JOURNAL OF MULTIVARIATE ANALYSIS 8, 487–517 (1978)

On the Invariance Principle for Sums of Independent Identically Distributed Random Variables

PÉTER MAJOR

Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary

Communicated by M. Rosenblatt

The paper deals with the invariance principle for sums of independent identically distributed random variables. First it compares the different possibilities of posing the problem. The sharpest results of this theory are presented with a sketch of their proofs. At the end of the paper some unsolved problems are given.

I

1. POSING THE PROBLEM

We begin with two examples.

Let $X_1, X_2, \ldots$ be independent identically distributed random variables (i.i.d.r.v.'s), $EX_1 = 0$, $EX_1^2 = 1$. Denote their partial sums by $S_k = \sum_{i=1}^k X_i$, $k = 1, 2, \ldots, n$.

EXAMPLE a. Define

$$\alpha_n = \frac{1}{n^{1/2}} \max_{k \leqslant n} S_k .$$

It can be proved that

$$\lim P(\alpha_n < x) = 2 \int_0^x \frac{1}{(2\pi)^{1/2}} e^{-x^2/2}\, dx$$

if $x \geqslant 0$, and 0 otherwise.

EXAMPLE b. Define $\beta_n = 1/n$ times the number of k's for which $S_k > 0$. It can be proved that

$$\lim_{n\to\infty} P(\beta_n < x) = \frac{2}{\pi} \arcsin x^{1/2}, \qquad 0 \leqq x \leqq 1.$$

Both of these results have a proof consisting of the following two steps.

Received April 1978.

AMS 1970 subject classification: Primary 60G10.

Key words and phrases: Invariance principle, weak convergence.

0047-259X/78/0084–0487$02.00/0

Step 1. These results hold if $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$.

Step 2. If there is a sequence of i.i.d.r.v.'s, $EX_1 = 0$, $EX_1^2 = 1$, for which these results hold, then they hold for any sequence of i.i.d.r.v.'s with expectation 0 and variance 1.

This means that the partial sums of independent random variables behave, in some sense, very similarly to each other. The above examples are only special cases of a more general law. Roughly speaking the following statement holds true: The limit distribution of any "reasonable" functional of the sequence $S_1, S_2, \ldots, S_n$ is independent of the initial distribution of X_1.

We state this result in a more precise form. Theorems of this type are called functional limit theorems or (weak) invariance principles. The word "weak" refers to the fact that we deal with convergence in distribution, i.e., with weak convergence.

Let us remark that it is the central limit theorem, which is behind the weak invariance principle. The sequence $S_1, S_2, \ldots, S_n$ behaves as if it were a sequence of normal random variables. One may be interested also in the strong laws (laws holding with probability 1, e.g., the law of large numbers, the law of iterated logarithm) of probability theory. Here again the same phenomenon appears. The limit of the partial sums does not depend on the initial distribution. Theorems explaining this phenomenon are called strong invariance principles.

2. The Notion of Weak Convergence in Metric Spaces

Let us be given a separable metric space (X, ρ) and a sequence of probability measures $\mu_0, \mu_1, \mu_2, \ldots$ on the Borel sets of X. We want to find a good definition of convergence for the sequence μ_i, $i = 1, 2, \ldots$, which is a generalization of the notion "convergence of distribution functions on the real line."

Theorem 2.1. *Definitions* 1, 2a, 2b, *and* 3 *given below are equivalent to each other.*

Definition 1. $\lim_{n\to\infty} \mu_n = \mu_0$ if and only if for every bounded continuous function $f(x)$ on X we have

$$\int f(x)\, d\mu_n(x) \to \int f(x)\, d\mu_0(x).$$

(i.e., in the language of functional analysis the μ_n's, as elements of the conjugate space of $C(X)$, tend weakly to μ_0).

Definition 2a. $\lim_{n\to\infty} \mu_n = \mu_0$ if and only if for every open set G, $G \subset X$ we have

$$\liminf_{n\to\infty} \mu_n(G) = \mu_0(G).$$

Definition 2b. $\lim_{n\to\infty} \mu_n = \mu_0$ if and only if for every closed set F, $F \subset X$ we have

$$\limsup_{n\to\infty} \mu_n(F) \leqq \mu_0(F).$$

Definition 3. $\lim_{n\to\infty} \mu_n = \mu_0$ if and only if for every subset A of X whose boundary has 0 μ_0 measure, we have

$$\lim_{n\to\infty} \mu_n(A) = \mu_0(A).$$

Fact. Let X be the real line R^1. Define $F_n(x) = \mu_n([-\infty, x))$. $\mu_n \to \mu_0$ if and only if $F_n(x) \to F_0(x)$ for every point of continuity of $F_0(x)$.

The following result is very important for us. Let (X, ρ_1) and (Y, ρ_2) be two separable metric spaces and $\mathscr{F}: X \to Y$ a continuous mapping. Let a sequence μ_n, $n = 0, 1, 2,...,$ of probability measures be given on X, $\mu_n \to \mu_0$. The mapping $\mathscr{F}$ induces a sequence of probability measures on Y in a natural way,

$$\nu_n(A) = \mu_n(x: \mathscr{F}x \in A) \qquad \text{for every Borel set } A \subset Y,\ n = 0, 1, 2,... .$$

Theorem (2.2). *We have* $\lim_{n\to\infty} \nu_n = \nu_0$. *Specifically, choosing* $Y = R^1$, *we obtain that given any continuous functional* $\mathscr{F}$ *on* X, $\lim \mu_n = \mu_0$ *implies that for* $F_n(u) = \mu_n (\mathscr{F}x < u)$, $n = 0, 1, 2,..., F_n \to F_0$ *in distribution.*

Remark. In the above theorem the condition about the continuity of $\mathscr{F}$ can be somewhat weakened. It is enough to assume that $\mathscr{F}$ is continuous with probability 1 with respect to the measure μ_0.

Finally we give a metrization of the weak convergence. We define the so-called Prochorov distance.

Definition. Let μ and ν be two probability measures on a metric space (X, ρ). Their Prochorov distance $\rho^P(\mu, \nu)$ is $\rho^P(\mu, \nu) = \inf\{\epsilon : \mu(A^\epsilon) + \epsilon \geqslant \nu(A)$ for every closed set $A\}$, where $A^\epsilon = \{x : \rho(x, A) \leqslant \epsilon\}$.

Let us remark that, at first sight, it cannot be seen that ρ^P is a metric, let alone that it metricizes the weak convergence.

Theorem (2.3). *If* X *is a separable metric space, then so is the space of probability measures on* X *with the Prochorov distance as metric.* $\rho^P(\mu_n, \mu_0) \to 0$ *if and only if* $\mu_n \to \mu_0$. *If* X *is complete, so is the space of probability measures with the Prochorov distance.*

3. The Weak Invariance Principle

First we need some definitions.

Definition of the Wiener Process

Let a stochastic process $W(t, \omega)$, $0 \leqslant t \leqslant T$ be given on a probability space space $(\Omega, \mathscr{A}, P)$. It is called a (standard) Wiener process on the interval $[0, T]$ if

(a) for any n and $0 \leqslant t_1 \leqslant t_2 \leqslant \cdots \leqslant t_n \leqslant T$ the random vector $(W(t_1), W(t_2), \ldots, W(t_n))$ is normally distributed,

(b) $EW(t) = 0$, $EW(t_1)\, W(t_2) = \min(t_1, t_2)$ for every $0 \leqslant t, t_1, t_2 \leqslant T$

(c) for every ω, $W(\cdot, \omega)$ is a continuous function on the interval $[0, T]$.

Having a random variable ξ in a probability space $(\Omega, \mathscr{A}, P)$ which takes its values in a metric space X, we can speak of the distribution of ξ. This is a probability measure μ on the Borel sets of X defined by the relation

$$\mu(A) = P(\xi \in A) \qquad \text{for every Borel set } A.$$

A Wiener process on the interval $[0, T]$ can be considered as a random variable taking values in the space $C[0, T]$ (The space of continuous functions on $[0, T]$ with the supremum norm.)

Definition of the Wiener Measure

Wiener process on $[0, 1]$ is a random variable taking values in $C[0, 1]$. Its distribution is called the Wiener measure, and is denoted by μ_w.

Let $X_1, X_2, \ldots, X_n$ be i.i.d.r.v.'s, $EX_1 = 0$, $EX_1^2 = 1$. Set $S_k = \sum_{i=1}^k X_i$, $k = 1, 2, \ldots, n$, and define a random polygon $S_n(t)$, $0 \leqslant t \leqslant 1$ in the following way:

$$S_n\left(\frac{k}{n}\right) = \frac{1}{n^{1/2}} S_k, \qquad k = 1, 2, \ldots, n \ (S_0 = 0),$$

$$S_n(t) = n\left[\left(t - \frac{k}{n}\right) S_n\left(\frac{k+1}{n}\right) + \left(\frac{k+1}{n} - t\right) S_n\left(\frac{k}{n}\right)\right] \quad \text{if } \frac{k}{n} \leqq t \leqq \frac{k+1}{n}.$$

$S_n(t)$ can be considered as a random variable taking values in $C[0, 1]$. Denote its distribution by μ_n.

Theorem (3.1). (*The weak invariance principle.*)

$$\lim_{n\to\infty} \mu_n = \mu_w.$$

Theorems (3.1) and (2.2) together explain why we get the same limit distribution for a lot of functionals of the sequence $(S_1, S_2, ..., S_n)$. In Example a of the first section we get our invariance upon choosing the functional

$$\mathscr{F}_1 x(t) = \sup_{0 \leqslant t \leqslant 1} x(t), \qquad x(t) \in C[0, 1].$$

In Example b one would choose the functional

$$\mathscr{F}_2 x(t) = \lambda(t: x(t) > 0) \qquad (\lambda \text{ means the Lebesgue measure}).$$

This functional is not continuous at the points $x \in C[0, 1]$, where

$$\lambda(t: x(t) = 0) > 0.$$

But it is not difficult to see that the set of points of dicontinuity has 0 Wiener measure, therefore the remark after Theorem (2.2) can be applied in this case.

Applying the functional $\mathscr{F}_2$ to $S_n(t)$, we get a result slightly different from that in Example b of the first section. To obtain the original result we define a random polygon $S_n^*(t)$, slightly different from $S_n(t)$, in the following way:

$$\text{If sign } S_k = \text{sign } S_{k+1}, \; S_n^*(t) = S_n(t) \text{ for } \frac{k}{n} \leqslant t \leqslant \frac{k+1}{n}.$$

$$\left.\begin{array}{l} \text{If sign } S_k \neq \text{sign } S_{k+1} \\ \text{or } S_k = 0 \text{ or } S_{k+1} = 0 \end{array}\right\} \Rightarrow \begin{array}{l} S_n^*\left(\frac{k}{n}\right) = \frac{S_k}{n^{1/2}} \; S_n^*\left(\frac{2k+1}{2n}\right) = 0 \\ S_n^*\left(\frac{k+1}{n}\right) = \frac{S_{k+1}}{n^{1/2}} \end{array}$$

and $S_n^*(t)$ is defined by linear interpolation in the intervals $(k/n, (2k+1)/(2n))$ and $((2k+1)/(2n), (k+1)/n)$.

The distribution of $S_n^*(t)$ tends also to the Wiener measure. (One can see it quite easily, e.g., by applying Theorem (4.3).)

On the other hand

$$\mathscr{F}_2 S_n^*(t) = \frac{1}{n} \text{ times the number of } k\text{'s for which } S_k = 0$$

Let us remark that we actually got a little more than we wanted to get in the first section. We know a priori that the sequence $\mathscr{F} S_n(t)$, $n = 1, 2, ...$ has a limit distribution, and it agrees with the distribution of $\mathscr{F} W(t)$.

In this section we dealt with weak convergence in $C[0, 1]$. We finish it with a result that explains what weak convergence in $C[0, 1]$ means.

THEOREM (3.2). *Let $X_n(t, \omega)$, $0 \leqslant t \leqslant 1$ be a sequence of random processes with continuous trajectories. The distributions of the processes $X_n(t, \omega)$ as measures on $C[0, 1]$ tend weakly to the distribution of $X_0(t, \omega)$ if and only if*

(a) *for every k and $0 \leqslant t_1 \leqslant t_2 \leqslant \cdots \leqslant t_k \leqslant 1$ the random vectors $(X_n(t_1),\ldots, X_n(t_k))$ tend in distribution to the random vector $(X_0(t_1),\ldots, X_0(t_k))$ as $n \to \infty$,*

(b) *for every $\epsilon > 0$*

$$\lim_{\delta\to 0} \sup_n P\Big(\sup_{|t_2-t_1|\leqslant\delta} |X_n(t_2) - X_n(t_1)| > \epsilon\Big) = 0.$$

4. The Connection Between Different Types of Convergence

How to Measure Speed of Convergence in the Invariance Principle?

The following diagram shows the connection among different types of convergence:

$$\begin{tikzcd}
\text{Convergence with probability } 1 \arrow[d] \\
\text{Convergence in probability} \arrow[d] \\
\text{Weak Convergence of the distributions.}
\end{tikzcd}$$

On the other hand the following result holds.

Theorem (4.1). *Let μ_n, $n = 0, 1, 2,\ldots$ be probability measures on a separable metric space X. Let* $\lim_{n\to\infty} \mu_n = \mu_0$. *Then one can construct a probability space $(\Omega, \mathscr{A}, P)$ and random variables ξ_n, $n = 0, 1, 2\ldots$ taking values in X in such a way that the distribution of ξ_n is μ_n, $n = 0, 1, 2,\ldots$ and* $\lim_{n\to\infty} \xi_n(\omega) = \xi_0(\omega)$ *with probability* 1.

The convergence in probability can be metricized in the following way:

If ξ and η are two random variables taking values on (X, ρ), their distance may be defined as

$$\rho_s(\xi, \eta) = \inf(\epsilon : P(\rho(\xi, \eta) > \epsilon) < \epsilon).$$

Then ρ_s metricizes convergence in probability.

The next theorem brings into connection the Prochorov distance and ρ_s.

Theorem (4.2). *If the random variable ξ has distribution μ, η has distribution ν, and $P(\rho(\xi, \eta) > \alpha) < \beta$, then we have $\mu(A^\alpha) + \beta > \nu(A)$ for any closed set A. (A^α is defined as $A^\alpha = (x : \rho(x, A) \leqslant \alpha)$).*

On the other hand, if μ and ν are two probability measures on a separable metric space X such that $\mu(A^\alpha) + \beta > \nu(A)$ for any closed set A, then there exists, for *every $\epsilon > 0$, a probability space $(\Omega, \mathscr{A}, P)$ and two random variables ξ and η with distributions μ and v on it in such a way that*

$$P(\rho(\xi, \eta) \geqslant \alpha + \epsilon) \leqslant \beta + \epsilon.$$

If the space X is complete, this relation holds even with $\epsilon = 0$.

Theorem (4.2) yields, as a special case, that

$$\rho^P(\mu, \nu) = \inf \rho_s(\xi, \eta), \qquad \xi \text{ has distr. } \mu, \eta \text{ has distr. } \nu.$$

Because of Theorem (4.1) and the diagram at the beginning of this section, Theorem (3.1) is equivalent to the following

THEOREM (4.3). *One can construct a probability space $(\Omega, \mathscr{A}, P)$ and processes $\tilde{S}_n(t)$ and $\tilde{W}_n(t)$, $n = 1, 2,\ldots$ on it in such a way that $\tilde{W}_n(t)$ is a Wiener process on $[0, 1]$, $\tilde{S}_n(t)$ is a random polygon whose distribution agrees with that of $S_n(t)$ for every n, and*

$$\lim_{n\to\infty} P\left(\sup_{0\leqslant t\leqslant 1} | S_n(t) - \tilde{W}_n(t)| > \epsilon\right) = 0 \qquad \text{for every} \quad \epsilon > 0.$$

($\tilde{W}_n(t)$ may be the same $W(t)$ for every n).

Let us remark that Theorem (4.3) is also equivalent to the following

THEOREM (4.3′). *Let $F(x)$ be a distribution function,*

$$\int x\, dF(x) = 0, \qquad \int x^2\, dF(x) = 1.$$

There exist two sequences of i.i.d.r.v's $X_1, X_2,\ldots, X_n$ and $Y_1, Y_2,\ldots, Y_n$ with distribution function $F(x)$ resp. $\phi(x)$ (from now on $\phi(x)$ denotes the standard normal distribution function) in such a way that the partial sums $S_k = \sum_{i=1}^k X_i$ and $T_k = \sum_{i=1}^k Y_i$, $k = 1, 2,\ldots, n$ satisfy the relation

$$P\left(\sup_{k\leqslant n} \frac{| S_k - T_k |}{n^{1/2}} > \epsilon\right) \to 0 \qquad \text{for every} \quad \epsilon > 0.$$

To show the equivalence between Theorem (4.3) and Theorem (4.3′) one has to observe that over the T_k's a Wiener process $W(t)$ on the interval $[0, n]$ can be spanned (i.e., one can construct a Wiener process $W(t)$, $0 \leqslant t \leqslant n$, such that $W(k) = T_k$ for $k = 1, 2,\ldots, n$ if the probability space is rich enough) and that $\sup_{k\leqslant n} \sup_{|t-k|<1} | W(t) - W(k)|$ is relatively small. Now if we are interested in the speed of convergence in the invariance principle, then it is natural to ask (preserving the notations of Theorem (4.3′)) for which sequences (α_n, β_n) can the relation

$$P\left(\sup_{k\leqslant n} \frac{|S_k - T_k|}{n^{1/2}} > \alpha_n\right) < \beta_n$$

be satisfied with an appropriate construction. Theorem (4.2) indicates that this question is closely related to the determination of the Prochorov distance between the Wiener measure and the distribution of $S_n(t)$.

Finally, the problem of the strong invariance principle can be formulated in the following way: For which sequences α_n, $n = 1, 2,...$ can the relation

$$\frac{S_n - T_n}{\alpha_n} \to 0 \qquad \text{with probability 1,}$$

or

$$\limsup \frac{|S_n - T_n|}{\alpha_n} \qquad \text{bounded with probability 1,}$$

be satisfied?

5. The First Estimates of Speed of the Convergence in the Invariance Principle: Estimates from Below

The following result was proven by V. Strassen.

Theorem (5.1). *Let $F(x)$ be a distribution function*

$$\int x\, dF(x) = 0, \qquad \int x^2\, dF(x) = 1.$$

There exist two sequences of i.i.d.r.v.'s $X_1, X_2, ...$ and $Y_1, Y_2, ...$, the first one with distribution $F(x)$, the second one with distribution function $\phi(x)$ in such a way that the partial sums $S_n = \sum_{i=1}^n X_i$, $T_n = \sum_{i=1}^n Y_i$ satisfy the relation

$$\frac{S_n - T_n}{(n \log\log n)^{1/2}} \to 0 \qquad \text{with probability 1.}$$

This result shows that the validity of the law of iterated logarithms for the sequence T_n implies the same also for the sequence S_n. Actually Strassen used this result in order to prove a sharpened form of the law of iterated logarithm.

Later he proved that if $\int x^4\, dF(x) < \infty$ (i.e., the fourth moment exists) then a construction satisfying

$$\limsup \frac{|S_n - T_n|}{n^{1/4}(\log n)^{1/2}(\log\log n)^{1/4}} < K \qquad \text{with probability 1}$$

is possible. (K is chosen appropriately.) In these constructions the so-called Skorochod embedding was used.

It turned out that these constructions cannot give a better approximation even if we impose some new restrictions on $F(x)$. Thus the question arose whether the last mentioned result is sharp.

A bound from below resulted from the solution of the so-called "stochastic geyser problem." The problem is the following one: Let an infinite sequence of random variables $S_1 + \epsilon_1, S_2 + \epsilon_2, \ldots$ be given, where $S_1, S_2, \ldots$ are the partial sums of i.i.d.r.v.'s with some distribution function $F(x)$, $\epsilon_1, \epsilon_2, \ldots$ (called error terms) are arbitrary random variables satisfying $\lim_{n\to\infty} \epsilon_n/f(n) = 0$ a.s. with some deterministic function $f(n)$. Question: is it possible to determine the unknown distribution function $F(x)$ with probability one by the sequence $S_1 + \epsilon_1, S_2 + \epsilon_2, \ldots$?

The following result was proven:

THEOREM (5.2). *If $\int e^{tx}\, dF(x) < \infty$ in some neighborhood of the origin (i.e., for $|t| < t_0$ some t_0) and $f(n) = \log n$, the answer is affirmative.*

Let us now change the problem a little. Let us assume that we know that $S_1, S_2, \ldots$ are the partial sums of i.i.d.r.v.'s either with a known distribution function $F(x)$ or with distribution function $\phi(x)$. We want to decide, with the help of the sequence $S_1 + \epsilon_1, S_2 + \epsilon_2, \ldots$, which one of the two cases happened. With a slight modification of the proof of Theorem (5.2) one gets that this problem can be solved with probability 1 if $\limsup |\epsilon_n|/(\log n) \leqslant c$ (c depends on $F(x)$).

This result has the following consequence:

If $S_1, S_2, \ldots$ are the partial sums of i.i.d.r.v.'s with distribution function $F(x)$, $T_1, T_2, \ldots$ are the partial sums of i.i.d.r.v.'s with distribution function $\phi(x)$, then

$$\limsup \frac{|S_n - T_n|}{\log n} > 2c \qquad \text{with probability 1.}$$

Indeed, defining $\epsilon_n = (S_n - T_n)/2$, we cannot decide whether we have the sequence $S_1 - \epsilon_1) S_2 - \epsilon_1, \ldots$ or $T_1 + \epsilon_1, T_2 + \epsilon_2, \ldots$. Therefore

$$\limsup |\epsilon_n|/(\log n) > c$$

must hold with probability 1.

It the tail behavior of the distribution function $F(x)$ is not very nice then the following argument, due to L. Breiman, gives a better lower bound.

THEOREM (5.3). *Let $X_1, X_2, \ldots$ be i.i.d.r.w.'s with distribution function $F(x)$; $Y_1, Y_2, \ldots$ i.i.d.r.v.'s with distribution function $\phi(x)$,*

$$S_n = \sum_{i=1}^{n} X_i, \qquad T_n = \sum_{i=1}^{n} Y_i, \qquad n = 1, 2, \ldots .$$

Assume that

$$\sum_{n=0}^{\infty} P(|X_1| > K_n) = \infty,$$

where K_n is a monotone numerical sequence, $\limsup K_n/(\log n) > 0$. *Then we have*

$$|S_n - T_n| > \frac{K_n}{4} \qquad \text{for infinitely many } n\text{'s with probability } 1.$$

Proof of Theorem (5.3). We have

$$\sum_{n=1}^{\infty} P(|X_n| > K_n) = \infty$$

therefore, by the Borel–Cantelli lemma, $|S_n - S_{n-1}| = |X_n| > K_n$ infinitely often with probability 1. On the other hand $|T_n - T_{n-1}| = |Y_n| < K_n/2$ for almost every n with probability 1. If both $|S_n - S_{n-1}| > K_n$ and $|T_n - T_{n-1}| < K_n/2$, then either $|S_n - T_n| > K_n/4$ or $|S_{n-1} - T_{n-1}| > K_n/4$, and this proves the statement.

Let us remark that the relation

$$\sum_{n=1}^{\infty} P(|X_1| > K_n) = \infty$$

is a moment type condition. For $K_n = n^{1/r}$, $r > 0$ it is equivalent to $E|X_1|^r = \infty$. It holds with $K_n = c \log n$ for every $c > 0$ if and only if $E(\exp t|X_1|) = \infty$ for every $t \neq 0$.

Let us now return to the "most regular case," to the case when $E\exp(tX_1) < \infty$ for $|t| < t_0$. We know that in this case, for an appropriate construction,

$$|S_n - T_n| = O(n^{1/4}(\log n)^{1/2}(\log\log n)^{1/4}) \qquad \text{with probability } 1.$$

On the other hand

$$\limsup \frac{S_n - T_n}{\log n} > c \qquad \text{with probability } 1$$

holds always for appropriate $c > 0$. One would like to know which is the real bound.

The first partial answer was given by Csörgő and Révész. They proved that under some conditions, the most important of which is $EX_1^3 = 0$, a construction satisfying

$$|S_n - T_n| = o(n^{1/(6-\epsilon)}), \qquad \epsilon > 0 \text{ is arbitrarily small}$$

is possible. In that construction they exploited the Edgeworth expansion of the central limit theorem. This states that

$$P(S_n < n^{1/2}x) = \phi(x) + \frac{1}{n^{1/2}} \frac{\mu_3}{6} (1 - x^2) \frac{1}{(2\pi)^{1/2}} e^{-x^2/2} + O\left(\frac{1}{n}\right),$$

where $\mu_3 = EX_1^3$. This specifically means that $| P(S_n < n^{1/2}x) - \phi(x)|$ has typically the magnitude $O(1/n^{1/2})$, but if the third moment of X_1 agrees with that of a standard normal random variable, it is only $O(1/n)$.

Now again the question arises, whether this result is sharp, whether the condition $EX_1^3 = 0$ is essential. The counterpart of this problem is the following one: For which $f(n)$ can the third moment of $F(x)$ be estimated by the sequence $S_1 + \epsilon_1$, $S_2 + \epsilon_2$,... in the stochastic geyser problem?

This question was investigated, but an estimate of the third moment was obtained only in the case $f(n) > \log n$. One has the feeling: Either the third moment of $F(x)$ can be estimated also in the case $f(n) = \log n$, or a construction satisfying $| S_n - T_n | = o(n^{1/6})$ or even more can be found. The second case turned out to be true.

6. Sharp Estimates for the Speed of Convergence in the Invariance Principle

Roughly speaking the results of this section state the following: The estimates from below given in the previous section are sharp. First we deal with the case when the moment-generating function exists.

Theorem (6.1). *If $\int x\, dF(x) = 0$, $\int x^2\, dF(x) = 1$, $\int \exp(tx)\, dF(x) < \infty$ for some $| t | < t_0$, then a sequence of i.i.d.r.v.'s X_1, X_2,... with distribution function $F(x)$ and another one Y_1, Y_2,... with distribution function $\phi(x)$ can be constructed in such a way that*

$$\limsup \frac{| S_n - T_n |}{\log n} < c \qquad \text{with probability } 1,$$

where

$$S_n = \sum_{i=1}^{n} X_i, \qquad T_n = \sum_{i=1}^{n} Y_i,$$

c is an appropriate constant.

This theorem is a consequence of the following more general theorem.

THEOREM (6.2). *Under the conditions of Theorem* (6.1) *the X_i's and Y_i's can be constructed in such a way that*

$$P(\max_{k\leqslant n} | S_k - T_k | > C \log n + x) < Ke^{-\lambda x},$$

where C, K, λ depend only on $F(x)$.

Theorem (6.1) explains why no estimate was found for the third moment with the help of $S_1 + \epsilon_1$, $S_2 + \epsilon_2$,... in the case $f(n) > \log n$. The fact that a distribution function $F(x)$, $\int x^3 \, dF(x) \neq 0$ can be found in such a way that

$$\limsup \frac{| S_n - T_n |}{\log n} < c \qquad \text{with probability } 1$$

(c actually can be arbitrarily small) explains this.

Theorem (5.3) shows that the existence of the moment-generating function is an essential condition in Theorem (6.1). If the moment-generating function does not exist, the next theorem describes the situation.

First some notation: Let $H(x)$, $x \geqslant 0$ be a monotone continuous function satisfying the relations:

(1) $H(x)/x^{2+\delta}$ is monotone increasing for some $\delta > 0$ and $x > x_0$.

(2) $\log H(x)/x$ is monotone decreasing for $x > x_0$.

Define K_n by the equation $H(K_n) = n$.

THEOREM (6.3). *If* $\int x \, dF(x) = 0$, $\int x^2 \, dF(x) = 1$, $\int H(| x |) \, dF(x) < \infty$, *then the i.i.d.r.v.'s X_i with distribution function $F(x)$ and Y_i with distribution function $\phi(x)$ can be constructed in such a way that*

$$P\left(\limsup \frac{| S_n - T_n |}{K_n} \leqq c\right) = 1$$

for appropriate c.

In order to compare Theorem (6.3) with Theorem (5.3) let us remark that $\int H(| x |) \, dF(x) < \infty$ is equivalent to the relation

$$\sum P(| X_1 | > K_n) < \infty.$$

In the case $H(x) = x^r$, $r > 2$, Theorem (6.3) yields, as a special case, the following

THEOREM (6.4). *If* $\int x \, dF(x) = 0$, $\int x^2 \, dF(x) = 1$, $\int | x |^r \, dF(x) < \infty$ *for some $r > 2$, then a construction satisfying* $\lim S_n - T_n/n^{1/r} = 0$ *with probability* 1 *is possible.*

To see how Theorem (6.3) implies Theorem (6.4), it is enough to observe that if $\int |x|^r dF(x) < \infty$, then $\int |x|^r f(|x|) dF(x) < \infty$ for an appropriate function $f(x), f(x) \to \infty$. So we can apply Theorem (6.3) with $H(x) = |x|^r f(x)$.

The weak invariance principle counterpart of Theorem (6.3) is the following

THEOREM (6.5). *Under the conditions of Theorem* (6.3) *for every* x, $K_n < x < C_1(n \log n^{1/2}$ *there exist two finite sequences* $X_1, X_2, \ldots, X_n$ *and* $Y_1, Y_2, \ldots, Y_n$ *such that*

$$P(\sup_{k \leqslant n} | S_k - T_k | > x) \leqq C_2 \frac{n}{H(ax)},$$

where C_1, C_2, *and* a *are positive constants depending only on* $F(x)$.

Let us now turn to the case when no more than the existence of two moments is assumed. Theorems (4.3′) and (5.1) speak of these cases. Theorem (5.3) and the fact that it turned out to be sharp in many cases would suggest that a construction satisfying

$$| S_n - T_n | = o(n^{1/2}) \qquad \text{with probability 1}$$

should exist. But this is not true. It can be shown that the result

$$| S_n - T_n | = o((n \log \log n)^{1/2}) \qquad \text{with probability 1}$$

cannot be improved.

Thus it might appear that Breiman's argument does not yield a sharp result in this case. However, this is only because we put the question in an improper way.

The following theorem generally holds true.

THEOREM (6.6). *Let* $\int x\, dF(x) = 0$, $\int x^2\, dF(x) = 1$. *There exists a sequence of i.i.d.r.v.'s* $X_1, X_2, \ldots$ *with distribution function* $F(x)$ *and a sequence of independent normal random variables* $Y_1, Y_2, \ldots$. $EY_n = 0$,

$$EY_n^2 = \int_{-n^{1/2}}^{n^{1/2}} x^2\, dF(x) - \left[\int_{-n^{1/2}}^{n^{1/2}} x\, dF(x)\right]^2, \qquad n = 1, 2, \ldots$$

in such a way that the partial sums $S_n = \sum_{i=1}^n X_i$, $T_n = \sum_{i=1}^n Y_i$, $n = 1, 2, \ldots$ *satisfy the relation*

$$\lim \frac{| S_n - T_n |}{n^{1/2}} = 0 \qquad \text{with probability 1.}$$

Theorems (4.3′) and (5.1) are easy consequences of this result. Theorem (6.6) also explains why the relation $| S_n - T_n | = o((n \log \log n)^{1/2})$ cannot be improved in the case when the T_n's are sums of independent standard normal random variables. At the first sight it may seem unnatural to approximate random variables with expectation 0 and variance 1 with nonstandard normal random variables. But the following argument may illuminate such a procedure.

Define

$$\begin{aligned} X_n' &= X_n \qquad \text{if} \quad |X_n| \leqslant n^{1/2} \\ &= 0 \qquad \text{if} \quad |X_n| > n^{1/2} \end{aligned}$$

and $S_n' = \sum_{i=1}^{n} X_i'$. Then $\sum P(X_n \neq X_n') < \infty$, and therefore $X_n = X_n'$ for almost every n with probability 1. This means that $| S_n - S_n' | < K(\omega)$ with probability 1. Therefore we may approximate the S_n''s instead of the S_n's in Theorem (6.6). But doing so, it is natural to couple X_n' with a Y_n whose first two moments agree with that of X_n'. Naturally EX_n' may differ from zero. On the other hand EX_n' is very near to 0, therefore Y_n can be substituted by $Y_n - EY_n$ in Theorem (6.6). But changing the variance of Y_n back to 1 may violate Theorem (6.6).

7. The Invariance Principle for the Empirical Distribution Function

The subject of this section is somewhat different from the previous ones. But the proofs and results are similar, and the importance of this subject in mathematical statistics may justify why we discuss it.

First we define the empirical distribution function. Let $X_1, X_2, \ldots, X_n$ be independent random variables uniformly distributed on the interval [0, 1], i.e., $P(X_i < t) = t$ for $0 \leqslant t \leqslant 1$. The empiricadistribution function is

$$F_n(t) = \frac{1}{n} \sum_{k=1}^{n} I_k(t) \qquad 0 \leqslant t \leqslant 1,$$

where

$$\begin{aligned} I_k(t) &= 1 \qquad \text{if} \quad X_k < t \\ &= 0 \qquad \text{if} \quad X_k \geqslant t. \end{aligned}$$

Let $W(t)$ be a Wiener process on the interval [0, 1]. We call the process $B(t) = W(t) - tW(1)$ a Brownian bridge. Let us observe that $W(1)$ and the process $B(t) = W(t) - tW(1)$ are independent of each other. (Since we deal with normal random variables, it is enough to check that $W(1)$ and $W(t) - tW(1)$ $0 \leqslant t \leqslant 1$ are uncorrelated.)

A Brownian bridge has the following properties.

(1) $(B(t_1), B(t_2), \ldots, B(t_k))$ is a normally distributed random vector for any k, $0 \leqslant t \leqslant t_2 \leqslant \cdots \leqslant t_k \leqslant 1$.

(2) $EB(t) = 0$, $EB(t_1) B(t_2) = t_1(1 - t_2)$ for any $0 \leqslant t \leqslant 1$, $0 \leqslant t_1 \leqslant t_2 \leqslant 1$.

(3) The trajectories of $B(t)$ are continuous functions for every ω.

If a process $B(t)$, $0 \leqslant t \leqslant 1$ satisfies (1), (2), and (3) and ξ is a standard normal random variable, independent of $B(t)$, then $W(t) = B(t) + t\xi$ is a Wiener process, and $B(t) = W(t) - tW(1)$.

Let us now return to the empirical distribution function $F_n(t)$.

We consider the process $n^{1/2}[F_n(t) - t]$, $0 \leqslant t \leqslant 1$. It can be seen that this process is asymptotically Gaussian, with expectation 0 and with the same covariance structure as that of $B(t)$. Therefore it is natural to expect that this process is near to an appropriate Brownian bridge $B(t)$. In fact the following statement holds true:

THEOREM (7.1). *If the probability space is rich enough, there exists a Brownian bridge $B_n(t)$ in such a way that*

$$P(\sup_{0 \leqslant t \leqslant 1} [n^{1/2} \mid n^{1/2}(F_n(t) - t) - B_n(t)| > C \log n + x) < Ke^{-\lambda x}$$

for all x, where C, K, λ are positive absolue constants.

Roughly speaking this result means that

$$| n^{1/2}[F_n(t) - t] - B_n(t)| = O((\log n)/n^{1/2}).$$

An argument similar to the proof of Theorem (5.2), may show that this result is sharp.

One would like to formulate an invariance principle for the empirical distribution function also in the language of measures, as it was done for sums of independent random variables. Here some minor difficulties arise. As $F_n(x)$ is not a continuous function, we cannot speak of the distribution of $n^{1/2}(F_n(t) - t)$ in the space $C[0, 1]$. There are two possibilities for getting rid of this inconvenience. One can either slightly change the definition of $F_n(t)$ to get a continuous function, or define a more general function space than the space $C[0, 1]$ and speak of convergence of measures in this space. (In the literature generally the second possibility is chosen, and the so-called D-space is defined with the Skorochod metric.)

But since the invariance principle for the empirical distribution function formulated in the language of measures has no such consequence which cannot already be seen directly from Theorem (7.1), we do not discuss it here.

We show an application of Theorem (7.1). Define

$$\gamma_n = n^{1/2} \sup_{0 \leqslant t \leqslant 1} | F_n(t) - t |$$

and

$$\delta_n = n \int_0^1 [F_n(t) - t]^2 \, dt.$$

THEOREM (7.2). *Both γ_n and δ_n have limit distributions as $n \to \infty$. The limit distribution of γ_n agrees with the distribution of* $\sup_{0 \leqslant t \leqslant 1} |B(t)|$, *and the limit distribution of δ_n with the distribution of $\int_0^1 B^2(t)\, dt$.*

γ_n is called the Kolmogorov–Smirnov and δ_n is called the von Mises statistics. Their limit distributions can be explicitly given and they are tabulated in every collection of statistical tables.

Finally we give a result in which we approximate the sequence $F_1(t), F_2(t),\ldots$ with Brownian bridges whose joint structure is similar to that of the sequence $F_1(t), F_2(t),\ldots$.

THEOREM (7.3). *Let $X_1, X_2, \ldots$ be an infinite sequence of i.i.d.r.v.'s with uniform distribution on* $[0, 1]$. *Let us define the empirical distribution functions $F_1(t), F_2(t),\ldots$ as we did at the beginning of this section. There exists a sequence of independent Brownian bridges $B_1(t)$, $B_2(t),\ldots$ in such a way that*

$$P(\sup_{1 \leqslant k \leqslant n} \sup_{0 \leqslant t \leqslant 1} | k(F_k(t) - t) - \sum_{j=1}^{k} B_j(t)| > (C \log n + x) \log n) < Ke^{-\lambda x}$$

for all x and n, where C, K, λ are positive absolute constants. Especially we have

$$P\left(\limsup_{n} \sup_{0 \leqslant t \leqslant 1} \frac{| n[F_n(t) - t] - \sum_{j \leqslant 1}^{n} B_j(t)|}{\log^2 n} < C\right) = 1$$

for appropriate C.

It is not known whether $\log n$ can be written instead of $\log^2 n$ in the last formula and whether the previous formula can be similarly improved or not.

The meaning of Theorem (7.3) may be probably more understandable with the help of the following remark.

Define $U_k(t) = I_k(t) - t$. Then the $U_k(t)$'s are independent for different k's, their covariance structure agrees with that of $B(t)$ and

$$n^{1/2}[F_n(t) - t] - \sum_{j=1}^{n} B_j(t) = \sum_{j=1}^{n} U_j(t) - \sum_{j=1}^{n} B_j(t).$$

II

8. HOW IS A GOOD APPROXIMATION OF THE PARTIAL SUMS MADE?

a. *The quantile transform.* Our aim is the following: Given a sequence of partial sums of i.i.d.r.v.'s $S_1, S_2,\ldots$ in a sufficiently rich probability space we want to approximate it with a sequence of normal random variables $T_1, T_2,\ldots$.

We often approximate the T_n's with the S_n's, or construct the sequence S_n and T_n at the same time instead of solving the original problem. But having solved these modified problems we can easily solve also the original one. All we need to do is to complete the sequence S_n to two sequences S_n, T_n with a prescribed joint distribution. Some standard theorems in measure theory like the existence of conditional distribution functions, the Tulcea–Ionescu theorem enable us to carry out this completition.

The first constructions were made with the help of the so-called Skorochod embedding. As it only rarely yields sharp results, and because the results obtained with its help can be generally derived otherwise, we do not discuss it.

First we speak of the so-called quantile transformation which can yield sharp results in the case when we have few moments. (Generally if we have four or more moments it ceases to give sharp results.)

First we discuss the following problem:

Let $F(x)$ and $G(x)$ be two distribution functions on the real line. Let us construct two random variables ξ and η. ξ with distribution function $F(x)$, η with distribution function $G(x)$, so that $|\xi - \eta|$ be small.

Let us make some remarks:

If ξ has distribution function $F(x)$, and $F(x)$ is strictly monotone, continuous, then $\alpha = F(\xi)$ is uniformly distributed on $[0, 1]$. On the other hand, if α is uniformly distributed on $[0, 1]$ and $F^{-1}(x)$ means the inverse of $F(x)$ then $\xi = F^{-1}(\alpha)$ has distribution function $F(x)$. In the general case one must be a little careful. If $F(x)$ has jumps, or if it is constant on an interval, then some problems arise. But these difficulties disappear upon introducing some slight modifications.

Fact 1. Let α be uniformly distributed on $[0, 1]$. Define the inverse of the distribution function $F(x)$ (we assume that the distribution functions are continuous from the left) as

$$F^{-1}(t) = \sup(x\colon F(x) \leqq t).$$

Then $F^{-1}(\alpha)$ has a distribution function $F(x)$.

Fact 2. Let ξ have distribution function $F(x)$ and let ϵ be a uniformly distributed tandom variable on $[0, 1]$, independent of ξ. Then

$$\alpha = \tilde{F}(\xi) = F(\xi) + \epsilon[F(\xi + 0) - F(\xi)]$$

is uniformly distributed on $[0, 1]$.

Now let us have two distribution functions $F(x)$ and $G(x)$. We propose two ways of constructing random variables ξ and η with distribution functions $F(x)$ and $G(x)$.

First method of construction. Let α be a uniformly distributed random variable on [0, 1]. Define

$$\xi = F^{-1}(\alpha), \qquad \eta = G^{-1}(\alpha).$$

Second method of construction. Let ξ have distribution function $F(x)$. Given the random variable ξ (and possibly an ϵ uniformly distributed on [0, 1], and independent of ξ) we define η as $\eta = G^{-1}(\tilde{F}(\xi))$.

Both constructions are called quantile transformations. We do not distinguish between them. We identify these two constructions because they produce the same joint distribution of the variables ξ and η, and actually this joint distribution is what we have to define. The next result shows an optimum property of the quantile transformation.

THEOREM (8.1). *Let $F(x)$ and $G(x)$ be two distribution functions*

$$\int |x| \, dF(x) < \infty, \qquad \int |x| \, dG(x) < \infty.$$

Let $f(x)$ be a convex function on the real line. Then we have

$$\inf E(f(\xi - \eta)) = \int_0^1 f(F^{-1}(x) - G^{-1}(x)) \, dx,$$

$$\xi \text{ has distribution funct. } \mu, \qquad \eta \text{ has distribution funct. } \nu.$$

Since $f(\xi - \eta) \geqq c(\xi - \eta) + d \geqq -c(|\xi| + |\eta|) + d$ with appropriate c and d, the (possibily infinite) expression $Ef(\xi - \eta)$ always has meaning. The expression on the right side is the value of $Ef(\xi - \eta)$ if we take the quantile transform.

Proof of Theorem (8.1). Let us first consider the special case when the measures determined by $F(x)$ and $G(x)$ are concentrated on a finite set $X = \{x_1, x_2, ..., x_n\}$. Then the minimum is attained for some pair ξ_0, η_0. Let us introduce the notation $p(x, y) = P(\xi_0 = x, \eta_0 = y)$; $x, y \in X$. We may assume that the following property (x) holds:

$$(x) \text{ For every } x_i > x_j, \quad y_i > y_j, \quad x_i, x_j, y_i, y_j \in X$$
$$\min[p(x_i, y_j), p(x_j, y_i)] = 0$$

Let us assume that contrary to our hypothesis there exist some $x_i > x_j$, $y_i > y_j$ in such a way that

$$p = \min[p(x_i, y_j), p(x_j, y_i)] > 0.$$

Define the random variables $\tilde{\xi}_0$, $\tilde{\eta}_0$ with the dollowing joint distribution $\tilde{p}(x, y)$, $x, y \in X$

$$\begin{aligned} \tilde{p}(x_i, y_i) &= p(x_i, y_i) + p, \\ \tilde{p}(x_j, y_j) &= p(x_j, y_j) + p, \\ \tilde{p}(x_i, y_j) &= p(x_i, y_j) - p, \\ p(x_j, y_i) &= p(x_j, y_i) - p, \\ \tilde{p}(x, y) &= p(x, y) \qquad \text{otherwise.} \end{aligned}$$

The distribution of $\tilde{\xi}_0$ agrees with that of ξ_0, the distribution of $\tilde{\eta}_0$ with that of η_0. On the other hand

$$\begin{aligned} &E[f(\tilde{\xi}_0 - \tilde{\eta}_0) - f(\xi_0 - \eta_0)] \\ &\quad = p[f(x_i - y_i) + f(x_j - y_j) - f(x_i - y_j) - f(x_j - y_i)]. \end{aligned}$$

Because of convexity, the relations

$$x_j - y_i \leqq \begin{matrix} x_j - y_j \\ x_i - y_i \end{matrix} \leqq y_j - x_i$$

and

$$\tfrac{1}{2}[(x_i - y_i) + (x_j - y_j)] = \tfrac{1}{2}[(x_j - y_i) + (x_i - y_j)]$$

imply that

$$f(x_i - y_i) + f(x_j - y_j) \leqslant f(x_i - y_j) + f(x_j - y_i).$$

Applying the last inequality we obtain that

$$Ef(\tilde{\xi}_0 - \tilde{\eta}_0) \leqq Ef(\xi_0 - \eta_0).$$

So if (x) does not hold for the pair (ξ_0, η_0) we can substitute it by the pair $(\tilde{\xi}_0, \tilde{\eta}_0)$ in this way.

Should the pair $(\tilde{\xi}_0, \tilde{\eta}_0)$ still not satisfy (x), we can continue the above procedure. We get, in finitely many steps, to a pair (ξ^*, η^*) which satisfies (x), and the minimum is taken for this (ξ^*, η^*).

But relation (x) determines the joint distribution of ξ and η. Since the quantile transformation has the property (x), we proved Theorem (8.1) in this case.

If $F(x)$ and $G(x)$ are concentrated in an interval $[-u, u]$, we can approximate them with the distribution functions

$$F_n(x) = F\left(\frac{[nx)}{n}\right), \qquad G_n(x) = G\left(\frac{[nx]}{n}\right), \qquad n = 1, 2,...$$

where [ ] means integer part.

Applying the already proved part of the theorem, and letting n to infinity we get that the result holds in this case too.

Taking limit in u, in a similar way, we get that the result holds in general.

The most interesting case of Theorem (8.1) is the case $f(x) = |x|^r$, especially when $r = 1$ or $r = 2$. Choosing $f(x) = (x - E\xi - E\eta)^2$ we see that the quantile transformation minimizes the variance. The case $r = 1$ shows that the quantile transformation gives the so-called Wasserstein distance in the one-dimensional case.

DEFINITION (Wasserstein distance). Let two probability measures μ and ν be given on the metric space (X, ρ). Their Wasserstein distance is

$$\rho_w(\mu, \nu) = \inf E\rho(\xi, \eta) \qquad \xi \text{ has distr. } \mu,\ \eta \text{ has distr. } \nu.$$

COROLLARY TO THEOREM (8.1). *On the real line with the usual metric the Wasserstein distance of the distributions $F(x)$ and $G(x)$ equals*

$$\int_0^1 |F^{-1}(x) - G^{-1}(x)|\, dx.$$

In our applications of the quantile transform the role of $F(x)$ is usually played by $F_n(x) = P(S_n < n^{1/2}x)$, where S_n is the sum of n i.i.d.r.v.'s with expectation 0, and the role of $G(x)$ is played by $\phi(x)$. By the central limit theorems $|F_n(x) - \phi(x)|$ is small, therefore the random variables $F_n^{-1}(\alpha)$ and $\phi^{-1}(\alpha)$ (α is uniformly distributed random variable on [0, 1]) are near each other. To estimate their distance we need an estimate of the speed of convergence in the central limit theorem. Let us emphasize that to get sharp results we must use non-uniform (depending on x) estimates for $|F_n(x) - \phi(x)|$. The reason for this is the following: Since the derivative of $\phi(x)$ is almost 0 at x if x is near infinity or minus infinity, thus $F_n^{-1}(\alpha) - \phi^{-1}(\alpha)$ may be very big at α near 0 or 1, despite of the fact that $|F_n(x) - \phi(x)|$ is small. Thus we must exploit the fact that $|F_n(x) - \phi(x))$ is much smaller at a large x than at an x near the origin. Otherwise estimating, e.g., the variance $E[F_n^{-1}(\alpha) - \phi^{-1}(\alpha)]^2$ we may get too weak results (the effect of $F_n^{-1}(\alpha) - \phi^{-1}(\alpha)$ for α near 0 or 1 would be too roughly estimated).

Let us now briefly describe how two sequences $S_1, S_2, \ldots$ and $T_1, T_2, \ldots$ of partial sums of i.i.d.r.v.'s with distribution function $F(x)$ and $\phi(x)$ are obtained with the help of the quantile transform. Naturally we want that

$$\sup_{k \leqslant n} |S_k - T_k| \qquad \text{be small.}$$

We choose an appropriate numerical sequence of integers $0 = n_0 < n_1 < n_2 < \cdots < n_k < \cdots$. We construct the random variables $(S_{n_k} - S_{n_{k-1}}, T_{n_k} - T_{n_{k-1}})$, $k = 1, 2, \ldots$ with the help of the quantile transform ($S_{n_0} = T_{n_0} = 0$).

We may assume that the pairs of random variables $(S_{n_k} - S_{n_{k-1}}, T_{n_k} - T_{n_{k-1}})$ are independent. Now, since the random variables S_{n_k}, T_{n_k}, $k = 1, 2, \ldots$ are already given, they can be completed to two sequences of random variables $S_1, S_2, \ldots$ and $T_1, T_2, \ldots$ in an arbitrary way. Now to estimate

$$\sup_{k \leqslant n} | S_k - T_k |$$

one has to investigate the following two expressions:

(a) $\sup | \sum [(S_{n_j} - S_{n_{j-1}}) - (T_{n_j} - T_{n_{j-1}})]$,

(b) $\sup\limits_{k} \sup\limits_{n_{k-1}<n<n_k} | S_n - S_{n_{k-1}} |$ and $\sup\limits_{k} \sup\limits_{n_{k-1}<n<n_k} | T_- - T_{n_{k-1}} |$.

With a good choice of the sequence n_k, the expressions in (a) and (b) have the same magnitude. Generally it is worth combining the above method with truncation of the summands at the beginning of the construction. The right level of truncation is suggested by Breiman's argument.

Let us now make a very rough calculation that may suggest what the magnitude of approximation with this method is like. We assume that the summands have many moments. First we estimate the variance.

$$a_k = E[(S_{n_k} - S_{n_{k-1}}) - (T_{n_k} - T_{n_{k-1}})]^2,$$

$$a_k = O(1).$$

One cannot expect a better result, since for the distribution function $F_{n_k - n_{k-1}}(x) = P(S_{n_k} - S_{n_{k-1}} < (n_k - n_{k-1})^{1/2}x)$ the inequality

$$| F_{n_k}(x) - \phi(x)| > \frac{c}{(n_k - n_{k-1})^{1/2}}$$

holds for a typical x. (Unless we have some special conditions.)

By the Kolmogorov inequality

$$\sup_{j \leqslant k} | S_{n_j} - T_{n_j} | = O(k^{1/2})$$

On the other hand,

$$\sup_{j \leqslant k-1} \sup_{n_j < n < n_{j+1}} | S_n - S_{n_j} | = O(n_k - n_{k-1})^{1/2})$$

if the sequence $n_k - n_{k-1}$ is monotone and tends to infinity fast enough. An analogous result holds for the T's for typical ω's.

So the good choice of n_k is $n_k = k^2$, and

$$\sup_{k \leqslant n} | S_k - T_k | = O(n^{1/4})$$

is obtained.

In an exact calculation some log factor also appears. At any rate the above calculation indicates that to get better results than $O(n^{1/4})$, a new technique must be worked out. This is described in the next section.

9. How Is a Good Approximation of the Partial Sums Made?

b. *The conditional quantile transform.* It is enough to discuss the following problem: Given an integer n, (for the sake of convenience we assume that $n = 2^m$) and a sequence $T_1, T_2, \ldots, T_n$ of partial sums of independent standard normal random variables, construct a finite sequence $S_1, S_2, \ldots, S_n$ of partial sums of .i.i.d.r.v.'s with distribution function $F(x)$, $\int x\, dF(x) = 0$, $\int x^2\, dF(x) = 1$ in such a way that

$$P(\sup_{k \leqslant n} | S_k - T_k | > x)$$

be small for $x > 0$.

We start the construction with the construction of S_n. It can be done simply using the quantile transform

$$\frac{S_n}{n^{1/2}} = F_n^{-1}\left(\phi\left(\frac{T_n}{n^{1/2}}\right)\right),$$

where $F_n(x) = P(S_n < n^{1/2}x)$.

The next task is to construct $S_{n/2}$. Now the joint distribution of $S_{n/2}$ and S_n is prescribed, and S_n is already given. Thus on the set $S_n = y$, the conditional distribution $P(S_{n/2} < x \mid S_n = y)$ is prescribed. That is the reason why we have to work with conditional distributions from now on.

For technical reasons we construct first the random variable $2S_{n/2} - S_n$ instead of $S_{n/2}$. Let us observe that $2S_{n/2} - S_n$ and S_n are uncorrelated, and their joint distribution is asymptotically normal. Therefore it is natural to expect that the conditional distribution $P(2S_{n/2} - S_n < x \mid S_n = y)$ is asymptotically normal with variance n. And in fact, the following result holds true.

Theorem (9.1). *If $\int \exp(tx)\, dF(x) < \infty$ for $| t | < t_0$, then*

$$P(S_n > n^{1/2}x) = [1 - \phi(x)] \exp\left[O\left(\frac{x^3 + 1}{n^{1/2}}\right)\right],$$

$$P(S_n < -n^{1/2}x) = \phi(-x) \exp\left[O\left(\frac{x^3 + 1}{n^{1/2}}\right)\right]$$

for $0 \leqslant x \leqslant \epsilon n^{1/2}$.

If, in addition, $F(x)$ has an integrable characteristic function then

$$P(2S_{n/2} - S_n > n^{1/2}x \mid S_n = n^{1/2}y)$$
$$= [1 - \phi(x)] \exp\left(O\left(\frac{x^3 + x^2 \mid y \mid + \mid y \mid + 1}{n^{1/2}}\right)\right),$$
$$P(2S_{n/2} - S_n < -n^{1/2}x \mid S_n = n^{1/2}y)$$
$$= \phi(-x) \exp\left(O\left(\frac{x^3 + x^2 \mid y \mid + \mid y \mid + 1}{n^{1/2}}\right)\right)$$

for $0 < x < \epsilon n^{1/2}$, $\mid y \mid < \epsilon n^{1/2}$. *$O(\cdot)$ is uniform in x and y.*

Let us assume for a while that the conditions for Theorem (9.1) are satisfied. We define $2S_{n/2} - S_n$ in the following way:

$$\frac{1}{n^{1/2}}(2S_{n/2} - S_n) = G_n^{-1}\left(\phi\left(\frac{1}{n^{1/2}}(2T_{n/2} - T_n)\right) \middle| \frac{S_n}{n^{1/2}}\right),$$

where $G_n(x \mid y) = P(2S_{n/2} - S_n < n^{1/2}x \mid S_n = n^{1/2}y)$ and $G_n^{-1}(x \mid y)$ means the inverse of the conditional distribution function $G_n(x \mid y)$ in parameter x with fixed y. (We define the inverse in the general case, as we did in Section 8.) $(1/n^{1/2})(2T_{n/2} - T_n)$ is independent of T_n, and therefore also of $S_n/n^{1/2}$, which is a function of T_n.

Thus it is easy to see that the so defined $2S_{n/2} - S_n$ has the prescribed conditional distribution with respect to the condition $S_n = n^{1/2}y$. Now $S_{n/2}$ is defined as $S_{n/2} = \frac{1}{2}(2S_{n/2} - S_n) + \frac{1}{2}S_n$, and we obtain that the pair $(S_{n/2}, S_n)$ has the prescribed joint distribution.

The next step is to construct $S_{n/4}$ and $S_{(3/4)n} - S_{n/2}$. It can be done in the same way. First we define the variables $2S_{n/4} - S_{n/2}$ and $2(S_{(3/4)n} - S_{n/2}) - (S_n - S_{2/n})$ as

$$G_{n/2}^{-1}\left(\phi\left(\frac{2^{1/2}}{n^{1/2}}(2T_{n/4} - T_{n/2})\right) \middle| \left(\frac{2}{n}\right)^{1/2} S_{n/2}\right)$$

and

$$G_{n/2}^{-1}\left(\phi\left(\left(\frac{2}{n}\right)^{1/2}[2(T_{(3/4)n} - T_{n/2}) - (T_n - T_{n/2})]\right) \middle| \left(\frac{2}{n}\right)^{1/2}(S_n - S_{n/2})\right).$$

Similarly to the argument about S_n, one obtains that both pairs $(S_{n/4}, S_{n/2})$ and $(S_{(3/4)n} - S_{n/2}, S_n - S_{n/2})$ have the right joint distributions. We claim that these pairs are independent, which implies that even the quadruple $(S_{n/4}, S_{n/2}, S_{(3/4)n}, S_n)$ has the right distribution. To see this, observe that the pairs $(S_{n/2}, S_n - S_{n/2})$ and $(2T_{n/4} - T_{n/2}, 2(T_{(3/4)n} - T_{n/2}) - (T_n - T_{n/2}))$ are independent. It is so because $S_{n/2}$ and S_n are the functions of T_n and $T_{n/2}$. So the random variables $S_{n/2}, S_n - S_{n/2}, 2T_{n/4} - T_{n/2}, 2(T_{(3/4)n} - T_{n/2}) -$

$(T_n - T_{n/2})$ are independent. Since the pair $(S_{n/4}, S_{n/2})$ is a function of the first and third variables while the pair $(S_{(3/4)n} - S_{n/2}, S_n - S_{n/2})$ is that of the second and fourth ones, they are independent as we claimed.

In the next step we define the variables $S_{[(2k+1)/8]n} - S_{(k/4)n}$, $k = 0, 1, 2, 3$ in the same way, and we go on till every S_j is defined. The procedure ends in $\log n$ steps, and at the end we obtain a sequence $S_1, S_2, \ldots, S_n$ with appropriate joint distribution.

Because of Theorem (9.1) $G_n^{-1}(\phi(x) \mid y) \sim x$. This relation will imply that $S_k \sim T_k$. A detailed calculation shows that roughly speaking the quantity $S_k - T_k$ grows up only with constant in every step of the construction. The construction finishes in $\log n$ steps, and this may explain why the result

$$\sup_{k \leqslant n} | S_k - T_k | = O(\log n)$$

is obtained.

The proof of the fact that $\sup_{k \leqslant n} | S_k - T_k |$ is small in this construction heavily depends on Theorem (9.1). So the question arises: What happens if the characteristic function of $F(x)$ is not integrable? One would hope that Theorem (9.1) holds also without this condition. This hope is however illusory. There are examples which show that the existence of the moment-generating function in itself does not guarantee the validity of Theorem (9.1).

The second part of Theorem (9.1) was proved by the integration of the conditional density function. The conditional density function can be expressed by the original densities, and thus the central limit theorem for density functions yields a good asymptotic for it.

Therefore the following idea seems natural. Take a sequence of i.i.d.r.v.'s $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ with normal distribution $E\epsilon_1 = 0$, and $E\epsilon_1^2$ is small. Define the random variables $\alpha_k = \sum_{j=1}^k \epsilon_j$, $k = 1, 2, \ldots, n$, and try to construct first a sequence $S_k + \alpha_k$, $k = 1, 2, \ldots, n$ near the sequence T_k, $k = 1, 2, \ldots, n$. (The α's and S's are independent.) If $E\epsilon_1^2$ is sufficiently small, then the closeness of the sequences $S_k + \alpha_k$ and T_k implies the closeness of the S_k's and T_k's. We apply the same construction that was described above. As $X_1 + \alpha_1$ has smooth density function, one may expect that Theorem (9.1) holds for its distribution and therefore a good construction can be obtained.

This construction does not always work but it gives a good approximation in several cases when the original one does not give the same. Namely the condition about the existence of integrable characteristic function can be substituted with the following weaker one: The distribution function $F(x)$ can be written in the form $F(x) = pF_1(x) + (1 - p)F_2(x)$, where $F_1(x)$ and $F_2(x)$ are distribution functions, $p > 0$, and $F_1(x)$ has density function. (If $F(x)$ has integrable characteristic function, then it has continuous bounded density.)

If the random variables $X_1, X_2, \ldots$ are bounded with probability 1, then there exists a modification of the conditional quantile transformation that yields a good approximation of the partial sums $S_k = \sum_{i=1}^k X_i$.

Let us now describe this construction.

We assume that we are given the random variables $T_1, T_2, \ldots, T_n$ and the set of random variables $\{X_1, X_2, \ldots, X_n\}$. We assume that $S_n = \sum_{k=1}^{n} X_k$ is the quantile transform of T_n. We may moreover assume that $\{X_1, \ldots, X_n\}$ is independent of the random variables $2(T_{(2k+1)2^{j-1}} - T_{k2^j}) - (T_{(k+1)2^j} - T_{k2^j}), j = 2, 3, \ldots, m, k = 0, 1, \ldots, n/2^j - 1$. In fact, we may define first the random variables $X_1, \ldots, X_n$ and T_n, then, independently of them, $2T_{n/2} - T_n$, then, independently of all the previous random variables, $2T_{n/4} - T_{n/2}$ and $2(T_{(3/4)n} - T_{n/2}) - (T_n - T_{n/2})$, etc.

We want to redefine the order of the random variables $X_1, X_2, \ldots, X_n$, i.e., to make a random permutation of the indices in such a way, that the partial sums of these random variables be close to the appropriate T_k's. Of course, we must be careful that our variables have the prescribed joint distribution.

Let us first remark that the density function $f(x_1, \ldots, x_k)$ of

$$P(X_1 = x_1, \ldots, X_k = x_k)$$

is invariant under any permutation of the set $\{x_1, \ldots, x_k\}$.

In the first step we tell which ones of our variables have an index less than or equal to $n/2$; that is, we define the sets $A_1 = \{X_1(\omega), \ldots, X_{n/2}(\omega)\}$ and $A_2 = \{X_{n/2+1}(\omega), \ldots, X_n(\omega)\}$. We must satisfy the following condition:

$$P(\{X_1, \ldots, X_{n/2}\} = \{x_{i_1}, \ldots, x_{i_{n/2}}\} \mid \{X_1, \ldots, X_n\} = \{x_1, \ldots, x_n\}) = 1\Big/\binom{n}{n/2}$$

for any set $\{x_1, \ldots, x_n\}$ and for any of its subsets $\{x_{i_1}, \ldots, x_{i_n}\}$ with $n/2$ elements.

Doing so, we guarantee that the joint distribution of $\{X_1, \ldots, X_{n/2}\}$ and $\{X_{n/2+1}, \ldots, X_n\}$ will be the prescribed one. On the other hand we want that

$$2S_{n/2} - S_n = \sum_{X_i \in A_1} X_i - \sum_{X_i \in A_2} X_i \text{ be near } 2T_{n/2} - T_n .$$

We choose sets A_1 and A_2 with an adaptation of the quantile transform techniques to this case.

Let us define for every subset $H = \{i_1, \ldots, i_{n/2}\}$ of the set $I = \{1, 2, \ldots, n\}$ containing $n/2$ elements the number $U_H(\omega) = \sum_{i \in H} X_i(\omega) - \sum_{i \in I-H} X_i(\omega)$. Let us put them into increasing order

$$U_{H_1} \leqslant U_{H_2} \leqslant \cdots \leqslant U_{H_p} \qquad p = \binom{n}{n/2}.$$

If two sums are equal, we order them at random. We define the disjoint intervals

$$I_1 = (-\infty, a_1), \qquad I_2 = [a_1, a_2), \ldots, I_p = [a_{p-1}, \infty)$$

by the equalities

$$\int_{I_k} \frac{1}{(2\pi)^{1/2}} e^{-t^2/2}\, dt = \frac{1}{p}.$$

If $1/n^{1/2}(2T_{n/2} - T_n) \in I_k$, we choose the X_i's with indices in H_k as the set A_1, and the X_i's with indices in $I - H_k$ as the set A_2. It is clear that we choose every subset of $\{X_1, ..., X_n\}$ with $n/2$ elements with equal probability to A_1.

If the distribution $P(U_{H_1}) = P(U_{H_2}) = \cdots = P(U_{H_p}) = 1/p$ is close to the normal distribution for arbitrary $\{X_1(\omega), ..., X_n(\omega)\}$, then $2S_{n/2} - S_n$ will be near to $2T_{n/2} - T_n$.

In the next step we may halve A_1 and A_2 in the same way, and we may go on till we have completely defined the new permutation.

The following theorem helps us to prove that the above construction gives a good approximation.

THEOREM (9.2). *We are given* $2N$ *real numbers* $x_1, x_2, ..., x_{2N}$ *satisfying*

$$\max |x_i| \leqslant K \qquad \text{and} \qquad \sigma^2 = \sum (x_i - \bar{x})^2 > cN, \qquad \bar{x} = \frac{1}{2N} \sum x_i .$$

Consider a random permutation π *of the indices* i, *where each permutation of the numbers* $1, 2, ..., 2N$ *is chosen with the same probability. Define the random sum*

$$U = (x_{\pi(1)} + \cdots + x_{\pi(N)}) - (x_{\pi(N+1)} + \cdots + x_{\pi(2N)}).$$

We have

$$P(U > x(N)^{1/2}) = \left(1 - \phi\left(\frac{N^{1/2}}{\sigma} x\right)\right) \exp\left[O\left(\frac{x^3 + 1}{N^{1/2}}\right)\right]$$

$$P(U < -x(N)^{1/2}) = \phi\left(-\frac{N^{1/2}}{\sigma} x\right) \exp\left[O\left(\frac{x^3 + 1}{N^{1/2}}\right)\right]$$

for all $0 \leqslant x \leqslant \epsilon(N)^{1/2}$ *with* $O(\cdot)$ *uniform in* x. ϵ *depends only on* K *and* c.

Let us briefly indicate why such a theorem holds. If we take pairs (x_{i_1}, x_{i_2}), $(x_{i_3}, x_{i_4}), ..., (x_{i_{2N-1}}, x_{i_{N2}})$ in every possible way, and consider the expression

$$\sum_{j=1}^{N} \epsilon_j (x_{i_{2j-1}} - x_{i_{2j}}),$$

where $\epsilon_1, \epsilon_2, ..., \epsilon_N$ are i.i.d.r.v.'s $P(\epsilon_1 = 1) = P(\epsilon_1 = -1) = \frac{1}{2}$ then U is the average of such expressions. On the other hand, every such expression is asymptotically normal. Their variances are different for different pairings, but typically close to σ, therefore the distribution of U, which is the mixture of these distributions, is also asymptotically normal.

We described how a construction satisfying Theorem (6.2) can be obtained in the cases when either $F(x)$ has an absolute continuous component, or $F(x)$ is concentrated on a finite interval.

The general case can be reduced to these two special situations with the help of the following simple theorem.

THEOREM (9.3). *Given the distribution functions $F_1(x), F_2(x)$ and $G_1(x)$, $G_2(x)$ let $S_1^{(i)}, S_2^{(i)},\ldots$ resp. $T_1^{(i)}, T_2^{(i)},\ldots$ be the partial sums of i.i.d.r.v.'s with distribution functions $F_i(x)$ resp. $G_i(x)$, $i = 1, 2$. For any $0 \leqslant p \leqslant 1$ there are two sequences $S_1, S_2, \ldots$ and $T_1, T_2, \ldots$ which are the partial sums of i.i.d.r.v.'s with distribution function $pF_1(x) + (1-p)F_2(x)$ resp. $pG_1(x) + (1-p)G_2(x)$ and satisfy the inequality*

$$P(\sup_{k\leqslant n} | S_k - T_k | > a + b) \leqq P(\sup_{k\leqslant n} | S_k^{(1)} - T_k^{(1)} | > a) + P(\sup_{k\leqslant n} | S_k^{(2)} - T_k^{(2)} | > b)$$

for all $a \geqslant 0$, $b \geqslant 0$, and n.

If $F(x)$ has no moment-generating function, then essentially the same construction works combined with some truncation. As the central limit theorem holds only in a smaller range in this case, we get a weaker approximation. The construction satisfying Theorem (7.1) can be done very similarly.

We have a Brownian bridge $B(t)$, and we have to put n points on the interval $[0, 1]$ in an appropriate way. We decide, with the help of $B(\frac{1}{2})$, how many of them to put in $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1]$, then, with the help of $2B(\frac{1}{4}) - B(\frac{1}{2})$ and $2B(\frac{3}{4}) - [B(1) - B(\frac{1}{2})]$, how many of them to put on the intervals $[0, \frac{1}{4})$, $[\frac{1}{4}, \frac{1}{2})$, $[\frac{1}{2}, \frac{3}{4})$, $[\frac{3}{4}, 1]$, etc. till we arrive at sufficiently small intervals.

10. HISTORY OF THE PROBLEM: COMMENTS

It was first observed by Erdös and Kac [11, 12] that functionals of sums of independent random variables have a limit distribution independent of the initial distribution, and that this can be directly proved. Another great impact on the theory of invariance principle was made by a paper of Doob [9]. Doob showed that the standardized empirical distribution function behaves similarly to the Brownian bridge. Later Donsker [8] justified this approach. These results gave the idea to work out the theory of invariance principle which was done by Prochorov [17], Skorochod [18], and others. This was done in the terms of probability measures. A very nice and readable work in this subject is Billingsley's book [4].

Theorem (4.1) was proved for a complete separable metric spaces by Skorochod [18] and Theorem (4.2) by Strassen [20]. Strassen applied the Banach–Hahn theorem. Later Dudley [10] found an elegant and simple proof of these theorems. He showed that theorem (4.2) is an almost immediate consequence of a famous combinatorial result, the König–Hall theorem (often called the marriage problem). Let us remark that Skorochod's original construction to Theorem (4.1) cannot be extended to noncomplete metric spaces.

Theorem (5.1) was proven by Strassen [21], who applied it to prove the following strengthened form of the law of iterated logarithm. Let $X_1, X_2, \ldots$ be i.i.d.r.v.'s $EX_1 = 0$, $EX_1^2 = 1$. Define the random polygon $S(t)$, $0 \leqslant t < \infty$ in the following way: $S(0) = 0$, $S(k) = X_1 + \cdots + X_k$, $k = 1, 2, \ldots$ and $S(t)$ is defined by linear interpolation between the integers. Consider the sequence $\tilde{S}_n(t) = S(nt)/(2n \log \log n)^{1/2}$, $n = 3, 4, 5$; $0 \leqslant t \leqslant 1$ as elements of the $C[0, 1]$ space. Strassen's theorem states the following:

The sequence $\tilde{S}_3(t), \tilde{S}_4(t), \ldots$ is relatively compact with the set of limit points H, $H = \{x(t)\colon \int_0^1 \dot{x}(t)^2\,dt \leqslant 1,\ x(0) = 0,\ x(t)$ is absolutely continuous$\}$ with probability 1. (This is meant in the $C[0, 1]$ space with the usual supremum metric.) This topic later became quite popular. Much work was done to describe more general situations. We would refer to the works of Oodaira [16] and Wichura [24].

The "stochastic geyser problem" was proposed by A. Rényi and solved by Bártfai [1]. There is a tale connected to it, which explains the name of the problem. The tale goes like this:

A shipwrecked man gets to an uninhabited island. There is a geyser on this island. The time intervals between subsequent outbursts are independent and identically distributed random variables.

Our man devotes all his life to the observation of the geyser. Having no watch he cannot measure the exact time of the outbursts, but seeing the sun rise and set, he can count how many outbursts took place each day. He leaves the record of his long, long observations to posterity. Can we determine the distribution of the time intervals between subsequent outbursts with the help of it?

In a more formal way, the question is the following: Let S_k denote the time of the kth outburst. (It is the sume of k i.i.d.r.v.'s.) Knowing the sequence $S_k + \epsilon_k$, $k = 1, 2, \ldots$; $\epsilon_k = (S_k] - S_k$ ([ ] means integer part) can we determine the unknown distribution of S_1?

Bártfai also recognized that the stochastic geyser problem is the converse of the problem about approximation of partial sums of i.i.d.r.v.'s. Therefore he investigated the latter problem, too, to get an upper bound in the stochastic geyser problem. Using the quantile transform technique he got essentially $O(n^{1/4})$ [2]. Then he made the brave remark that despite of this result, he believed, the result $O(\log n)$ in the geyser problem is sharp.

There are several methods for solving the stochastic geyser problem. One common feature of all of them is that they are based on large deviation results. All the "estimates of classical type" fail to work.

Theorem (5.3) was obtained by Breiman [6]. There seems to be some interesting philosophy behind this simple result. It is known that in many problems the sums of independent random variables do not behave so nicely if they have only few moments. The proof of Theorem (5.3) (and some other results too) would suggest that this bad behavior is caused by some individual terms that take up exceptionally big values.

The approximation $| S_k - T_k | = o(n^{1/(6-\epsilon)})$ under the condition $EX_1^3 = 0$

was obtained by Révész and M. Csörgő [7]. They applied quantile transform techniques.

Theorems (6.1), (6.2), (6.3), (6.4), (6.5) and also Theorems (7.1) and (7.3) were obtained in a paper consisting of two parts, written by Komlós, Major and Tusnády [13]. The case when the third moment does not exist was omitted because of some technical problems. This gap was filled by a paper of Major [14]. A result similar to Theorem (6.5) in the case when the third moment does not exist was proven earlier by Borovkov [5]. Theorem (6.6) was proved by Major [15].

The results of Sections 6 and 7 also yield an estimate about the speed of convergence of distribution of a functional applied to partial sums or empirical distribution function. In "nice" cases they give a rate of convergence $O((\log n)/n^{1/2})$. In many cases this is the best result available; all the same, the $\log n$ factor seems superficial. There must exist a theorem stating: For every "nice" functional $\mathscr{F}$ the distributions $P(\mathscr{F}S_n(t) < x)$ or $P(\mathscr{F}n^{1/2}[F_n(t) - t] < x)$ have a distance $O(1/n^{1/2})$ from their limits. Probably there is even an Edgeworth type expansion for these distributions by the powers of $1/n^{1/2}$.

What this theorem should look like is not at all clear. All the same, this still undiscovered theorem may be the reason why all the "classical estimates" fail to solve the stochastic geyser problem.

Another application of the results in Sections 6 and 7 may be the determination of the distribution of $\mathscr{F}_nS_n(t)$ and $\mathscr{F}_nn^{1/2}[F_n(t) - t]$; i.e., the case when also the functional is changing with n. If we have a good approximation $S_n(t) \sim W(t)$ or $n^{1/2}[F_n(t) - t] \sim B(t)$ we may investigate the sequence $\mathscr{F}_nW(t)$ or $\mathscr{F}_nB(t)$ instead of the original problem. This idea was exploited by Bickel and Rosenblatt [3] in an investigation about density function estimates. Their method shows how useful it would be in many applications to find the multidimensional analog of Theorem (7.1).

The exact formulation of this problem is the following: Let $X_1, X_2, \ldots, X_n$ be i.i.d.r.v.'s uniformly distributed on the d-dimensional unit cube. Define the empirical distribution function $F_n(x_1, \ldots, x_d) = 1/n$ (the number of X_i's in the cube $[0, x_1) \times \cdots \times [0, x_d)$). There is an obvious candidate among the Gaussian processes $B(t_1, \ldots, t_d)$ which one can approximate $n^{1/2}[F_n(t_1, \ldots, t_d) - t_1t_2 \cdots t_d]$. Question: What is the magnitude of

$$\sup_{t_1, \ldots, t_d} [B(t_1, \ldots, t_d) - n^{1/2}[F_n(t_1, \ldots, t_d) - t_1t_2, \ldots, t_d]]$$

in the case of the optimal approximation? Tusnády [22] has proved that in the case $d = 2$ an approximation with magnitude $O(\log^2 n)/n^{1/2})$ is possible. It is not known whether $\log^2 n$ can be substituted by $\log n$. The case of $d > 2$ is even less known. Csörgő and Révész—using quantile transform techniques—proved that a construction yielding $O(n^{-1/(2(k+1))} \log^{3/2} n)$ [7] (d is the dimension of the space) is possible. But the real question is whether the order of magnitude in the optimal construction is $O((\log^\alpha n)/n^{1/2})$ or $\alpha \geqslant 1$, $-\frac{1}{2} < \beta < 0$. We know nothing about this problem.

Another interesting and still unsolved problem is to prove the analog of Theorem (6.2) for independent non-identically distributed random variables.

Let us give an example, that shows that there are cases when a simple adaptation of the present technique does not work.

Let the sequence $\alpha_1, \alpha_2, \ldots, 2 \leqslant \alpha_i \leqslant 4$, be real numbers linearly independent over the integers, i.e., let the relation $\sum_{i=1}^n k_i\alpha_i = 0$ (k_i's are integers) imply that $k_1 = k_2 = \cdots = k_n = 0$. Let the random variable X_k have the distribution

$$P(X_k = \alpha_k) = P(X_k = -\alpha_k) = \frac{1}{2\alpha_k{}^2}, \qquad P(X_k = 0) = 1 - \frac{1}{\alpha_k{}^2}.$$

Now if we know the value of the sum of independent random variables $S_n = X_1 + \cdots + X_n$, then, because of the linear independence of the α_k's, we also know the values of $X_1, X_2, \ldots, X_n$. Thus the conditional quantile transform technique—at least in its original form—fails to work.

Theorem (8.1) was proved in the special case $f(x) = x^2$ by Bártfai [2] and in the case $f(x) = |x|$ by Vallander [23]. The author of these notes learned about its general form from a lecture of Stout. As he had not seen the original proof, he does not know whether it differs from the present one.

The quantile transform technique seems to be rather a folklore. It was discovered by many authors indepedently of each other. There was however a period, when it was (unjustly) neglected, and the Skorochod embedding was preferred instead. The conditional quantile transform technique was worked out in the paper of Komlós, Major, and Tusnády [13].

References

[1] Bártfai, P. (1966). Die Bestimmung der zu einem widerkehrenden Prozess gehörenden Verteilungsfunktion aus den mit Fehlern behafteten Daten einer einzigen Realisation. *Studia Sci. Math. Hungar.* **1** 161–168.

[2] Bártfai, P. (1970). Über die Entfernung der Irrfahrtswege. *Studia Sci. Math. Hung.* **5** 41–49.

[3] Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095.

[4] Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York/London/Sydney/Toronto.

[5] Borovkov, A. A. (1973). On thc rate of convergence for the invariance principle (in Russian), *Theor. Probability Appl.* **18** 217–234.

[6] Breiman, L. (1967). On the tail behaviour of sums of independent random variables. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **9** 20–24.

[7] Csörgö, M. and Révész, P. (1975). A new method to prove Strassen type laws of Invariance Principle, I. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **31** 255–259; II 261–269.

[8] Donsker, M. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov–Smirnov theorems. *Ann. Math. Statist.* **23** 277–281.

[9] Doob, J. L. (1949). Heuristic approach to the Kolmogorov–Smirnov theorems. *Ann. Math. Statist.* **20** 393–403.

[10] Dudley, R. M. (1968). Distances of probability measures and random variables. *Ann. Math. Statist.* **39** 1563–1572.

[11] Erdős, P., and Kac, M. (1946). On certain limit theorems in the theory of probability. *Bull. Amer. Math. Soc.* **52** 292–302.

[12] Erdős, P., and Kac, M. (1947). On the number of positive sums of independent random variables. *Bull. Amer. Math. Soc.* **53** 1011–1020.

[13] Komlós, J., Major, P., and Tusnády, G. (1975, 1976). An approximation of partial sums of independent RV's and the sample DF. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete.* I **32** 111–131; II **34** 33–58.

[14] Major, P. (1976). The approximation of partial sums of independent RV's. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **35** 213–220.

[15] Major, P. (1978). An improvement of Strassen's invariance principle. *Ann. Probability*, to appear.

[16] Oodaira, H. (1973). The law of iterated logarithm for Gaussian processes. *Ann. Probability* **1** 964–967.

[17] Prochorov, Ju. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theor. Probability Appl.* **1** 157–314.

[18] Skorochod, A. V. (1956). Limit theorems for stochastic processes. *Theor. Probability Appl.* **1** 261–290.

[19] Strassen, V. (1964). An invariance principle for the law of iterated logarithm. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **3** 211–226.

[20] Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.* **36** 423–439.

[21] Strassen, V. (1967). Almost sure behaviour of sums of independent random variables and martingales. *Proc. 5th Berkeley Sympos. Math. Statist. Probab.*, 1965, Vol. II, Part 1, pp. 315–343, Univ of California Press, Berkeley.

[22] Tusnády, G. (1977). A remark on the approximation of the sample DF in the multidimensional case. *Periodica Math. Hungar.* **8** 53–55.

[23] Vallander, S. S. (1973). Calculation of Wassenstein distance between probability distributions on the line (in Russian). *Theor. Probability Appl.* **18** 824–827.

[24] Wichura, M. J. (1973). Some Strassen type laws of the iterated logarithm for multiparameter stochastic processes. *Ann. Probability* **1** 272–296.