (n \to \infty) \ .$$

FIRST PROOF. It is sufficient to show that for each $m \in I\!N$

$$(1.3) \qquad \limsup_{n\to\infty} \sup_{t\in I\!R} |F_n(t) - F(t)| \le 1/(m + 1) \text{ a.s.}$$

Choose $-\infty = t_{0,m} < t_{1,m} \le ... \le t_{m,m} < t_{m+1,m} = \infty$ such that $F(t_{k,m}-) \le k/(m + 1)$ $\le F(t_{k,m}), k \in \{1, ..., m\}$, and observe that

$$|F_n(t) - F(t)| = \max\ (F_n(t) - F(t), F(t) - F_n(t)) \ .$$

For any $t \in [t_{k-1,m}, t_{k,m}), k \in \{1, ..., m + 1\}$ we have

$$F_n(t) - F(t) \le F_n(t_{k,m}-) - F(t_{k-1,m})$$

$$= F_n(t_{k,m}-) - F(t_{k,m}-) + F(t_{k,m}-) - F(t_{k-1,m})$$

$$\le |F_n(t_{k,m}-) - F(t_{k,m}-)| + 1/(m + 1) \ .$$

Similarly we have

$$F(t) - F_n(t) \leq |F_n(t_{k-1,m}) - F(t_{k-1,m})| + 1/(m+1) \ .$$

Because $F_n(-\infty) - F(-\infty) = F_n(\infty) - F(\infty) = 0$ this yields

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \leq \max_{k \in \{1,...,m\}} |F_n(t_{k,m}-) - F(t_{k,m}-)| + \max_{k \in \{1,...,m\}} |F_n(t_{k,m}) - F(t_{k,m})| + 1/(m+1) \ . \tag{1.4}$$

The ordinary strong law of large numbers entails that

$$F_n(t_{k,m}-) \to F(t_{k,m}-) \text{ and } F_n(t_{k,m}) \to F(t_{k,m}) \text{ a.s. } (n \to \infty) \ .$$

Because (1.4) involves only the maximum of only a finite number of random variables, (1.3) follows. □

SECOND PROOF. We begin this proof with two symmetrization procedures. Let $X_1', ..., X_n'$ be another sample from P, defined on the same probability space and independent of $X_1, ..., X_n$. Then we have for any $\varepsilon > 0$ and $n \geq 2/\varepsilon^2$ that

$$\mathbb{P}(\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| > \varepsilon) \leq 2\mathbb{P}(\sup_{t \in \mathbb{R}} |F_n(t) - F_n'(t)| > \varepsilon/2) \ , \tag{1.5}$$

where F_n' is the empirical distribution function of the new sample. In order to prove (1.5), first observe that by Chebyshev's inequality we have that for any $t \in \mathbb{R}$ and for $n \geq 2/\varepsilon^2$, since $F(t)(1 - F(t)) \leq \frac{1}{4}$,

$$\mathbb{P}(|F_n'(t) - F(t)| \leq \varepsilon/2) \geq 1 - \frac{1}{n\varepsilon^2} \geq \tfrac{1}{2} \ . \tag{1.6}$$

Now select a random variable τ, independent of F_n', for which $|F_n(\tau) - F(\tau)| > \varepsilon$ on the event $\{\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| > \varepsilon\}$. Note that (1.6) remains true when we replace t by τ, even when we condition on F_n. This yields almost immediately (1.5):

$$\begin{aligned} \tfrac{1}{2}\mathbb{P}(\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| > \varepsilon) &\leq \mathbb{P}(|F_n'(\tau) - F(\tau)| \leq \varepsilon/2 \mid \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| > \varepsilon) \\ &\quad \cdot \mathbb{P}(\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| > \varepsilon) \end{aligned} \tag{1.7}$$

$$= I\!P(|F_n'(\tau) - F(\tau)| \leq \varepsilon/2,\ \sup_{t\in I\!R} |F_n(t) - F(t)| > \varepsilon)$$

$$= I\!P(|F_n'(\tau) - F(\tau)| \leq \varepsilon/2,\ |F_n(\tau) - F(\tau)| > \varepsilon)$$

$$\leq I\!P(|F_n(\tau) - F_n'(\tau)| > \varepsilon/2) \leq I\!P(\sup_{t\in I\!R} |F_n(t) - F_n'(t)| > \varepsilon/2)\ .$$

Now we continue with the second symmetrization. Therefore we define on our probability space n independent and identically distributed random variables $\eta_1, ..., \eta_n$ with

$$I\!P(\eta_1 = 1) = I\!P(\eta_1 = -1) = \tfrac{1}{2}\ ;$$

the η_i's are independent from the X_i and X_i'. Observe that

$$\mathcal{L}(1_{(-\infty,t]}(X_i) - 1_{(-\infty,t]}(X_i'), i = 1, ..., n,\ \ t \in I\!R)$$

$$= \mathcal{L}(\eta_i(1_{(-\infty,t]}(X_i) - 1_{(-\infty,t]}(X_i')), i = 1, ..., n,\ \ t \in I\!R)\ .$$

Hence

$$\text{(1.8)} \qquad I\!P\left(\sup_{t\in I\!R} |F_n(t) - F_n'(t)| > \varepsilon/2\right)$$

$$= I\!P\left(\sup_{t\in I\!R} \left|\frac{1}{n}\sum_{i=1}^n \eta_i(1_{(-\infty,t]}(X_i) - 1_{(-\infty,t]}(X_i'))\right| > \varepsilon/2\right)$$

$$\leq I\!P\left(\sup_{t\in I\!R} \left|\frac{1}{n}\sum_{i=1}^n \eta_i 1_{(-\infty,t]}(X_i)\right| > \varepsilon/4\right)$$

$$+ I\!P\left(\sup_{t\in I\!R} \left|\frac{1}{n}\sum_{i=1}^n \eta_i 1_{(-\infty,t]}(X_i')\right| > \varepsilon/4\right)$$

$$= 2I\!P(\sup_{t\in I\!R} |F_n^s(t)| > \varepsilon/4)\ ,$$

where $F_n^s(t) = \dfrac{1}{n}\sum_{i=1}^n \eta_i 1_{(-\infty,t]}(X_i),\ t \in I\!R$. Combination of (1.5) and (1.8) yields that for $\varepsilon > 0$ and $n > 2/\varepsilon^2$

$$\text{(1.9)} \qquad I\!P\left(\sup_{t\in I\!R} |F_n(t) - F(t)| > \varepsilon\right) \leq 4I\!P\left(\sup_{t\in I\!R} |F_n^s(t)| > \varepsilon/4\right)\ .$$

To bound the right hand side of (1.9), we work conditionally on the observations $X_1, ..., X_n$, leaving only the randomness contributed by the η_i. Once the locations of the X_i are fixed the supremum on the right in (1.9), reduces to a maximum taken over a strategically chosen set of point t_j, $j = 1, ..., n$ (we can e.g. choose $t_j = X_j(\omega), j = 1, ..., n$). We find

$$(1.10) \qquad I\!P\left(\sup_{t \in I\!R} |F_n^s(t)| > \varepsilon/4 \mid X_i = x_i, i = 1, ..., n \right)$$

$$\leq \sum_{j=1}^{n} I\!P(|F_n^s(t_j)| > \varepsilon/4 \mid X_i = x_i, i = 1, ..., n) .$$

Let $Y_1, ..., Y_n$ be independent random variables with mean zero and bounded range $a_i \leq Y_i \leq b_i$. Then Hoeffding's inequality asserts that for each $\lambda > 0$

$$(1.11) \qquad I\!P\left(\left| \sum_{i=1}^{n} Y_i \right| \geq \lambda \right) \leq 2 \exp\left(-2\lambda^2 / \sum_{i=1}^{n} (b_i - a_i)^2 \right) .$$

Applying this inequality with $Y_i = \eta_i 1_{(-\infty, t]}(X_i)$ we readily obtain that

$$I\!P(|F_n^s(t)| > \varepsilon/4 \mid X_i = x_i, i = 1, ..., n)$$

$$\leq 2 \exp(-2(n\varepsilon/4)^2 / (4nF_n(t))) \leq 2 \exp(-n\varepsilon^2/32) .$$

Use this for each t_j in (1.10):

$$(1.12) \qquad I\!P\left(\sup_{t \in I\!R} |F_n^s(t)| > \varepsilon/4 \mid X_i = x_i, i = 1, ..., n \right)$$

$$\leq 2n \exp(-n\varepsilon^2/32) .$$

Noting that the right hand side of (1.12), does not depend on the X_i, we see from (1.9) and (1.12) that for all $\varepsilon > 0$ and all $n \in I\!N$

$$I\!P\left(\sup_{t \in I\!R} |F_n(t) - F(t)| > \varepsilon \right) \leq 8n \exp(-n\varepsilon^2/32) ,$$

yielding that $\sum_{n=1}^{\infty} I\!P\left(\sup_{t \in I\!R} |F_n(t) - F(t)| > \varepsilon\right) < \infty$, for all $\varepsilon > 0$. An application of the Borel-Cantelli lemma completes the proof. □

Apart from the fact that we are now completely sure that F_n converges uniformly to F almost surely, there is another reason for proving Theorem 1.1 twice. The first proof is incorporated because it is more or less standard and easy since it is based on the strong law of large numbers for a sequence of independent random variables. The second proof is more modern and has therefore the attendant advantage that it can be extended to a much more general setup, which we will present now. Until further notice, we will assume that our X_i are not random variables defined on $I\!R$, but independent and identically distributed stochastic elements in some measurable space $(\mathcal{S}, \mathcal{E})$, entailing that $X_i : \Omega \to \mathcal{S}$ is a $(\mathcal{F}, \mathcal{E})$-measurable mapping; typically $\mathcal{S}$ is a space with some topological structure. The empirical probability measure P_n is now defined by

$$(1.13) \qquad P_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i), \quad A \in \mathcal{E}$$

and as before the common "true" probability measure of the X_i will be denoted by P. Obviously, for fixed $A \in \mathcal{E}$, $P_n(A)$ has similar properties as mentioned before for $F_n(t)$, in particular

$$(1.14) \qquad P_n(A) \to P(A) \text{ a.s. } (n \to \infty) \ .$$

But we want to show that (1.14) holds uniformly over a subset $\mathcal{A}$ of $\mathcal{E}$. A little reflection shows that choosing $\mathcal{A} = \mathcal{E}$ is often not sensible; e.g. if $(\mathcal{S}, \mathcal{E}) = (I\!R, \mathcal{B})$ and P a probability measure with a continuous distribution function, then $\sup_{A \in \mathcal{E}} |P_n(A) - P(A)| = 1$ a.s. for all $n \in I\!N$.

The following two properties are sufficient for almost sure convergence uniformly over $\mathcal{A}$:

(CG) $\mathcal{A}$ is countably generated, meaning that there exists a countable subclass $\mathcal{D}$ of $\mathcal{A}$, such that for any $A \in \mathcal{A}$ there is a sequence $\{D_n, n \in I\!N\}$ in $\mathcal{D}$ with $1_{D_n}(x) \to 1_A(x)$ for all $x \in \mathcal{S}$.

(VC) $\mathcal{A}$ is a Vapnik-Cervonenkis class (or a class of polynomial discrimination), meaning that there exists a polynomial p such that from every set of N points in $\mathcal{S}$, the class picks out at most $p(N)$ distinct subsets. Formally, if $\{s_1, ..., s_N\} \subset \mathcal{S}$, then there are at most $p(N)$ distinct sets of the form $\{s_1, ..., s_N\} \cap A$ with $A \in \mathcal{A}$.

Now we are ready to present the generalization of the Glivenko-Cantelli theorem.

THEOREM 1.2 (uniform strong law of large numbers). For every $\mathcal{A} \subset \mathcal{E}$ satisfying CG and VC we have

$$(1.15) \qquad \sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \to 0 \text{ a.s. } (n \to \infty) \ .$$

Rather then giving a formal proof of (1.15) we only mention that it follows closely the lines of the second proof of Theorem 1.1. Apart from the obvious changes, the main adaptation is that the n multiplier (the number of t_j's) has to be replaced by $p(n)$.

Property CG is only assumed to avoid measurability problems and is often easy to check. Much more interesting is the question which classes $\mathcal{A}$ satisfy th VC property. Here are, without proof, some examples:

(a) the class $\{(-\infty, t] : t \in \mathbb{R}^d\}$ of lower quadrants in $\mathbb{R}^d$, $d \in \mathbb{N}$, where $(-\infty, t]$ $= \prod_{i=1}^{d} (-\infty, t_i]$,

(b) all rectangles in $\mathbb{R}^d$,

(c) all ellipsoids in $\mathbb{R}^d$,

(d) all half spaces in $\mathbb{R}^d$,

(e) all polyhedra in $\mathbb{R}^d$ with a bound on the number of sides,

(f) any subset of an $\mathcal{A}$ satisfying VC,

(g) any finite union of $\mathcal{A}$'s satisfying VC,

(h) the class $\{A^c : A \in \mathcal{A}\}$, where $\mathcal{A}$ satisfies VC,

(i) the class $\{A \cap \widetilde{A} : A \in \mathcal{A}, \widetilde{A} \in \widetilde{\mathcal{A}}\}$, where $\mathcal{A}$ and $\widetilde{\mathcal{A}}$ satisfy VC,

(j) the class $\{A \cup \widetilde{A} : A \in \mathcal{A}, \widetilde{A} \in \widetilde{\mathcal{A}}\}$, where $\mathcal{A}$ and $\widetilde{\mathcal{A}}$ satisfy VC,

(k) the class of sets of the form $\{\{x : g(x) \geq 0\}, g \in G\}$, where G is a finite-dimensional vector space of real functions on $\mathbb{S}$.

We conclude this extension to Vapnik-Cervonenkis indexing classes by mentioning a remarkable fact. Polynomials (in N) increase much more slowly than 2^N. For large N a Vapnik-Cervonenkis class must fail to pick out a least one of the 2^N subsets for each collection of N points in $\mathbb{S}$. Surprisingly this characterizes the VC property, meaning that if the number of sets picked out by a class is always less than 2^N (N large), then its is bounded by a polynomial.

Now it is time to return to $\mathbb{R}$ again, i.e. we assume that our X_i take values in $\mathbb{R}$, like in the beginning of this chapter. The random variables

$$\min_{1 \leq i \leq n} X_i = X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n} = \max_{1 \leq i \leq n} X_i \tag{1.16}$$

are called the order statistics and, for $i = 1, \ldots, n$, the random variable

$$R_{in} = \#\{1 \leq j \leq n : X_j \leq X_i\} = \sum_{j=1}^{n} 1_{[0,\infty)}(X_i - X_j) \tag{1.17}$$

is called the rank of X_i.

Ranks and order statistics are particular useful when the underlying distribution function F is continuous on $\mathbb{R}$ because then the sample elements are all different almost surely. For an arbitrary right-continuous function $G : \mathbb{R} \to [0, 1]$ that is non-decreasing let us define the generalized inverse

$$G^{-1}(u) = \inf\{t : G(t) \geq u\}, u \in [0, 1] \; , \tag{1.18}$$

with the convention that $\inf(\phi) = \infty, \inf(\mathbb{R}) = -\infty$. Note that this inverse is left-continuous on $(0,1]$. The following relations are immediate for continuous F:

$$(1.19) \qquad F_n(X_i) = \frac{1}{n} R_{in},\ F_n(X_{i:n}) \overset{\text{a.s.}}{=} \frac{i}{n},\ F_n^{-1}\left(\frac{i}{n}\right) = X_{i:n} \ .$$

Let $\mathbb{O}_n = \{\langle t_1, ..., t_n\rangle \in \mathbb{R}^n : t_1 \leq ... \leq t_n\}$ and $\mathcal{R}_n$ the set of all permutations of $\langle 1, ..., n\rangle$.

THEOREM 1.3. Provided that F is continuous we have, for any $O \subset \mathbb{O}_n$ with $O \in \mathcal{B}^n$ and any $\langle r_1, ..., r_n\rangle \in \mathcal{R}_n$,

$$(1.20) \qquad \mathbb{P}(\langle X_{1:n}, ..., X_{n:n}\rangle \in O) = n!\mathbb{P}(\langle X_1, ..., X_n\rangle \in O) \ ,$$

$$(1.21) \qquad \mathbb{P}(\langle R_{1n}, ..., R_{nn}\rangle = \langle r_1, ..., r_n\rangle) = 1/n! \ ,$$

$$(1.22) \qquad \langle X_{1:n}, ..., X_{n:n}\rangle \text{ and } \langle R_{1n}, ..., R_{nn}\rangle \text{ are independent.}$$

PROOF. Let $\langle \rho_1, ..., \rho_n\rangle$ be the inverse permutation of $\langle r_1, ..., r_n\rangle$. All statements follow from the relation

$$(1.23) \qquad \mathbb{P}(\langle R_{1n}, ..., R_{nn}\rangle = \langle r_1, ..., r_n\rangle,\ \langle X_{1:n}, ..., X_{n:n}\rangle \in O)$$

$$= \mathbb{P}(\langle X_{\rho_1}, .., X_{\rho_n}\rangle \in O) = \mathbb{P}(\langle X_1, ..., X_n\rangle \in O) \ .$$

Relation (1.20) is obtained from (1.23) by summing over all $n!$ permutations and exploiting the circumstance that the expression on the right in (1.23) is independent of the particular permutation $\langle r_1, ..., r_n\rangle$.

We obtain (1.21) by substituting $O = \mathbb{O}_n$ in (1.23) and noticing that

$$(1.24) \qquad \mathbb{P}(\langle X_1, ..., X_n\rangle \in \mathbb{O}_n) = \mathbb{P}(X_1 \leq ... \leq X_n) = 1/n! \ .$$

Combining (1.20) and (1.21) we see that $\mathbb{P}(\langle X_1, ..., X_n\rangle \in O) = \mathbb{P}(\langle R_{1n}, ..., R_{nn}\rangle = \langle r_1, ..., r_n\rangle) \cdot \mathbb{P}(\langle X_{1:n}, ..., X_{n:n}\rangle \in O)$. Substitution in (1.23) leads at once to (1.22). □

The joint distribution of the rank and order statististics being known, it is in principle possible to determine the marginal distributions. It is easy to see that (1.21) entails

$$\mathcal{L}(R_{in}) = UN\{1, ..., n\} \ .$$

It takes a bit more work to see from (1.20) that

$$(1.25) \qquad \mathbb{P}(X_{i:n} \leq t) = \frac{n!}{(i-1)!(n-i)!} \int_0^{F(t)} s^{i-1}(1-s)^{n-i} ds, \quad t \in \mathbb{R} \ .$$

It is interesting to observe that there are two more ways to find the distribution of $X_{i:n}$. The first one assumes the existence of the derivative f of F on $\mathbb{R}$ and leads to an expression for the density that corresponds to (1.25). A standard combinatorial argument based on the multinomial distribution shows

$$\begin{aligned}
&\mathbb{P}(t < X_{i:n} \le t+\Delta) = \mathbb{P}(\#\{j : X_j \le t\} = i-1 \ , \\
&\#\{j : t < X_j \le t+\Delta\} = 1, \#\{j : X_j > t+\Delta\} = n-i) + O(\Delta^2) \\
&= \frac{n!}{(i-1)!1!(n-i)!}\{F(t)\}^{i-1}\{F(t+\Delta) - F(t)\}\{1-F(t+\Delta)\}^{n-i} + O(\Delta^2) \ .
\end{aligned}$$

Dividing by Δ and taking the limit for $\Delta \downarrow 0$ we find that the density g of the distribution of $X_{i:n}$ equals

$$g(t) = \frac{n!}{(i-1)!(n-i)!}\{F(t)\}^{i-1}\{1-F(t)\}^{n-i}f(t), t \in \mathbb{R} \ .$$

The second one, which does not require any assumption about F, leads to a different expression. It is clear that

$$(1.26) \qquad \{X_{i:n} \le t\} = \left\{F_n(t) \ge \frac{i}{n}\right\} \ ,$$

from which we have at once

$$(1.27) \qquad \mathbb{P}(X_{i:n} \le t) = \sum_{k=i}^{n} \binom{n}{k} \{F(t)\}^k\{1-F(t)\}^{n-k}, t \in \mathbb{R} \ .$$

As a consequence we find the equality of the right-hand sides in (1.25) and (1.27), a well-known result in the theory of special functions.

It appears to be quite often convenient to make a transformation to the unit interval. Since Φ, the distribution function of the standard normal distribution, is one-to-one on $\mathbb{R}$ with range (0,1) it is clear that "without loss of information" we might consider the transformed random variabele $\xi_i = \Phi(X_i)$ with values in (0,1). Let us note that

$$\mathbb{P}(\xi_i \le t) = \mathbb{P}(X_i \le \Phi^{-1}(t)) = F(\Phi^{-1}(t)), t \in [0,1] \ .$$

Other transformations are of course also possible. A thus obtained distribution function that concentrates mass 1 on (0,1) will be denoted by Γ; the corresponding empirical distribution function at stage $n \in \mathbb{N}$ is defined by

$$(1.28) \qquad \Gamma_n(t) = \frac{1}{n}\#\{1 \le i \le n : \xi_i \le t\}, t \in [0,1] \ .$$

This course will concentrate on stochastic processes of the type

(1.29) $U_{n,\Gamma}(t) = n^{1/2}(\Gamma_n(t) - \Gamma(t)), \quad t \in [0,1]$;

any such process is called a (reduced) empirical process. When F is continuous, transformation by means of F itself turns out to be particularly interesting, since then we have for $\xi_i = F(X_i)$ that

(1.30) $\mathcal{L}(\xi_i) = UN(0,1)$.

This property is easily checked, the transformation is called the probability integral transformation. It is an important observation for nonparametric statistics. In this case we have

(1.31) $\Gamma(t) = \mathbb{P}(\xi_i \leq t) = \mathbb{P}(F(X_i) \leq t) = t = I(t), t \in [0,1]$,

and we will simply write

(1.32) $U_{n,\Gamma} = U_n, \text{ if } \Gamma = I$.

The process in (1.32) is called the uniform empirical process. Here is a consequence of (1.31).

THEOREM 1.4. For continuous F we have

(1.33) $\mathcal{L}(\sup_{t\in\mathbb{R}} n^{1/2}|F_n(t) - F(t)|) = \mathcal{L}(\sup_{t\in[0,1]} |U_n(t)|)$.

PROOF. According to (1.31) we have that $\xi_1, ..., \xi_n$ is a random sample of size n from the $UN(0,1)$-distribution. Restricting ourselves for convenience to continuous F that are also strictly increasing, we easily see that

$$\begin{aligned}
&n^{1/2} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \\
&= \sup_{t\in\mathbb{R}} n^{1/2}|\frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(X_i) - F(t)| \\
&= \sup_{t\in\mathbb{R}} n^{1/2}|\frac{1}{n}\sum_{i=1}^{n} 1_{[0,F(t)]}(F(X_i)) - F(t)| \\
&= \sup_{t\in[0,1]} n^{1/2}|\frac{1}{n}\sum_{i=1}^{n} 1_{[0,t]}(\xi_i) - t| \\
&= \sup_{t\in[0,1]} |U_n(t)| ,
\end{aligned}$$

which proves (1.33) in this case. Only minor modifications are required for arbitrary continuous F. $\square$

It can be shown that

$$(1.34)\qquad \mathbb{P}\Big(\sup_{t\in[0,1]} |U_n(t)| \le x\Big) \to 1 + 2\sum_{k=1}^{\infty}(-1)^k \exp(-2k^2x^2) = H(x) \quad (n\to\infty),\ x>0\ .$$

In fact a proof of (1.34) can be given by establishing first that $U_n, n = 1, 2, \ldots$, converges in distribution and then deriving the distribution of the supremum of the absolute value of the limit. This is another motivation for the study of convergence in distribution of empirical processes as random elements in a function space.

We now consider a straightforward application of Theorem 1.4 and (1.34). Suppose we wish to test the null hypothesis that $X_1, \ldots, X_n$ is a random sample from a completely specified continuous distribution function F against the alternative that the underlying distribution function is still continuous but not equal to F. This statistical problem is called testing for goodness-of-fit. It is natural to use

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \quad \text{or} \quad \sup_{t\in\mathbb{R}} n^{1/2}|F_n(t) - F(t)|$$

as a test statistic and to reject for large values. Under the null hypothesis the exact distribution is known and (1.33) and (1.34) give the asymptotic distribution. An asymptotically level α test ($\alpha \in (0,1)$) is obtained when we reject if

$$\sup_{t\in\mathbb{R}} n^{1/2}|F_n(t) - F(t)| \ge H^{-1}(1-\alpha)\ .$$

An asymptotically $1-\alpha$ confidence band is given by the set of all continuous F such that

$$F_n(t) - n^{-1/2}H^{-1}(1-\alpha) < F(t) < F_n(t) + n^{-1/2}H^{-1}(1-\alpha), \quad t\in\mathbb{R}\ .$$

The test is called the Kolmogorov-Smirnov test.

Other test statistics are also used. The Cramér-von Mises test e.g. is based on the statistic

$$(1.35)\qquad n\int_{-\infty}^{\infty} (F_n(t) - F(t))^2 dF(t)\ .$$

It is easy to see that this statistic has the same distribution as $\int_0^1 U_n^2(t)dt$; the asymptotic distribution is of χ^2-type with a countably infinite number of degrees of freedom.

We conclude this chapter with a (less straightforward) application to the two-sample problem. Suppose that $X_1, \ldots, X_{n_1}$ is a random sample of size n_1 from a continuous distribution F_1 on $\mathbb{R}$ and that $X_{n_1+1}, \ldots, X_{n_1+n_2}$ is a random sample of size n_2 from a continuous distribution function F_2 on $\mathbb{R}$. The problem of testing the null hypothesis that $F_1 = F_2$ against the alternative of inequality is called the nonparametric two-sample problem. Let $n = n_1 + n_2$ and $R_{1n}, \ldots, R_{nn}$ be the ranks of the pooled sample elements. It is intuitively clear that

$$(1.36) \qquad \sum_{i=1}^{n_1} R_{in} \quad \text{or} \quad \sum_{i=1}^{n_1} \frac{R_{in}}{n+1}$$

is a useful test statistic and that the null hypothesis should be rejected for relatively small and large values. Let us from now on assume that the null hypothesis is true and that F is the common distribution function. In this case the pooled sample is just a sample of size n from F and (1.21) implies that the statistic in (1.36) has one and the same distribution for each continuous F. The critical values may be obtained asymptotically from the relation

$$(1.37) \qquad \mathbb{P}\left(\left(\frac{12n}{n_1 n_2}\right)^{1/2} \sum_{i=1}^{n_1} \left(\frac{R_{in}}{n+1} - \tfrac{1}{2}\right) \le x \right) \to \Phi(x), \quad \text{as} \quad n_1 \wedge n_2 \to \infty, \ x \in \mathbb{R} \ .$$

This is Wilcoxon's two-sample test.

Let us briefly sketch a proof of a slightly more general version of (1.37). Suppose that for each n the real numbers $c_{1n}, ..., c_{nn}$ satisfy

$$(1.38) \qquad \sum_{i=1}^{n} c_{in} = 0, \ \sum_{i=1}^{n} c_{in}^2 = 1, \ \max_{1 \le i \le n} c_{in} \to 0 \quad (n \to \infty) \ ,$$

and introduce

$$W_n = (12)^{1/2} \sum_{i=1}^{n} c_{in} \frac{R_{in}}{n+1}, \ S_n = (12)^{1/2} \sum_{i=1}^{n} c_{in} \xi_i \ ,$$

where the $\xi_i = F(X_i)$ are independent with common $UN(0,1)$-distribution under the null hypothesis. It will be shown that

$$(1.39) \qquad \mathbb{P}(W_n \le x) \to \Phi(x) \quad (n \to \infty), \ x \in \mathbb{R} \ .$$

First note that the central limit theorem applies to S_n and yields

$$(1.40) \qquad \mathbb{P}(S_n \le x) \to \Phi(x) \ (n \to \infty), \quad x \in \mathbb{R} \ .$$

Next we will show that W_n behaves asymptotically like S_n by demonstrating that

$$(1.41) \qquad \mathbb{E}(W_n - S_n)^2 \to 0, \quad \text{as} \quad n \to \infty \ ,$$

so that $\mathbb{P}(|W_n - S_n| \ge \varepsilon) \to 0$, as $n \to \infty$, for each $\varepsilon > 0$. Hence $W_n = S_n + o_{\mathbb{P}}(1)$, as $n \to \infty$, and consequently W_n has the same limiting distribution as S_n.

To see (1.41) observe that the random vector

$$(V_1, ..., V_n) = \left(\frac{R_{1n}}{n+1} - \xi_1, ..., \frac{R_{nn}}{n+1} - \xi_n \right)$$

has exchangeable components. This entails in particular that $\mathcal{L}(V_i) = \mathcal{L}(V_1)$ for all i and that $\mathcal{L}(V_i V_j) = \mathcal{L}(V_1 V_2)$ for all $i \ne j$. We now find

$$\tfrac{1}{12}\mathbb{E}(W_n - S_n)^2 = \mathbb{E}\left(\sum_{i=1}^{n} c_{in}V_i\right)^2 = \sum\sum_{i\neq j} c_{in}c_{jn}\mathbb{E}V_iV_j + \sum c_{in}^2\mathbb{E}V_i^2$$

$$= \left(\sum\sum_{i\neq j} c_{in}c_{jn}\right)\mathbb{E}V_1V_2 + \left(\sum c_{in}^2\right)\mathbb{E}V_1^2$$

$$= \left(\sum c_{in}\right)^2 \mathbb{E}V_1V_2 + \left(\sum c_{in}^2\right)(\mathbb{E}V_1^2 - \mathbb{E}V_1V_2) \leq 2\mathbb{E}V_1^2 \ .$$

To complete the proof of (1.41) note that

$$\mathbb{E}V_1^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\left(\frac{R_{1n}}{n+1} - \xi_1\right)^2 |R_{1n} = i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{i}{n+1} - \xi_{i:n}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} \text{Var } \xi_{i:n} = \sum_{i=1}^{n}\frac{i(n+1-i)}{n(n+1)^2(n+2)} \to 0 \ (n\to\infty) \ .$$

Here we use (1.25) in the special case that there $F = I$, see (1.31).

To see that (1.37) is implied by (1.39) just observe that

$$\left(\frac{12n}{n_1n_2}\right)^{1/2}\sum_{i=1}^{n_1}\left(\frac{R_{in}}{n+1} - \tfrac{1}{2}\right) = (12)^{1/2}\sum_{i=1}^{n} c_{in}\frac{R_{in}}{n+1} \ ,$$

with $c_{in} = (n_2/(nn_1))^{1/2}$ for $i = 1, ..., n_1$ and $c_{in} = -(n_1/(nn_2))^{1/2}$ for $i = n_1+1, ..., n_1+n_2$. These c_{in} satisfy (1.38).

PROBLEMS

1.1 It can be shown that the class of lower quadrants in $\mathbb{R}^2$ picks out at most $1 + \frac{1}{2}N + \frac{1}{2}N^2$ subsets from a set of N points in $\mathbb{R}^2$ (see the definition of VC). Find a set of N points for which this bound is achieved.

1.2 Show that the class of closed convex sets in $\mathbb{R}^2$ does not satisfy VC.

1.3 Let P be the standard normal law on $\mathbb{R}$ and let $\mathcal{A}$ contain exactly those subsets of $\mathbb{R}$ which can be written as a union of finitely many closed intervals. Does (1.15) hold for this class $\mathcal{A}$?

1.4 Check that the central limit theorem is applicable in (1.40).

1.5 Calculate Var $\xi_{i:n}$ $(i = 1, ..., n)$, when $\Gamma = I$, with the aid of (1.25).

2. FUNDAMENTAL TOOLS

In this chapter we establish what will be called the "fundamental inequality", a probability inequality concerning "local" fluctuations of the empirical process. In the subsequent chapters it will be shown to be an extremely useful and elegant tool for proving various results in empirical process theory. We will once more extend our setup, now to the multidimensional case; similarly as in chapter 1 we assume for convenience that the random vectors take values in $(0,1)^d, d \in \mathbb{N}$. If we need to display the coordinates of $t \in \mathbb{R}^d$, we write $t = \langle t_1, ..., t_d\rangle = \langle t_j \rangle$ and if $t_j = u$ for all $1 \le j \le d$ we simply write $\langle u \rangle$. For $s, t \in \mathbb{R}^d$ we define $s \le t$ to mean that $s_j \le t_j$ for all $1 \le j \le d$; $s < t$ means that $s_j < t_j$ for all $1 \le j \le d$.

To establish our goal we need to derive a number of auxiliary results which are of independent interest. We begin with two inequalities for partial sums of independent random variables indexed by elements of $\mathbb{N}^d$. Let $d \in \mathbb{N}$ and fix $v \in \mathbb{N}^d$. For $t \in \mathbb{N}^d$ with $t \le v$ let Y_t be a random variable with mean 0 and finite variance σ_t^2 and assume that these random variables are independent. For arbitrary $s, u \in \mathbb{N}^d$ with $s \le u \le v$ define partial sums by

$$S_u = \sum_{t \le u} Y_t, \; S_{s,u} = \sum_{s \le t \le u} Y_t \tag{2.1}$$

and write $\sigma^2 = \text{Var } S_v = \sum\limits_{t \le v} \sigma_t^2$.

THEOREM 2.1. For all $\lambda \in \mathbb{R}$ we have

$$\mathbb{P}(\max_{u \le v} S_u \ge \lambda) \le 2^d \mathbb{P}(S_v \ge \lambda - d\sqrt{2\sigma^2}) \; . \tag{2.2}$$

PROOF. Let $A = \{\max\limits_{u \le v} S_u \ge \lambda\}$ and $B = \{\max\limits_{u \le v; u_d = v_d} S_u \ge \lambda - \sqrt{2\sigma^2}\}$. We first prove that

$$\mathbb{P}A \le 2\mathbb{P}B \; . \tag{2.3}$$

Let $C = \left\{ k \in \mathbb{N} : k \le \prod\limits_{j=1}^{d} v_j \right\}$, $D = \{t \in \mathbb{N}^d : t_j \le v_j, \text{ for all } 1 \le j \le d\}$ and $g : C \to D$ a fixed bijective function with the property: $k < l$ implies $(g(k))_d \le (g(l))_d$. Write $\widetilde{g(k)}$ $= \langle (g(k))_1, ..., (g(k))_{d-1}, v_d \rangle, E_k = \{S_{g(k)} \ge \lambda\}$ and $A_k = E_k E_{k-1}^c ... E_1^c$ $(k \in C)$. Note that the A_k are pairwise disjoint and that $A = \bigcup\limits_{k \in C} A_k$.

Now we have by Chebyshev's inequality and the independence of A_k and $\{S_{\widetilde{g(k)}} - S_{g(k)}$ $\ge -\sqrt{2\sigma^2}\}$ that

$$\mathbb{P}A_k \le 2\mathbb{P}A_k \mathbb{P}(S_{\widetilde{g(k)}} - S_{g(k)} \ge -\sqrt{2\sigma^2})$$

$$= 2\mathbb{P}(A_k \cap \{S_{\widetilde{g(k)}} - S_{g(k)} \ge -\sqrt{2\sigma^2}\}) \le 2\mathbb{P}A_k B \; .$$

Hence

$$\mathbb{P}A = \sum_{k\in C} \mathbb{P}A_k \le 2\sum_{k\in C} \mathbb{P}A_k B = 2\mathbb{P}AB \le 2\mathbb{P}B \ ,$$

which proves (2.3) . It is now easily seen that the proof of Theorem 2.1 can be completed by an induction argument. □

THEOREM 2.2. For all $\lambda \in \mathbb{R}$ we have

$$\mathbb{P}(\max_{s\le u\le v} S_{s,u} \ge \lambda) \le 2^{2d}\mathbb{P}(S_v \ge \lambda - 2d\sqrt{2\sigma^2}) \ . \tag{2.4}$$

PROOF. From the Y_t, $t \in \mathbb{N}^d$, we derive a collection of random variables indexed by elements of $\mathbb{N}^{2d}$, which has the properties needed for application of Theorem 2.1. To distinguish the elements of $\mathbb{N}^{2d}$ and $\mathbb{N}^d$ we denote the elements of $\mathbb{N}^{2d}$ by $\bar{s}, \bar{t}$ etc. We consider $\bar{t} \le \bar{w} = \langle v_1, ..., v_d, v_1, ..., v_d\rangle$. For $\bar{t}$ with $\bar{t}_j + \bar{t}_{d+j} = v_j + 1$, for all $1 \le j \le d$, we define $Y_{\bar{t}} = Y_{\langle \bar{t}_1, \bar{t}_2, ..., \bar{t}_d\rangle}$; $Y_{\bar{t}} = 0$ otherwise. Now it is easily seen that $S_{s,u} = S_{\langle u_1,...,u_d, v_1+1-s_1,...,v_d+1-s_d\rangle}$, which combined with an application of Theorem 2.1 (with d replaced by $2d$) yields

$$\mathbb{P}(\max_{s\le u\le v} S_{s,u} \ge \lambda) \le \mathbb{P}(\max_{\bar{u}\le\bar{w}} S_{\bar{u}} \ge \lambda)$$

$$\le 2^{2d}\mathbb{P}(S_{\bar{w}} \ge \lambda - 2d\sqrt{2\sigma^2}) = 2^{2d}\mathbb{P}(S_v \ge \lambda - 2d\sqrt{2\sigma^2}) \ .$$ □

Now we return to the (reduced) empirical process as defined in (1.29), but based on random vectors instead of random variables. To be more precise, our sequence $\xi_1, \xi_2, ...$ is now a sequence of independent random vectors taking values in $(0,1)^d, d \in \mathbb{N}$ with common law P and distribution function Γ. As indexing sets we will either use quadrants of the form $[0, t_1] \times ... \times [0, t_d], t = \langle t_j\rangle \in [0,1]^d$, which are identified with the points t, or half-open rectangles of the form $(s_1, t_1] \times ...(s_d, t_d], s = \langle s_j\rangle < t = \langle t_j\rangle, s, t \in [0,1]^d$, denoted by $R(s,t)$ or simply R. The class of all these half-open rectangles is denoted by $\mathcal{R}$.

The empirical distribution function can be defined again as in (1.28):

$$\Gamma_n(t) = \frac{1}{n}\#\{1 \le i \le n : \xi_i \le t\}, t \in [0,1]^d \ , \tag{2.5}$$

and $P_n(R)$, the empirical measure of $R \in \mathcal{R}$, is defined in (1.13). The multivariate empirical process, indexed by points, is now given by

$$U_{n,\Gamma}(t) = n^{1/2}(\Gamma_n(t) - \Gamma(t)), t \in [0,1]^d \ , \tag{2.6}$$

and the multivariate empirical process, indexed by rectangles, by

$$U_{n,\Gamma}(R) = n^{1/2}(P_n(R) - P(R)), R \in \mathcal{R} \ . \tag{2.7}$$

Note that the process in (2.6) is tied down at the lower boundary $L_d = \{t \in [0,1]^d : t_1 \wedge ... \wedge t_d = 0\}$ of $[0,1]^d$ and at its upper vertex $\langle 1\rangle$, meaning that $U_{n,\Gamma}(t) = 0$ if $t \in L_d$ or $t = \langle 1\rangle$. This property follows from the fact that both Γ_n and Γ are 0 on L_d and 1 at $\langle 1\rangle$.

Our next step is relating $U_{n,\Gamma}$ to the process Z_n defined by

$$(2.8) \qquad Z_n(t) = n^{-1/2}(N_n(t) - n\Gamma(t)), t \in [0,1]^d ,$$

where N_n is a Poisson process on $[0,1]^d$ with $\mathbb{E}N_n(t) = n\Gamma(t)$ for all $t \in [0,1]^d$. We have that conditionally on $N_n(\langle 1\rangle) = n$ the processes $U_{n,\Gamma}$ and Z_n are equal in law. Rather than proving this we only mention that a version of N_n can be constructed as follows:

$$N_n(t) = \#\{1 \le i \le \nu_n : \xi_i \le t\}, t \in [0,1]^d ,$$

where ν_n is a Poisson(n) random variable independent of the ξ_i. For $R \in \mathcal{R}, N_n(R)$ and $Z_n(R)$ will have their obvious meaning. We now give a special version of a useful inequality based on the above relation.

THEOREM 2.3. For any $R \in \mathcal{R}$ with $P(R) \le \frac{1}{2}$, all $\lambda \in \mathbb{R}$ and either choice of sign we have

$$(2.9) \qquad \mathbb{P}(\sup_{S\subset R} \pm U_{n,\Gamma}(S) \ge \lambda) \le 2\mathbb{P}(\sup_{S\subset R} \pm Z_n(S) \ge \lambda) ,$$

where $S \in \mathcal{R}$.

PROOF. Let $A_1 = \{\sup_{S\subset R} \pm U_{n,\Gamma}(S) \ge \lambda\}$ and $A_2 = \{\sup_{S\subset R} \pm Z_n(S) \ge \lambda\}$. Notice that $N_n(R)$ and $N_n(R^c)$ are independent. It follows that

$$(2.10) \qquad \mathbb{P}A_1 = \mathbb{P}(A_2 \mid N_n(\langle 1\rangle) = n)$$

$$= \mathbb{P}(A_2 \cap \{N_n(\langle 1\rangle) = n\})/\mathbb{P}(N_n(\langle 1\rangle) = n)$$

$$= \sum_{k=0}^{n} \mathbb{P}(A_2 \cap \{N_n(R) = k\} \cap \{N_n(R^c) = n-k\})/\mathbb{P}(N_n(\langle 1\rangle) = n)$$

$$= \sum_{k=0}^{n} (\mathbb{P}(N_n(R^c) = n-k)/\mathbb{P}(N_n(\langle 1\rangle) = n))\mathbb{P}(A_2 \cap \{N_n(R) = k\})$$

$$\le \frac{\mathbb{P}(N_n(R^c) = [nP(R^c)])}{\mathbb{P}(N_n(\langle 1\rangle) = n)} \cdot \mathbb{P}A_2 ,$$

because $\mathbb{P}(Z = k) \le \mathbb{P}(Z = [\tau])$ if Z is a Poisson (τ) random variable ($[\tau]$ = largest integer $\le \tau$). An application of Stirling's formula completes the proof. □

Note that the restriction $P(R) \le \frac{1}{2}$ is used only at the end of the proof. It is easy to see that the number $\frac{1}{2}$ in the statement of Theorem 2.3 may be replaced by any number $a < 1$, provided only that we change the number 2 in (2.9) into a number depending on a. We will, however, only need this inequality for rectangles R with $P(R) \le \frac{1}{2}$.

By invoking the independent increment property of Z_n and Theorem 2.2, we can now get rid of the supremum in the right hand side of (2.9).

THEOREM 2.4. For any $R \in \mathcal{R}$, all $\lambda \in \mathbb{R}$ and either choice of sign we have

$$(2.11) \qquad I\!P(\sup_{S\subset R} \pm Z_n(S) \geq \lambda) \leq 2^{2d} I\!P(\pm Z_n(R) \geq \lambda - 2d\sqrt{2P(R)}) \ ,$$

where $S \in \mathcal{R}$.

PROOF. For $k \in I\!N$ let $\mathcal{P}_k \subset \mathcal{R}$ be the partition of R consisting of 2^{kd} rectangles of equal Lebesgue measure with the property that the length of the i-th side of an element of $\mathcal{P}_k$ is 2^{-k} times the length of the i-th side of R. Observe now that the collection $\{\pm Z_n(E)\}_{E\in\mathcal{P}_k}$ can be seen as a collection of random variables $\{Y_t\}_{t\leq v}$ (with the obvious indexation) to which Theorem 2.2 may be applied with $\sigma^2 = P(R)$.

We also have

$$(2.12) \qquad \lim_{k\to\infty} \sup{}^{*} \pm Z_n(S) = \sup_{S\subset R} \pm Z_n(S) \ \text{ a.s. } ,$$

where $\sup^*$ denotes the supremum over all $S \subset R, S \in \mathcal{R}$, such that S is a union of elements of $\mathcal{P}_k$. Combination of the observation just above (2.12) and (2.12) itself yields for all $\lambda \in I\!R$

$$(2.13) \qquad I\!P(\sup_{S\subset R} \pm Z_n(S) > \lambda) \leq 2^{2d} I\!P(\pm Z_n(R) \geq \lambda - 2d\sqrt{2P(R)}) \ .$$

Since (2.13) holds for all $\lambda \in I\!R$, (2.11) easily follows. $\square$

The proof of the aforementioned fundamental inequality requires sharp bounds for the tail probabilities of a Poisson random variable. For this purpose we define $\psi : [-1,\infty) \to [0,\infty)$ by

$$(2.14) \qquad \begin{cases} \psi(\lambda) = 2\lambda^{-2} \int_0^\lambda \log\ (1+\sigma)d\sigma = 2\lambda^{-2}\{(1+\lambda)\log(1+\lambda) - \lambda\} \ , \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \lambda \in (-1,0)\cup(0,\infty) \ ; \\ \psi(0) = 1; \psi(-1) = 2 \ . \end{cases}$$

This function has the following properties:

$$(2.15) \qquad \begin{cases} \psi \text{ is decreasing and continuous} \ , \\ \psi(\lambda) \sim 2 \ \log\ \lambda/\lambda \text{ as } \lambda \to \infty \ , \\ \psi(\lambda) \geq 1/(1+\lambda/3), \ \lambda \in [-1,\infty) \ , \\ I\psi \text{ is increasing} \ . \end{cases}$$

Let $\tau > 0$, then we have for a Poisson (τ) random variable Z and either choice of sign

$$(2.16) \qquad I\!P(\pm(Z-\tau) \geq \lambda) \leq \exp\left(\frac{-\lambda^2}{2\tau}\psi\left(\frac{\pm\lambda}{\tau}\right)\right), \ \lambda \geq 0 \ .$$

(For the minus part of (2.16) we also assume $\lambda \leq \tau$ because of the definition of ψ.) These bounds follow easily from computing the moment generating function, applying the Markov inequality and then minimizing the thus obtained upper bound.

THEOREM 2.5 (fundamental inequality). Let $n \in I\!N$ and $R \in \mathcal{R}$ with $0 < P(R) \leq \frac{1}{2}$. Then we have for any $\varepsilon \in (0,1)$

$$(2.17) \qquad I\!P(\sup_{S \subset R} U_{n,\Gamma}(S) \geq \lambda) \leq C^+ \exp\left(\frac{-(1-\varepsilon)\lambda^2}{2P(R)}\psi\left(\frac{\lambda}{n^{1/2}P(R)}\right)\right), \lambda \geq 0 ,$$

$$(2.18) \qquad I\!P(\sup_{S \subset R} -U_{n,\Gamma}(S) \geq \lambda) \leq C^- \exp\left(\frac{-(1-\varepsilon)\lambda^2}{2P(R)}\psi\left(\frac{-(1-\varepsilon)\lambda}{n^{1/2}P(R)}\right)\right), 0 \leq \lambda \leq n^{1/2}P(R) ,$$

$$(2.19) \qquad I\!P(\sup_{S \subset R} |U_{n,\Gamma}(S)| \geq \lambda) \leq C \exp\left(\frac{-(1-\varepsilon)\lambda^2}{2P(R)}\psi\left(\frac{\lambda}{n^{1/2}P(R)}\right)\right), \lambda \geq 0 ,$$

where $S \in \mathcal{R}$ and $C^+, C^-, C \in (0, \infty)$ only depend on d and ε.

PROOF. Let us first consider (2.17) and (2.18). Combination of Theorem 2.3, Theorem 2.4 and (2.16) immediately yields for $\lambda \geq 4d\sqrt{2P(R)}/\varepsilon$ and either choice of sign

$$(2.20) \qquad I\!P(\sup_{S \subset R} \pm U_{n,\Gamma}(S) \geq \lambda) \leq 2^{2d+1} \exp\left(\frac{-(1-\varepsilon)\lambda^2}{2P(R)}\psi\left(\frac{\pm\lambda(1-\frac{1}{2}\varepsilon)}{n^{1/2}P(R)}\right)\right) .$$

(Again for the minus part we also assume $\lambda \leq n^{1/2}P(R)$.) For the other values of λ for which we need to prove (2.17) and (2.18) we can, of course, bound the probabilities by 1. From this remark and (2.20) we immediately obtain the bounds given in (2.17) and (2.18). Finally, observe that (2.19) follows from (2.17) and (2.18). □

Theorem 2.5 is the main result of this chapter. It is a rather sharp inequality as we will see later on. Often it will be applied to

$$(2.21) \qquad I\!P(\sup_{s \leq t} |U_{n,\Gamma}(s)| \geq \lambda) ,$$

which is a special case of the left hand side of (2.19), but it is also a pleasant, though natural consequence of our approach that the plus and minus parts can be kept separated. Theorem 2.5 remains true for rectangles S and R that are both open or both closed. A proof follows from a slight modification of the proof of Theorem 2.4.

This chapter will be concluded with two applications of the fundamental inequality. The first one yields another fundamental result for multivariate empirical processes, namely a sharp probability inequality for the global fluctuations of the empirical process. Let $X_1, ..., X_n$ be independent random vectors taking values in $I\!R^d$, with common but arbitrary distribution function F (and law P) and write $U_{n,F}$ for the corresponding multivariate empirical process, cf. (2.6) and (2.7).

THEOREM 2.6. Let $d \in I\!N$ and $n \in I\!N$. For all $\varepsilon \in (0,1)$ we have

$$(2.22) \qquad I\!\!P\Big(\sup_{t\in I\!\!R^d} |U_{n,F}(t)| \geq \lambda\Big) \leq C\exp(-2(1-\varepsilon)\lambda^2),\ \lambda \geq 0\ ,$$

where $C \in (0,\infty)$ does not depend on F, n or λ, but only on d and ε; when $d=1$, ε may be taken 0 and C equal to 2.

PROOF. We only prove the part where $\varepsilon \in (0,1)$. Without loss of generality we will assume that the X_i take values in $(0,1)^d$ and that F has uniform marginals (and hence is continuous); we will not prove that we indeed may do so.

For $1/\delta \in I\!\!N$, let $\mathcal{T}_\delta$ be the collection of vectors

$$\{\langle k_1\delta, k_2\delta, ..., k_d\delta\rangle : k_j \text{ integer},\ 0 \leq k_j \leq 1/\delta \text{ for all } 1 \leq j \leq d\}\ .$$

Let us also introduce for $1 \leq j \leq d$ and $1 \leq k \leq 1/\delta$ ($k \in I\!\!N$) the slices

$$C_{j,k} = (0,1]^{j-1} \times ((k-1)\delta, k\delta] \times (0,1]^{d-j}\ .$$

For arbitrary $t \in [0,1]^d$ define $\bar{t}, \bar{\bar{t}} \in \mathcal{T}_\delta$ by $0 < t_j - \bar{t}_j < \delta$ for all $1 \leq j \leq d$ and $0 \leq \bar{\bar{t}}_j - t_j < \delta$ for all $1 \leq j \leq d$. Note that $\bar{t} \leq t \leq \bar{\bar{t}}$. Now observe that we trivially have (in the obvious notation)

$$(2.23) \qquad U_{n,F}(t) = U_{n,F}(\bar{\bar{t}}) - U_{n,F}\left(\prod_{j=1}^{d}[0,\bar{\bar{t}}_j] \setminus \prod_{j=1}^{d}[0,t_j]\right)$$

and

$$(2.24) \qquad -U_{n,F}(t) = -U_{n,F}(\bar{t}) - U_{n,F}\left(\prod_{j=1}^{d}[0,t_j] \setminus \prod_{j=1}^{d}[0,\bar{t}_j]\right)\ .$$

Combining (2.23) and (2.24) with

$$|U_{n,F}(t)| = \max(U_{n,F}(t), -U_{n,F}(t))\ ,$$

we easily obtain that

$$\sup_{t\in[0,1]^d} |U_{n,F}(t)| \leq \max_{t\in\mathcal{T}_\delta} |U_{n,F}(t)| + \max_{k_1,\ldots,k_d} \sum_{j=1}^{d} \sup_{\substack{R\subset C_{j,k_j}\\ R\in\mathcal{R}}} -U_{n,F}(R)\ ,$$

and hence

$$\sup_{t\in[0,1]^d} |U_{n,F}(t)| \leq \max_{t\in\mathcal{T}_\delta} |U_{n,F}(t)| + \sum_{j=1}^{d} \max_{0\leq k\leq 1/\delta} \sup_{\substack{R\subset C_{j,k}\\ R\in\mathcal{R}}} -U_{n,F}(R)\ .$$

This yields for $0 < \varepsilon < 1$,

$$\text{(2.25)} \qquad I\!\!P(\sup_{t\in[0,1]} |U_{n,F}(t)| \geq \lambda) \leq I\!\!P(\max_{t\in\mathcal{T}_\delta} |U_{n,F}(t)| \geq (1-\varepsilon)\lambda)$$

$$+ I\!\!P\left(\sum_{j=1}^{d} \max_{0\leq k\leq 1/\delta} \sup_{\substack{R\subset C_{j,k}\\ R\in\mathcal{R}}} - U_{n,F}(R) \geq \varepsilon\lambda\right) .$$

By the inequality in (2.18) we have, since $P(C_{j,k}) = \delta$,

$$\text{(2.26)} \qquad I\!\!P\left(\sum_{j=1}^{d} \max_{0\leq k\leq 1/\delta} \sup_{\substack{R\subset C_{j,k}\\ R\in\mathcal{R}}} - U_{n,F}(R)\right)$$

$$\leq \sum_{j=1}^{d}\sum_{k=0}^{1/\delta} I\!\!P\left(\sup_{\substack{R\subset C_{j,k}\\ R\in\mathcal{R}}} - U_{n,F}(R) \geq \varepsilon\lambda/d\right)$$

$$\leq C_1 \exp(-\varepsilon^2\lambda^2/(4d^2\delta)) ,$$

since $\psi(-\frac{1}{2}\varepsilon\lambda/(dn^{1/2}\delta)) \geq 1$, where $C_1 = C_1(d,\delta) < \infty$. Moreover, employing Hoeffding's inequality in (1.11) we have that

$$\text{(2.27)} \qquad I\!\!P(\sup_{t\in\mathcal{T}_\delta} |U_{n,F}(t)| \geq (1-\varepsilon)\lambda) \leq \sum_{t\in\mathcal{T}_\delta} I\!\!P(|U_{n,F}(t)| \geq (1-\varepsilon)\lambda)$$

$$\leq 2(1/\delta+1)^d \exp(-2(1-\varepsilon)^2\lambda^2) .$$

Choosing $1/\delta = [8d^2/\varepsilon^2] + 1$, (2.25)-(2.27) yield as an upper bound for the left hand side of (2.25),

$$C_2 \exp(-2(1-\varepsilon)^2\lambda^2) ,$$

where $C_2 = C_2(d,\varepsilon) < \infty$. Relabeling $(1-\varepsilon)^2$ by $(1-\varepsilon)$ gives the right hand side of (2.22).□

The second application which we will present deals with density estimation. Let now $X_1, ..., X_n$ be a random sample from a distribution function F (pertaining to a law P) on $I\!\!R$, with uniformly continuous derivative $F' = f$. Set $M = \max_{t\in I\!\!R} f(t)$. The problem is to estimate the unknown density f. We will restrict ourselves to so-called naïve estimators which are very directly related to empirical measures. For each $t \in I\!\!R$ consider the half-open intervals (one might as well take the closed or open intervals)

$$(t_n^-, t_n^+] = (t - \tfrac{1}{2}l_n, t + \tfrac{1}{2}l_n] \ ,$$

where $\{l_n\}_{n=1}^{\infty}$ is a sequence of positive real numbers converging to zero. Because $f(t) = F'(t) = \lim_{n\to\infty} P((t_n^-, t_n^+])/l_n$, a natural estimator for $f(t)$ is

$$\hat{f}_n(t) = P_n((t_n^-, t_n^+])/l_n, t \in I\!R \ .$$

It is easy to see that $\hat{f}_n$ is a density indeed, since $\hat{f}_n \geq 0$ and

$$\int_{-\infty}^{\infty} \hat{f}_n(t)dt = \frac{1}{l_n}\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} 1_{(t_n^-, t_n^+]}(X_i)dt = \frac{1}{l_n}\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} 1_{[X_{in}^-, X_{in}^+)}(t)\ dt = 1 \ .$$

Observe, however, that $\hat{f}_n$ is not an unbiased estimator of f; as a matter of fact

$$I\!E\hat{f}_n(t) = P(t_n^-, t_n^+])/l_n = f_n(t), \ t \in I\!R \ .$$

On the other hand we do have that there exists a $c \in (0,\infty)$ such that

$$\text{(2.28)} \qquad \limsup_{n\to\infty} \left(\frac{nl_n}{\log n}\right)^{1/2} \sup_{t\in I\!R} |\hat{f}_n(t) - f_n(t)| \leq c \text{ a.s. } ,$$

provided that $nl_n/\log n \to \infty$, as $n \to \infty$;

$$\text{(2.29)} \qquad \sup_{t\in I\!R} |f_n(t) - f(t)| \to 0, \text{ as } n \to \infty \ .$$

Combination of (2.28) and (2.29) yields

$$\text{(2.30)} \qquad \sup_{t\in I\!R} |\hat{f}_n(t) - f(t)| \to 0 \text{ a.s. } (n \to \infty) \ .$$

The rate of convergence in (2.29) - and hence in (2.30) - may be improved by imposing further smoothness conditions on f.

Let us now prove (2.28). To avoid inessential technicalities we will assume that F is strictly increasing on $I\!R$. Write

$$F(t_n^{\pm}) = u_n^{\pm}, \ F(t) = u \ ,$$

U_n for the empirical process based on the $F(X_i) = \xi_i$ (as before), and note that

$$(u_n^-, u_n^+] \subset (0,1), \ |u_n^+ - u_n^-| \leq Ml_n \ \text{ for all } u \in (0,1) \ .$$

Choosing $m_n = [1/Ml_n)] \in I\!N$, let us partition $(0,1]$ into the subintervals

$$(2.31)\qquad \left(\frac{j-1}{m_n}, \frac{j}{m_n}\right] = (u_{j-1,n}, u_{j,n}],\ j \in \{1, ..., m_n\}\ .$$

Because the length of the intervals in (2.31) is at least Ml_n, any interval $(u_n^-, u_n^+]$ intersects at most two adjacent subintervals in (2.31).

These considerations entail that

$$\left(\frac{nl_n}{\log n}\right)^{1/2} \sup_{t\in\mathbb{R}} |\hat{f}_n(t) - f_n(t)| = (l_n \log n)^{-1/2} \sup_{u\in(0,1)} |U_n((u_n^-, u_n^+])|$$

$$\leq 2(l_n \log n)^{-1/2} \max_{j=1,\dots,m_n} \sup_{u_{j-1,n}\leq v<w\leq u_{j,n}} |U_n((v,w])|\ .$$

Application of (2.19) (with $\varepsilon = \frac{1}{2}$) yields

$$p_n = \mathbb{P}\left(\left(\frac{nl_n}{\log n}\right)^{1/2} \sup_{t\in\mathbb{R}} |\hat{f}_n(t) - f_n(t)| \geq c\right)$$

$$\leq \sum_{j=1}^{m_n} \mathbb{P}(\sup_{u_{j-1,n}\leq v<w\leq u_{j,n}} |U_n((v,w])| \geq \tfrac{1}{2}c(l_n \log n)^{1/2})$$

$$\leq m_n C \exp\left(\frac{-c^2 l_n \log n}{16(1/m_n)}\psi\left(\frac{c(l_n \log n)^{1/2}}{2n^{1/2}(1/m_n)}\right)\right)\ .$$

It is immediate from the assumptions on l_n and the definition of m_n that for n large

$$m_n \leq n \text{ and } m_n l_n \geq \frac{1}{2M}\ ;$$

$$\psi\left(\frac{c(l_n \log n)^{1/2}}{2n^{1/2}(1/m_n)}\right) \to \psi(0) = 1\ (n\to\infty), \text{ for all } c > 0\ .$$

Hence, choosing $c = 10M^{1/2}$ it follows that for n large

$$p_n \leq Cn \exp(-3\log n) = Cn^{-2}\ .$$

Consequently $\sum_{n=1}^{\infty} p_n < \infty$ which implies (2.28) according to the Borel-Cantelli lemma.

Density estimation is e.g. used in nonparametric pattern recognition (discriminant analysis) when the densities of the feature vectors are unknown and are to be estimated from training samples.

PROBLEMS

2.1 Prove that $\mathbb{P}(Z = k) \leq \mathbb{P}(Z = [\tau])$ if Z is a Poisson (τ) random variable (see the proof of Theorem 2.3).

2.2 Prove that $\mathbb{P}(N_n(R^c) = [nP(R^c)])/\mathbb{P}(N_n(\langle 1\rangle) = n) \leq 2$ if $P(R) \leq \frac{1}{2}$, using Stirling's formula: for $m \in \mathbb{N}$

$$e^{\frac{1}{12m+1}} < \frac{m!e^m}{m^m\sqrt{2\pi m}} < e^{\frac{1}{12m}} .$$

(See the proof of Theorem 2.3.)

2.3 Prove (2.16).

2.4 Assume the density f (see the second application) has bounded first and second derivatives. Show that now (2.30) can be improved to

$$\limsup_{n\to\infty} \left(\frac{n}{\log n}\right)^{2/5} \sup_{t\in\mathbb{R}} |\hat{f}_n(t) - f(t)| \leq \tilde{c} \ \text{ a.s. },$$

for some $\tilde{c} \in (0, \infty)$.

3. THE UNIFORM EMPIRICAL PROCESS

In this chapter we consider strong and weak convergence of the empirical process. In between we derive some properties of the oscillation modulus of the empirical process, a concept which is of fundamental importance for, e.g., the weak convergence proofs. Here and in the subsequent chapters we exclusively deal with the one dimensional uniform empirical process U_n as defined in (1.32). Many of the obtained results can be easily transformed to the case where the random variables have a continuous distribution function F, cf. (1.30) and e.g. Theorem 1.4.

We begin with specifying the rate of almost sure convergence in Theorem 1.1. This leads to the celebrated law of the iterated logarithm. Let U_n be based on the first n of a sequence ξ_i, $i \in I\!N$, of independent $UN(0,1)$, random variables.

THEOREM 3.1 (law of the iterated logarithm). We have

$$\limsup_{n\to\infty} \sup_{t\in[0,1]} \frac{|U_n(t)|}{(2\log\log n)^{1/2}} = \tfrac{1}{2} \text{ a.s.} \tag{3.1}$$

The proof of this result is based on Theorem 2.6 in conjunction with a so-called maximal inequality which we will state without proof.

THEOREM 3.2. Let $\varepsilon \in (0,1)$ and write $n_k = [(1+\frac{1}{2}\varepsilon)^k]$, $k \in I\!N$. Then we have for all $k \in I\!N$ and for $\lambda > 2/\varepsilon^{1/2}$

$$I\!P\Big(\max_{n_k<n\le n_{k+1}} \sup_{t\in[0,1]} |U_n(t)| \ge \lambda\Big) \le 2I\!P\Big(\sup_{t\in[0,1]} |U_{n_{k+1}}(t)| \ge (1-\varepsilon)\lambda\Big) . \tag{3.2}$$

The proof of the inequality in (3.2) resembles a bit the proof of Theorem 2.1. Theorem 3.2 is needed to deal with the strong dependence between U_n and U_{n+1}, say.

PROOF OF THEOREM 3.1. Using that

$$\sup_{t\in[0,1]} |U_n(t)| \ge |U_n(\tfrac{1}{2})| ,$$

it easily follows from the ordinary law of the iterated logarithm for Bernoulli (or 0-1) random variables that almost surely the left hand side of (3.1) is bounded from below by $\frac{1}{2}$. So we focus on the upper bound.

It is easily seen from the Borel-Cantelli lemma that it suffices to show that for every $\varepsilon \in (0,1)$

$$\sum_{k=1}^{\infty} I\!P\Big(\max_{n_k<n\le n_{k+1}} \sup_{t\in[0,1]} |U_n(t)| \ge \tfrac{1}{2}\frac{1+\varepsilon}{1-\varepsilon}(2\log\log n_k)^{1/2}\Big) < \infty . \tag{3.3}$$

Using Theorem 3.2 and then Theorem 2.6 we see that

$$I\!P\left(\max_{n_k<n\le n_{k+1}} \sup_{t\in[0,1]} |U_n(t)| \ge \tfrac{1}{2}\frac{1+\varepsilon}{1-\varepsilon}(2\log\log n_k)^{1/2}\right) \tag{3.4}$$

$$\leq 2I\!P(\sup_{t\in[0,1]} |U_{n_{k+1}}(t)| \geq \tfrac{1}{2}(1+\varepsilon)(2\log\log n_k)^{1/2})$$

$$\leq 4\exp(-(1+\varepsilon)^2 \log\log n_k) = 4(\log n_k)^{-(1+\varepsilon)^2} \; .$$

From the definition of n_k and (3.4) we see that (3.3) holds. □

There even exists a so called functional law of the iterated logarithm. Let $D = D[0,1]$ be the linear space of all functions on [0,1] that are right-continuous and possess left-hand limits at each point $t \in [0,1]$ and endow D with the supremum-norm metric ρ according to

$$\rho(f,g) = \sup_{t\in[0,1]} |f(t) - g(t)|; \; f,g \in D \; . \tag{3.5}$$

The subclass

$$F = \{f \in D : \; f \text{ absolutely continuous, } f(0) = f(1) = 0 \text{ and } \int_0^1 (\dot{f}(t))^2 \leq 1\} \; ,$$

where $\dot{f}$ denotes the Lebesgue derivative of f, plays an important role in the formulation of that law. Let us more explicitely write $U_{n,\omega}, \omega \in \Omega$, instead of U_n. Without proof we mention the following result.

THEOREM 3.3 (functional law of the iterated logarithm). There exists an $\Omega_0 \in \mathcal{F}$ with $I\!P\Omega_0 = 1$, such that for each $\omega \in \Omega_0$ the sequence $\{U_{n,\omega}/(2\log\log n)^{1/2}\}_{n=2}^{\infty}$ is relatively compact in D with respect to ρ and the set of limit points is equal to F.

Note that Theorem 3.3 extends Theorem 3.1. In fact Theorem 3.3 entails that

$$\limsup_{n\to\infty} \sup_{t\in[0,1]} \frac{|U_n(t)|}{(2\log\log n)^{1/2}} = \sup_{f\in F} \sup_{t\in[0,1]} |f(t)| \text{ a.s.} \tag{3.6}$$

To find (3.1) it only remains to show that the number on the right in (3.6) is equal to $\frac{1}{2}$. The functional form of the law of the iterated logarithm, however, also informs us about the precise limiting behaviour of other functionals. As a second example of application of Theorem 3.3 let us mention that

$$\limsup_{n\to\infty} \frac{\int_0^1 U_n^2(t)dt}{2\log\log n} = 1/\pi^2 \text{ a.s. ,}$$

a result of interest in view of (1.35). It is proved by showing that

$$\sup_{f\in F} \int_0^1 f^2(t)dt = 1/\pi^2 \; .$$

To prove this one may use a calculus of variation argument.

Now we turn to the oscillation modulus of the empirical process, defined by

$$(3.7)\qquad \omega_n(\alpha) = \sup_{\substack{R\in\mathcal{R}\\ |R|\le\alpha}} |U_n(R)|,\ 0<\alpha\le 1\ .$$

Of course, here in the one dimensional situation $\mathcal{R}$ is the class of half-open intervals; for $R\in\mathcal{R}$ the length of R is denoted by $|R|$.

THEOREM 3.4. Let $\varepsilon\in(0,\frac{1}{2}]$ and $\alpha\in(0,1/4)$. Then we have

$$(3.8)\qquad I\!P(\omega_n(\alpha)\ge\lambda\alpha^{1/2})\le C\frac{1}{\alpha}\exp\left(\frac{-(1-\varepsilon)\lambda^2}{2}\psi\left(\frac{\lambda}{(n\alpha)^{1/2}}\right)\right),\lambda\ge 0\ ,$$

where $C=C(\varepsilon)\in(0,\infty)$.

PROOF. For $k\in I\!N$, to be specified later, let us consider the subintervals of [0,1]

$$\left(\frac{j\alpha}{k},\frac{(j+1)\alpha}{k}+\alpha\right]\cap[0,1]=(s_{j,k},t_{j,k}],\ j=0,1,...,\left[\frac{k}{\alpha}\right]\ .$$

Note that for each $R=(s,t]$ with $|R|\le\alpha$ there exists $j=j_s$ such that $s=s_{j_s,k}+\delta_s$, with $0\le\delta_s<\alpha/k$, and hence $R\subset(s_{j_s,k},t_{j_s,k}]$.

It follows from Theorem 2.5 that for each $0<\eta<1$ there exists $C(\eta)\in(0,\infty)$ such that

$$\begin{aligned}
&I\!P(\omega_n(\alpha)\ge\lambda\alpha^{1/2})\\
&\le I\!P\left(\max_{j=0,1,\ldots,[k/\alpha]}\ \sup_{s_{j,k}\le s<t\le t_{j,k}}|U_n((s,t])|\ge\lambda\alpha^{1/2}\right)\\
&\le\left(\left[\frac{k}{\alpha}\right]+1\right)C(\eta)\exp\left(\frac{-(1-\eta)\lambda^2\alpha}{2\alpha(1+1/k)}\psi\left(\frac{\lambda\alpha^{1/2}}{n^{1/2}\alpha(1+1/k)}\right)\right)\\
&\le\left(\left[\frac{k}{\alpha}\right]+1\right)C(\eta)\exp\left(\frac{-(1-\eta)\lambda^2}{2(1+1/k)}\psi\left(\frac{\lambda}{(n\alpha)^{1/2}}\right)\right)\ .
\end{aligned}$$

From this we obtain the upper bound in (3.8) by choosing $\eta=\frac{1}{2}\varepsilon$, $k=[2/\varepsilon]+1$, and noting that $(1-\frac{1}{2}\varepsilon)/(1+1/k)\ge 1-\varepsilon$. □

Note the similarities between this proof and the proof of relation (2.28). Here, however, we have been more precise by carefully selecting our subintervals, resulting in an inequality which is sharp enough for obtaining exact almost sure limits for the oscillation modulus. To establish this we need an analogue of Theorem 3.2, which will not be proved.

THEOREM 3.5. Let $\varepsilon\in(0,1)$, $\alpha\in(0,1]$ and write $n_k=[(1+\frac{1}{2}\varepsilon)^k]$, $k\in I\!N$. Then we have, for all $k\in I\!N$ and for $\lambda>2(\alpha/\varepsilon)^{1/2}$

$$(3.9) \qquad \mathbb{P}\left(\max_{n_k<n\le n_{k+1}} \omega_n(\alpha)\ge\lambda\right)\le 2\mathbb{P}(\omega_{n_{k+1}}(\alpha)\ge(1-\varepsilon)\lambda)\ .$$

We will now investigate the almost sure behavior of $\omega_n(\alpha)$ choosing $\alpha=a_n\in(0,1]$ depending on n. The almost sure behavior of $\omega_n(a_n)$ can be established for sequences $\{a_n\}_{n=1}^{\infty}$ of five different orders of magnitude that essentially constitute the complete spectrum of possiblities. Here we will present only one these five cases, namely the in various respects most natural and most important case.

THEOREM 3.6. Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of number in $(0,1]$ such that $a_n\downarrow 0$, $na_n\uparrow$, $\log(1/a_n)/\log\log n\to\infty$, and $na_n/\log n\to\infty$ $(n\uparrow\infty)$. Then we have

$$(3.10) \qquad \lim_{n\to\infty}\frac{\omega_n(a_n)}{(a_n\log(1/a_n))^{1/2}}=2^{1/2}\text{ a.s.}$$

PROOF. We restrict ourselves to the proof that $2^{1/2}$ is almost surely an upper bound for the expression on the left in (3.10).

By the Borel-Cantelli lemma and Theorem 3.5 it suffices to show that for every "small" $\varepsilon>0$ we have $\sum_{k=1}^{\infty}\mathbb{P}A_k<\infty$, where

$$A_k=\{\omega_{n_{k+1}}(a_{n_k})\ge((1+3\varepsilon)2a_{n_{k+1}}\log(1/a_{n_{k+1}}))^{1/2}\}$$

and $n_k=[(1+\frac{1}{2}\varepsilon)^k]$. By an application of Theorem 3.4 we have

$$(3.11) \qquad \begin{aligned}\mathbb{P}(A_k)\le &C\frac{1}{a_{n_k}}\exp\left(\frac{-(1-\varepsilon)}{2}(1+3\varepsilon)2\frac{a_{n_{k+1}}}{a_{n_k}}\log\frac{1}{a_{n_{k+1}}}\right.\\ &\left.\cdot\psi\left(\frac{((1+3\varepsilon)2\frac{a_{n_{k+1}}}{a_{n_k}}\log\frac{1}{a_{n_{k+1}}})^{1/2}}{(n_{k+1}a_{n_k})^{1/2}}\right)\right)\ .\end{aligned}$$

By the conditions of the present theorem it follows that the argument of the function ψ tends to 0 as $k\to\infty$. Hence we have by (3.11) for large k

$$(3.12) \qquad \mathbb{P}(A_k)\le C\frac{1}{a_{n_k}}\exp\left(-(1+\varepsilon)\log\frac{1}{a_{n_{k+1}}}\right)\le C\frac{1}{a_{n_{k+1}}}(a_{n_{k+1}})^{1+\varepsilon}=Ca_{n_{k+1}}^{\varepsilon}\ .$$

Now using $\log(1/a_n)/\log\log n\to\infty$ $(n\to\infty)$, the definition of n_k and (3.12), shows the finiteness of $\sum_{k=1}^{\infty}\mathbb{P}A_k$. □

Now we consider the important and delicate issue of weak convergence of the empirical processes U_n. First we present some general preparatory material on weak convergence.

We have already implicitely observed that U_n is a mapping from Ω to D. Before the question of weak convergence can be even raised we need to equip D with a σ-field of subsets and make sure that U_n is measurable with respect to this σ-field. Since our main purpose to study weak convergence of the processes U_n is to obtain weak convergence of functionals $\varphi(U_n)$ for a large class of measurable $\varphi : D \to I\!R$, the σ-field should not be taken too small. On the other hand, if we take the σ-field too large U_n may fail to be measurable.

Recall that $\mathcal{B}$ denotes the σ-field of Borel sets in $I\!R$. The first σ-field in D that we will consider is $\mathcal{D}$, the smallest σ-field containing the class of subsets

$$\{\{f \in D : f(t) \in B\}, t \in [0,1], B \in \mathcal{B}\}\ ,$$

called the projection σ-field. Because $U_n(t)$ is a random variable for each $t \in [0,1]$ it is obvious that U_n is $(\mathcal{F}, \mathcal{D})$-measurable.

It is interesting to note that $\mathcal{D}$ can be obtained in a different manner. Let us again equip D with the metric ρ defined in (3.5) and define $\mathcal{D}_\rho$ as the smallest σ-field containing the family of subsets

$$\{\{g \in D : \rho(f,g) < r\}, r \in (0,\infty), f \in D\}\ .$$

Without proof we state the result: $\mathcal{D} = \mathcal{D}_\rho$, and hence U_n is $(\mathcal{F}, \mathcal{D}_\rho)$-measurable.

There is yet another σ-field, the smallest σ-field $\mathcal{B}_\rho(D)$ that contains the class of all ρ-open subsets of D, called the Borel σ-field (w.r.t ρ). It may be shown that $\mathcal{D}_\rho$ is strictly smaller than $\mathcal{B}_\rho(D)$. This σ-field is too large for our purposes as it turns out that U_n is not measurable w.r.t. $\mathcal{B}_\rho(D)$. One might, however, define another metric on D (making D separable) such that U_n is measurable with respect to the Borel σ-field obtained from this new metric. By using $\mathcal{D} = \mathcal{D}_\rho$ we can get around these problems and have, moreover, the advantage that we can use the simple supremum-norm metric ρ.

We need a proper definition of weak convergence. Let (M, δ) be a metric space with metric δ and $\mathcal{M}_\delta$ the σ-field generated by the δ-open balls. Let $Z, Z_1, Z_2, ... : (\Omega, \mathcal{F}, I\!P) \to (M, \mathcal{M}_\delta)$ with induced probability measures $P, P_1, P_2, ...$ We say that Z_n converges weakly to Z (or P_n to P) if

$$I\!E f(Z_n) \to I\!E f(Z), \text{ as } n \to \infty, \text{ for all } f \in \mathcal{C}^*_\delta(M)\ , \tag{3.13}$$

where $\mathcal{C}^*_\delta(M)$ is the set of real valued functions on M that are bounded, δ-uniformly continuous and $(\mathcal{M}_\delta, \mathcal{B})$-measurable. We denote this by writing

$$Z_n \Rightarrow Z \text{ (or } P_n \Rightarrow P\text{), as } n \to \infty, \text{ in } (M, \mathcal{M}_\delta, \delta)\ .$$

Note that we did not impose the condition that M be separable. The reason is that in fact the space D isn't separable with respect to the metric ρ induced by the supremum-norm. Many important metric spaces, however, are separable. Not only $I\!R^d$ is an example but also C the space of all continuous functions on $[0,1]$, is separable (w.r.t ρ). In case of separability the theory simplifies. First of all, note that when M is separable $\mathcal{M}_\delta = \mathcal{B}_\delta(M)$, where $\mathcal{B}_\delta(M)$ is the σ-field generated by all δ-open subsets of M. Here is a general result without proof.

THEOREM 3.7. Let (M, δ) be a separable metric space. Then the following statements are equivalent:

$P_n \Rightarrow P$, as $n \to \infty$;

$$\int_M f dP_n \to \int_M f dP \ (n \to \infty), \text{ for all } f \in \mathcal{C}_\delta(M) \ ,$$

where $\mathcal{C}_\delta(M)$ is the set of real valued functions on M that are bounded and continuous;

$$\limsup_{n\to\infty} \ P_n(C) \le P(C), \text{ for all closed } C \ ;$$

$$\limsup_{n\to\infty} \ P_n(O) \ge P(O), \text{ for all open } O \ ;$$

$$\limsup_{n\to\infty} \ P_n(B) = P(B), \text{ for all } B \in \mathcal{B}_\delta(M) \text{ with } P(\partial B) = 0 \ .$$

In the special case that M is equal to $\mathbb{R}$ with the Euclidean metric, weak convergence of P_n to P is equivalent with

$$G_n(t) \to G(t), \text{ for each } t \in \mathbb{R} \text{ where } G \text{ is continuous,}$$

where G_n and G are the distribution functions corresponding to P_n and P respectively.

Let us now return to weak convergence in its general setting on arbitrary metric spaces $(M, \mathcal{M}_\delta, \delta)$. One of the important things is that it entails almost immediately the weak convergence of certain functionals.

THEOREM 3.8. Suppose that $Z_n \Rightarrow Z$ on $(M, \mathcal{M}_\delta, \delta)$ and let $\varphi : M \to \mathbb{R}$ be an $(\mathcal{M}_\delta, \mathcal{B})$-measurable and δ-uniformly continuous mapping. Then we have (δ_1 is the Euclidean metric on $\mathbb{R}$)

$$\text{(3.14)} \qquad \varphi(Z_n) \Rightarrow \varphi(Z), \text{ as } n \to \infty, \text{ in } (\mathbb{R}, \mathcal{B}, \delta_1) \ .$$

When M is separable it suffices when φ is continuous.

PROOF. It suffices to prove that

$$\mathbb{E} f(\varphi(Z_n)) \to \mathbb{E} f(\varphi(Z)), \text{ for all } f \in \mathcal{C}^*_{\delta_1}(\mathbb{R}) \ .$$

Note that for such f obviously $f \circ \varphi \in \mathcal{C}^*_\delta(M)$, so that (3.14) is immediate from the definition of weak convergence. □

We are now ready for application of the general weak convergence theory to the empirical processes U_n. We only need to specify the prospective weak limit. First we claim the existence of a Gaussian stochastic process

$$W(t), \ t \in [0,1] \ ,$$

in $(D, \mathcal{D}_\rho)$, with $\mathbb{E}W(t) = 0$ and covariance $\mathbb{E}W(s)W(t) = s \wedge t$ for all $s, t \in [0,1]$. This process is called a Wiener process or one dimensional Brownian motion. Observe that $W(0) = 0$ almost surely and Var $W(t) = t$. We may and will assume that the sample paths of W are continuous. Now our candidate limit is U, defined by

$$U(t) = W(t) - tW(1),\ t \in [0,1]\ . \tag{3.15}$$

The process U is called a Brownian bridge or tied-down Wiener process. It is easy to see that $\mathbb{E}U(t) = 0$ and that $\mathbb{E}U(s)U(t) = s \wedge t - st$, whence Var $U(t) = t(1-t), s, t \in [0,1]$. Moreover, U has also continuous sample paths and $U(0) = U(1) = 0$ almost surely.

The following elementary properties are the first step towards weak convergence of U_n. Let us fix arbitrary $0 \le t_1 < \ldots < t_r \le 1$ and note that

$$\begin{aligned}
&\mathbb{E}U_n(t_j)U_n(t_k) \\
&= n\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(1_{[0,t_j]}(X_i) - t_j)\frac{1}{n}\sum_{l=1}^{n}(1_{[0,t_k]}(X_l) - t_k) \\
&= \frac{1}{n}\ \sum\sum_{i \neq l}\ \mathbb{E}(1_{[0,t_j]}(X_i) - t_j)\mathbb{E}(1_{[0,t_k]}(X_l) - t_k) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(1_{[0,t_j]}(X_i) - t_j)(1_{[0,t_k]}(X_i) - t_k) \\
&= \mathbb{E}(1_{[0,t_j]}(X_1) - t_j)(1_{[0,t_k]}(X_1) - t_k) \\
&= t_j \wedge t_k - t_j t_k\ .
\end{aligned}$$

Recall that

$$\mathbb{E}U(t_j)U(t_k) = t_j \wedge t_k - t_j t_k\ .$$

It follows that the vectors $\langle U_n(t_1), \ldots, U_n(t_r)\rangle$ and $\langle U(t_1), \ldots, U(t_r)\rangle$ have the same covariance matrix

$$\Sigma = \begin{pmatrix} t_1(1-t_1) & \ldots & t_r \wedge t_1 - t_r t_1 \\ t_1 \wedge t_2 - t_1 t_2 & & \vdots \\ \vdots & & t_r \wedge t_{r-1} - t_r t_{r-1} \\ t_1 \wedge t_r - t_1 t_r & \ldots & t_r(1-t_r) \end{pmatrix}.$$

The multivariate central limit theorem yields

(3.16) $\langle U_n(t_1), ..., U_n(t_r)\rangle \Rightarrow \langle U(t_1), ..., U(t_r)\rangle$,

where, according to the definition, $\mathcal{L}(\langle U(t_1), ..., U(t_r)\rangle) = \mathcal{N}(0, \Sigma)$.

THEOREM 3.9 (functional central limit theorem). We have

(3.17) $U_n \Rightarrow U$, as $n \to \infty$, in $(D, \mathcal{D}_\rho, \rho)$.

PROOF. We will try to directly verify the definition of weak convergence. For each $m \in \mathbb{N}$ define $A_m : D \to D$ by

$$(A_m(x))(t) = \begin{cases} x((j-1)/m) & , t \in [(j-1)/m, j/m),\ j \in \{1, ..., m\} ; \\ x(1) & , t = 1 . \end{cases}$$

It is clear that $A_m(x)$ is an approximation of x. Choose $f \in \mathcal{C}^*_\rho(D)$; write $0 \le \sup_{x \in D} |f(x)| = C < \infty$ and $0 \le \delta_f(\varepsilon) = \sup\{|f(x) - f(y)| :\ \rho(x, y) < \varepsilon\} < \infty$. The finiteness of both C and $\delta_f(\varepsilon)$ follows from the definition of $\mathcal{C}^*_\rho(D)$. On the subset $\{\rho(U_n, A_m(U_n)) < \varepsilon\}$ we have $|f(U_n) - f(A_m(U_n))| \le \delta_f(\varepsilon)$, and on $\{\rho(U_n, A_m(U_n)) \ge \varepsilon\}$ we have $|f(U_n) - f(A_m(U_n))| \le 2C$. A similar remark can be made regarding U, so that

$$\begin{aligned}
&|\mathbb{E}f(U_n) - \mathbb{E}f(U)| \\
&\le |\mathbb{E}f(U_n) - \mathbb{E}f(A_m(U_n))| \\
&\quad + |\mathbb{E}f(A_m(U_n)) - \mathbb{E}f(A_m(U))| \\
&\quad + |\mathbb{E}f(A_m(U)) - \mathbb{E}f(U)| \\
&\le \delta_f(\varepsilon) + 2C\,\mathbb{P}(\rho(U_n, A_m(U_n)) \ge \varepsilon) \\
&\quad + |\mathbb{E}f(A_m(U_n)) - \mathbb{E}f(A_m(U))| \\
&\quad + \delta_f(\varepsilon) + 2C\,\mathbb{P}(\rho(U, A_m(U)) \ge \varepsilon) .
\end{aligned}$$

Now observe that

$$\rho(U_n, A_m(U_n)) \le \omega_n(1/m) ,$$

so that according to (3.8) (with $\varepsilon = \eta \in (0, \frac{1}{2}]$)

$$\begin{aligned}
&\mathbb{P}(\rho(U_n, A_m(U_n)) \ge \varepsilon) \le \mathbb{P}(\omega_n(1/m) \ge \varepsilon) \\
&\le Cm \exp\left(\frac{-(1-\eta)\varepsilon^2 m}{2} \psi\left(\frac{\varepsilon m}{n^{1/2}}\right)\right) .
\end{aligned}$$

It follows that, for each $\varepsilon > 0$,

$$(3.18) \qquad \limsup_{n\to\infty} \; \mathbb{P}(\rho(U_n, A_m(U_n)) \geq \varepsilon)$$

$$\leq Cm \exp(-\tfrac{1}{2}(1-\eta)\varepsilon^2 m) \to 0, \text{ as } m \to \infty \; .$$

Since the sample paths of U are continuous on $[0,1]$ and hence uniformly continuous, we have $\rho(U, A_m(U)) \to 0$, almost surely $(m \to \infty)$, and hence

$$(3.19) \qquad \mathbb{P}(\rho(U, A_m(U)) \geq \varepsilon) \to 0, \text{ as } m \to \infty \; ,$$

for each $\varepsilon > 0$.

We call $x \in D$ a generator of $\langle x_0, x_1, ..., x_m\rangle \in \mathbb{R}^{m+1}$ if $x(0) = x_0, x(1/m) = x_1, ...,$ $x(1) = x_m$, and define the function $\varphi : \mathbb{R}^{m+1} \to \mathbb{R}$ by setting

$$\varphi(x_0, x_1, ..., x_m) = f(A_m(x)) \; ,$$

where $x \in D$ is any generator of $\langle x_0, x_1, ..., x_m\rangle$. Let δ_{m+1} denote Euclidean distance in $\mathbb{R}^{m+1}$ and let $x, y \in D$ be generators of $\langle x_0, x_1, ..., x_m\rangle$ respectively $\langle y_0, y_1, ..., y_m\rangle$. Suppose that $\delta_{m+1}(\langle x_0, x_1, ..., x_m\rangle, \langle y_0, y_1, ..., y_m\rangle)$ is small, then $\rho(A_m(x), A_m(y))$ is small and hence $|f(A_m(x)) - f(A_m(y))|$ is small. It follows that φ is continuous (even δ_{m+1}-uniformly continuous) on $\mathbb{R}^{m+1}$. It is, moreover, bounded, because f is. Because $\mathbb{R}^{m+1}$ is separable, application of (3.16) with $r = m+1$ and $t_1 = 0,\ t_2 = 1/m, ..., t_{m+1} = 1$ yields

$$(3.20) \qquad \mathbb{E}f(A_m(U_n)) = \mathbb{E}\varphi(U_n(0), U_n(1/m), ..., U_n(1))$$

$$\to \mathbb{E}\varphi(U(0), U(1/m), ..., U(1)) = \mathbb{E}f(A_m(U)), \text{ as } n \to \infty \text{ for all } m \in \mathbb{N}.$$

Straightforward combination of (3.18)-(3.20) yields that $\mathbb{E}f(U_n) \to \mathbb{E}f(U)$, as $n \to \infty$. □

It should be noted that weak convergence is a property of the image probability measures of the random elements $Z, Z_1, Z_2, ...$. Therefore these random elements need not be defined on the same probability space, but should only map into the same metric space. It is, however, always possible to define them on the same probability space, and we did for convenience. But even then weak convergence doesn't tell us anything about convergence of sample paths. Given weak convergence, a special construction of probabilistically equivalent processes is possible, called the Skorohod (-Dudley-Wichura) construction, that appears to be very useful in theoretical arguments.

THEOREM 3.10. Suppose that

$$Z_n \Rightarrow Z, \text{ as } n \to \infty, \text{ in } (M, \mathcal{M}_\delta, \delta) \; ,$$

and that for some δ-separable $M^* \in \mathcal{M}_\delta$, (3.21) $\quad I\!\!P(Z \in M^*) = 1$.

Then there exists a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{I\!\!P})$ and $(\tilde{\mathcal{F}}, \mathcal{M}_\delta)$-measurable mappings $\tilde{Z}, \tilde{Z}_1, \tilde{Z}_2, ...$: $\tilde{\Omega} \to M$ with $\tilde{Z} \stackrel{d}{=} Z, \tilde{Z}_1 \stackrel{d}{=} Z_1, \tilde{Z}_2 \stackrel{d}{=} Z_2, ...$, such that

(3.22) $\quad \delta(\tilde{Z}_n, \tilde{Z}) \to 0$ a.s. $\quad (n \to \infty)$.

Rather than proving this theorem let us just apply it to obtain, almost without effort, a strengthening of (3.14) provided that (3.21) is fulfilled.

THEOREM 3.11. Let the conditions of Theorem 3.10 be fulfilled. Let $\varphi : M \to I\!\!R$ be $(\mathcal{M}_\delta, \mathcal{B})$-measurable and let φ be δ-continuous on M^*.

Then we have the weak convergence

$$\varphi(Z_n) \Rightarrow \varphi(Z), \text{ as } n \to \infty, \text{ in } (I\!\!R, \mathcal{B}, \delta_1) .$$

PROOF. Let us introduce the equivalent processes $\tilde{Z}, \tilde{Z}_1, \tilde{Z}_2, ...$ of the previous theorem. Write $E_1 = \{\tilde{\omega} \in \tilde{\Omega} : \tilde{Z}_n(\tilde{\omega}) \not\to \tilde{Z}(\tilde{\omega})\}$ and $E_2 = \{\tilde{\omega} \in \tilde{\Omega} : \tilde{Z}(\tilde{\omega}) \in M \backslash M^*\}$. It is clear that $\varphi(\tilde{Z}_n(\tilde{\omega})) \to \varphi(\tilde{Z}(\tilde{\omega}))$, as $n \to \infty$, for all $\tilde{\omega} \notin E_1 \cup E_2$. Since $\tilde{I\!\!P}(E_1 \cup E_2) = 0$, it follows that $\varphi(\tilde{Z}_n) \to \varphi(\tilde{Z})$, almost surely $(n \to \infty)$. But this implies that $\varphi(\tilde{Z}_n) \Rightarrow \varphi(\tilde{Z})$, and finally $\varphi(Z_n) \Rightarrow \varphi(Z)$, because $\varphi(\tilde{Z}) \stackrel{d}{=} \varphi(Z), \varphi(\tilde{Z}_n) \stackrel{d}{=} \varphi(Z_n)$. □

Let us now specialize these results to the case of the empirical processes.

THEOREM 3.12. There exists a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{I\!\!P})$ and $(\tilde{\mathcal{F}}, \mathcal{D}_\rho)$-measurable mappings $\tilde{U}, \tilde{U}_1, \tilde{U}_2, ... : \tilde{\Omega} \to D$ with

(3.23) $\quad \tilde{U} \stackrel{d}{=} U, \tilde{U}_n \stackrel{d}{=} U_n \ (n \in I\!\!N)$, and $\rho(\tilde{U}_n, \tilde{U}) \to 0$ a.s. $(n \to \infty)$.

Let $\varphi : D \to I\!\!R$ be $(\mathcal{D}_\rho, \mathcal{B})$-measurable and ρ-continuous on C, the space of all continuous functions on $[0,1]$. Then we have

(3.24) $\quad \varphi(U_n) \Rightarrow \varphi(U)$, as $n \to \infty$, in $(I\!\!R, \mathcal{B}, \delta_1)$.

PROOF. The conditions for application of Theorem 3.10 are satisfied because $C \in \mathcal{D}_\rho$ is ρ-separable. □

Before we turn to applications of the theory of this chapter, we present (without proof) some important refinements of Theorem 3.12, known as “strong approximations”.

THEOREM 3.13. There exists a sequence of Brownian bridges $U^{(1)}, U^{(2)}, ...$, such that for some $C \in (0, \infty)$

(3.25) $$\limsup_{n\to\infty} \frac{n^{1/2}}{\log n} \sup_{t\in[0,1]} |U_n(t) - U^{(n)}(t)| \le C \text{ a.s.}$$

Note that here U_n is based on the first n random variables of the sequence $\xi_1, \xi_2, ...$, whereas it easily follows from Theorem 3.1 that $\tilde{U}_n$ (in (3.23)) is not. On the other hand, in (3.23) the Brownian bridge limit does not depend on n, like in (3.25). However, for inferring precise almost sure results like laws of the iterated logarithm, in contrast to weak limit theorems, both results are useless. Statement (3.23) because there is no sequence structure as just indicated; (3.25) because the simultaneous distribution of the $U^{(n)}$, $n \in \mathbb{N}$, is not known.

The following theorem helps us to overcome these problems. Let K be a real-valued Gaussian stochastic process on $[0,\infty)\times[0,1]$ with continuous sample paths and $\mathbb{E}K(x,t) = 0$ and covariance $\mathbb{E}K(x,t)K(y,s) = (x \wedge y)(t \wedge s - ts)$, $x, y \in [0,\infty)$, $t, s \in [0,1]$. This process is called a Kiefer process. We consider $K(n,t)$, $n \in \mathbb{N}$, $t \in [0,1]$. Observe that for each fixed $n \in \mathbb{N}$, $n^{-\frac{1}{2}}K(n,I) \stackrel{d}{=} U \stackrel{d}{=} U^{(n)}$. It can also be shown that $\sum_{i=1}^{n} U_{(i)} \stackrel{d}{=} K(n,I)$, simultaneously in $n \in \mathbb{N}$, where $U_{(1)}, U_{(2)}, ...$ is a sequence of independent Brownian bridges.

THEOREM 3.14. There exists a Kiefer process K such that for some $C \in (0,\infty)$

$$\text{(3.26)} \qquad \limsup_{n\to\infty} \frac{n^{1/2}}{(\log n)^2} \sup_{t\in[0,1]} |U_n(t) - n^{-1/2}K(n,t)| \leq C \text{ a.s.}$$

Now we turn to applications. First let us apply (3.24) of Theorem 3.12. Consider the functional $\varphi_1 : D \to \mathbb{R}$ defined by

$$\text{(3.27)} \qquad \varphi_1(f) = \sup_{t\in[0,1]} |f(t)| \ .$$

This functional is $(\mathcal{D}_\rho, \mathcal{B})$-measurable and easily seen to be ρ-continuous (even uniformly ρ-continuous) on D. As a matter of fact for $f, g \in D$ with $\rho(f,g)$ small we have $|\varphi_1(f) - \varphi_1(g)|$ $\leq \sup_{t\in[0,1]}|f(t) - g(t)| = \rho(f,g)$ small. According to (3.24) we have

$$\sup_{t\in[0,1]} |U_n(t)| \Rightarrow \sup_{t\in[0,1]} |U(t)|, \text{ as } n \to \infty \ .$$

To obtain the distribution of the right hand side of (3.27) one makes use of reflection principles. One then finds the distribution already mentioned in (1.34).

Also consider the functional $\varphi_2 : D \to \mathbb{R}$ defined by

$$\varphi_2(f) = \int_0^1 f^2(t)dt \ ,$$

cf. (1.35). This functional is also $(\mathcal{D}_\rho, \mathcal{B})$-measurable and ρ-continuous on the entire space D. Let again $f, g \in D$ with $\rho(f,g)$ small. Then $|\varphi_2(f) - \varphi_2(g)| = |\int (f-g)(f+g)|$ $\leq \rho(f,g)\{\rho(f,0) + \rho(g,0)\}$ is also small; hence the functional is continuous although it is not uniformly continuous on D. Again (3.24) applies and yields

$$\int_0^1 U_n^2(t)dt \Rightarrow \int_0^1 U^2(t)dt, \text{ as } n \to \infty \ .$$

Our second application deals with the famous bootstrap. Using (3.23) it is easy to show that the empirical process can be bootstrapped. This goes as follows. Let $\mathbb{X}_n = \langle X_1, ..., X_n \rangle$ be a random sample from F with empirical distribution function F_n. (For convenience we assume that the X_i take values in $(0,1)$ instead of $\mathbb{R}$. This leads to processes in D, where we have our weak convergence machinery.) Suppose now that the distribution of some random variable $R(\mathbb{X}_n, F)$, depending on both $\mathbb{X}_n$ and F is desired. The bootstrap principle suggests approximating the distribution of $R(\mathbb{X}_n, F)$ by that of $R(\mathbb{X}_n^*, F_n)$, where $\mathbb{X}_n^* = \langle X_1^*, ..., X_n^* \rangle$, with the $X_i^*, 1 \le i \le n$, being an independent random sample from the (random) distribution function F_n. Often it is hard to compute the distribution of $R(\mathbb{X}_n^*, F_n)$, but it is possible to obtain as many independent values of $R(\mathbb{X}_n^*, F_n)$ as desired by Monte Carlo sampling, and thus estimate the distribution of $R(\mathbb{X}_n^*, F_n)$ to whatever accuracy is desired. This distribution is then used as an approximation to the distribution of $R(\mathbb{X}_n, F)$.

The fact that this procedure is asymptotically correct in many situations is well established in the literature. Here we will simply show that the empirical process $n^{1/2}(F_n - F)$ itself can be bootstrapped. First observe that when ξ is a $UN(0,1)$ random variable, then for any distribution function F, $F^{-1}(\xi)$ is distributed according to F (even when F has discontinuities, cf. (1.30)). Now let $\xi_{ni}^*, 1 \le i \le n, n \in \mathbb{N}$, be a triangular array of $UN(0,1)$ random variables for which (3.23) holds ($\tilde{U}_n$ is based on $\xi_{ni}^*, ..., \xi_{nn}^*$). Set $X_{ni}^* = F_n^{-1}(\xi_{ni}^*)$ and denote the empirical distribution function of the $X_{ni}^*, 1 \le i \le n$, with F_n^*. Thus

$$\begin{aligned}
(3.28) \qquad & \sup_{t \in [0,1]} |n^{1/2}(F_n^*(t) - F_n(t)) - \tilde{U}(F(t))| \\
&= \sup_{t \in [0,1]} |\tilde{U}_n(F_n(t)) - \tilde{U}(F(t))| \\
&\le \sup_{t \in [0,1]} |\tilde{U}_n(F_n(t)) - \tilde{U}(F_n(t))| + \sup_{t \in [0,1]} |\tilde{U}(F_n(t)) - \tilde{U}(F(t))| \\
&\le \sup_{t \in [0,1]} |\tilde{U}_n(t) - \tilde{U}(t)| + \sup_{t \in [0,1]} |\tilde{U}(F_n(t)) - \tilde{U}(F(t))| \,.
\end{aligned}$$

The right hand side of (3.28) converges to 0, almost surely, because of (3.23) and the uniform continuity of $\tilde{U}$ in conjunction with the Glivenko-Cantelli theorem.

This leads to the following result. We have almost surely (with respect to $(\Omega, \mathcal{F}, \mathbb{P})$, on which the X_i are defined), given $X_1, ..., X_n$, that the conditional distribution of $n^{1/2}(F_n^* - F_n)$ converges weakly to $U \circ F$, as $n \to \infty$, in $(D, \mathcal{D}_\rho, \rho)$. As a consequence for a measurable, ρ-continuous functional φ, we have almost surely, given $X_1, ..., X_n$, that the conditional distribution of $\varphi(n^{1/2}(F_n^* - F_n))$ converges weakly to $\varphi(U \circ F)$, as $n \to \infty$.

Finally we apply Theorem 3.4 to obtain almost sure results for uniform k-spacings. Let the $\xi_n, n \in \mathbb{N}$, be $UN(0,1)$ and independent again. Using a slightly different definition than usual we call the random variables

$$D_{i,n}(k) = \xi_{i+k:n} - \xi_{i:n}, i \in \{1, ..., n-k\} \,,$$

the uniform k-spacings. The minimal k-spacing is the random variable

$$m_{k,n} = \min_{1 \le i \le n-k} D_{i,n}(k) \,,$$

and the maximal k-spacing is the random variable

$$M_{k,n} = \max_{1 \le i \le n-k} D_{i,n}(k) \ .$$

These statistics can be defined in a similar way when we sample from another distribution than the uniform, and can then be used for inference about the maximum respectively minimum of the underlying density.

Let us focus on $m_{k,n}$ and sketch how for fixed $k \in \mathbb{N}$ an almost sure limit can be found. For this purpose let us introduce ($\alpha \in [0,1]$)

$$\Delta_n^+(\alpha) = \sup_{\substack{R \in \mathcal{R} \\ |R| = \alpha}} nP_n(R) \ .$$

The relation with $m_{k,n}$ becomes clear from the simple event identities

$$\begin{aligned} \{\Delta_n^+(\alpha) > k\} &= \{\text{there exists a half-open interval of length } \alpha \\ &\quad \text{containing at least } k+1 \text{ observations}\} \\ &= \{\text{there exists a } k\text{-spacing} < \alpha\} = \{m_{k,n} < \alpha\} \ . \end{aligned} \tag{3.29}$$

The relation with the oscillation modulus is

$$\Delta_n^+(\alpha) = n^{1/2}\Big(\sup_{\substack{R \in \mathcal{R} \\ |R| = \alpha}} U_n(R) + n^{1/2}\alpha\Big) \le n^{1/2}\omega_n(\alpha) + n\alpha \ . \tag{3.30}$$

It can be shown that

$$\lim_{n\to\infty} \frac{\log m_{k,n}}{\log(1/n)} = 1 + \frac{1}{k} \text{ a.s.} \tag{3.31}$$

For the sake of brevity, we will restrict ourselves to proving that for some $K \in (1, \infty)$

$$\limsup_{n\to\infty} \frac{\log m_{k,n}}{\log(1/n)} \le K \text{ a.s.} \tag{3.32}$$

In fact, with a bit more effort we could show, using as in the proof of Theorem 3.1 summation over a subsequence, that $1 + 1/k$ is almost surely an upper bound for the expression on the left hand side in (3.31). For (3.32) it suffices to prove that

$$\sum_{n=2}^{\infty} \mathbb{P}\left(\frac{\log m_{k,n}}{\log(1/n)} > K\right) < \infty \ . \tag{3.33}$$

Because $\log(1/n) < 0$ we have, using (3.29), (3.30) and (3.8),

$$p_n = I\!P(\log m_{k,n}/\log(1/n) > K) = I\!P(\log m_{k,n} < K\log(1/n))$$

$$= I\!P(m_{k,n} < n^{-K}) = I\!P(\Delta_n^+(n^{-K}) > k) = I\!P(\Delta_n^+(n^{-K}) \geq k+1)$$

$$\leq I\!P(\omega_n(n^{-K}) \geq n^{-\frac{1}{2}}(k+1-n^{1-K}))$$

$$\leq Cn^K \exp(-\tfrac{1}{2}(1-\varepsilon)(n^{K-1}(k+1-n^{1-K})^2\psi(n^{(K-1)}(k+1-n^{1-K}))) \ .$$

Let us now observe that

$$n^{K-1}(k+1-n^{1-K}) = (k+1)n^{K-1} - 1 \to \infty \quad (n\to\infty) \ ,$$

so that the value of the ψ-function tends to 0 as $n\to\infty$. Hence in this situation the rate of convergence to 0 becomes important. Using the second property in (2.15) we see that for n large

$$\psi((k+1)n^{K-1}-1) \geq 2(1-\varepsilon)\frac{\log((k+1)n^{K-1}-1)}{(k+1)n^{K-1}-1} \ .$$

Substitution yields for n large

$$p_n \leq Cn^K \exp(-(1-\varepsilon)^2(k+1-n^{1-K})\log((k+1)n^{K-1}-1))$$

$$\leq Cn^K \exp(-(1-\varepsilon)^3(k+1)\log((k+1)n^{K-1}))$$

$$\leq Cn^K \exp(-(1-\varepsilon)^3(k+1)(K-1)\log n)$$

$$\leq Cn^{K-(1-\varepsilon)^3(k+1)(K-1)} \leq C/n^2 \ ,$$

for $\varepsilon > 0$ small and K sufficiently large. This implies (3.33).

PROBLEMS

3.1 Compute $\sup\limits_{\omega\in\Omega}\ \sup\limits_{t\in[0,1]} |U_n(t)|$ and $\inf\limits_{\omega\in\Omega}\ \sup\limits_{t\in[0,1]} |U_n(t)|$.

3.2 Show that in (3.11) the argument of the function ψ tends to 0 as $k\to\infty$.

3.3 Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of numbers in $(0,1]$ such that $a_n \downarrow 0$, $na_n\uparrow$, and $\log(1/a_n)/\log\log n \to c \in [0,\infty)$, as $n\uparrow\infty$. Prove that

$$\limsup_{n\to\infty} \frac{\omega_n(a_n)}{(a_n\log\log n)^{1/2}} \leq (2(1+c))^{1/2} \quad \text{a.s.}$$

(Cf. Theorem 3.6.)

3.4 Define a stochastic process V on $[0,\infty)$ by

$$V(t) = (1+t)U\left(\frac{t}{1+t}\right) .$$

Show that $EV(t) = 0, t \geq 0$, and $EV(s)V(t) = s \wedge t,\ s,t \geq 0$. (Hence V is a Wiener process on $[0,\infty)$.)

3.5 Show that there does not exist a Brownian bridge U such that

$$\rho(U_n, U) \to 0 \text{ a.s. } (n \to \infty) .$$

(Cf. (3.23).)

3.6 Show that in the two-sample situation (see last application in chapter 1) we have if F_1, F_2 are continuous and $F_1 = F_2$ that

$$\sup_{t \in \mathbb{R}} \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} |F_{1,n_1}(t) - F_{2,n_2}(t)| \Rightarrow \sup_{t \in [0,1]} |U(t)|, \text{ as } n_1 \to \infty \text{ and } n_2 \to \infty ,$$

where F_{1,n_1} and F_{2,n_2} are independent empirical distribution functions of sizes n_1 and n_2 from F_1 and F_2 respectively, and where U denotes a Brownian bridge.