Studia Scientiarum Mathematicarum Hungarica

AUXILIO
CONSILII INSTITUTI MATHEMATICI
ACADEMIAE SCIENTIARUM HUNGARICAE

TOMUS II.
FASC. 1–2.
1967

AKADÉMIAI KIADÓ, BUDAPEST

Studia Scientiarum Mathematicarum Hungarica

A Magyar Tudományos Akadémia matematikai folyóirata

Szerkesztőség: Budapest V., Reáltanoda u. 13—15.

Technikai szerkesztő: Katona Gy.

Kiadja az Akadémiai Kiadó, Budapest V., Alkotmány u. 21.

A *Studia Scientiarum Mathematicarum Hungarica* angol, német, francia vagy orosz nyelven közöl eredeti értekezéseket a matematika tárgyköréből. Félévenként jelenik meg, évi egy kötetben.

Előfizetési ára belföldre 120,— Ft, külföldre 165,— Ft. Megrendelhető a belföld számára az Akadémiai Kiadónál, a külföld számára pedig a Kultúra Könyv és Hírlap Külkereskedelmi Vállalatnál (Budapest II., Fő u. 32).

Cserekapcsolatok felvétele ügyében kérjük az MTA Matematikai Kutató Intézete Könyvtárához (Budapest V., Reáltanoda u. 13—15) fordulni.

Közlésre szánt dolgozatokat kérjük két példányban a szerkesztőség címére küldeni.

Studia Scientiarum Mathematicarum Hungarica is a journal of the Hungarian Academy of Sciences publishing original papers on mathematics, in English, German, French or Russian.

It is published semiannually, making up one volume per year.

Editorial Office: Budapest V., Reáltanoda u. 13—15, Hungary.
Technical Editor: Gy. Katona

Subscription rate: Ft 165 per volume. Orders may be placed with *Kultúra* Trading Co. for Books and Newspapers, Budapest 62, P. O. B. 149 or with its representatives abroad.

For establishing exchange relations please write to the Library of the Mathematical Institute (Budapest V., Reáltanoda u. 13—15).

Papers intended for publication should be sent to Editor in 2 copies.

Studia Scientiarum Mathematicarum Hungarica

Auxilio
Consilii Instituti Mathematici
Academiae Scientiarum Hungaricae

Redigit

A. Rényi

Adiuvantibus

M. Arató, L. Fejes Tóth, T. Frey, G. Freud, L. Kalmár,
A. Prékopa, K. Tandori

Tomus II

Akadémiai Kiadó, Budapest
1967

STUDIA SCIENTIARUM MATHEMATICARUM HUNGARICA

Tomus 2

INDEX

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 1—6.

NOTES ON A NONLINEAR INTEGRAL EQUATION

by

I. BIHARI

1. Introduction

Certain investigations concerning the stability and asymptotic behaviour of the solutions of the *system*

$$\dot{x} = Ax + f(t, x), \quad A = \text{const}, \ f(t, 0) = 0, \ t \geqq 0 \tag{1}$$

lead to the integral equation

$$x(t) = y(t) + \int_0^t Y(t-s) f(s, x(s))\, ds, \quad y(t) = Y(t)a, \quad x(0) = y(0) = a \tag{2}$$

equivalent to (1). Here

$$\dot{Y} = AY, \quad Y(0) = I. \tag{3}$$

Let every solution of

$$\dot{y} = Ay \tag{4}$$

be bounded for $t \geqq 0$, i.e. the real parts of the characteristic roots λ of A are non-positive and the roots of the minimal polynom of A are simple. Unifying every block of Y with $\operatorname{Re} \lambda < 0$ in a hyperblock Z_1 and those with $\operatorname{Re} \lambda = 0$ in a hyperblock Z_2, Y can be decomposed as

$$Y = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} = Y_1 + Y_2 \tag{5}$$

where $Y_1 \to 0$ as $t \to +\infty$, while Y_2 is bounded for every t (for $t < 0$, too). Making use of (5) equation (2) may be replaced by

$$x(t) = z(t) + \int_0^t Y_1(t-s) f(s, x(s))\, ds - \int_t^\infty Y_2(t-s) f(s, x(s))\, ds \tag{6}$$

$$z(t) = Y(t)c, \quad c = a + b, \quad b = \int_0^\infty Y_2(-s) f(s, x(s))\, ds.$$

This equation has been investigated and solved in [1], in connection with asymptotic problems. The present paper deals with the following generalization of (6)

$$x(t) = z(t) + \int_0^t k_1(t, s) f_1(s, x(s))\, ds + \int_t^\infty k_2(t, s) f_2(s, x(s))\, ds \tag{7}$$

the solutions of which will be studied in certain function space. Here $k_1(t, s)$ and $k_2(t, s)$ are real $n \times n$ matrices, continuous on $0 \leqq s \leqq t < \infty$ and $0 \leqq t \leqq s < \infty$, respectively. $f_i(t, x)$, $x(t)$, $z(t)$ are vectors in R^n.

Equation (7) without the second integral has been investigated by CORDUNEANU [2] who obtained considerable results which may be easily extended to (7) (at least if $f_1 \equiv f_2 \equiv f$ [1]) only his operator $Ux(t)$ must be replaced here by

$$Ux(t) = z(t) + \int_0^t k_1(t, s) f(s, x(s))\, ds + \int_t^\infty k_2(t, s) f(s, x(s))\, ds \tag{8}$$

which is a compact (completely continuous) map assuming suitable conditions (s. later).

2. As an extension of the results of A. STOKES [3] the statements of [1] can be reestablished and generalized under less restrictive conditions. However, this will be done on the loss of uniqueness of the solutions of (6) and (7). Also it fails now the possibility of a successive approximation-process.

The tool used here is the fixed point theorem of SCHAUDER—TYCHONOV in the form as follows:

Let E be a complete, locally convex, linear topological space. Suppose A is a closed, convex subset of E and $U: E \to E$ a continuous map of E into itself such that $U(A)$ is relatively compact in E. Then U admits at least one fixed point in A.

THEOREM 1. *Let the following conditions be satisfied:*

1° *$k_1(t, s)$ and $k_2(t, s)$ are continuous and bounded in $0 \leqq s \leqq t < \infty$ and $0 \leqq t \leqq s < \infty$ respectively, i.e. $\|k_i(t, s)\| \leqq K_i$, $i = 1, 2$,* [2]

2° *$f(t, x)$ continuous for $t \geqq 0$, $x \in R^n$ and*

$$\| f(t, x) \| \leqq G(t, \|x\|), \quad t \geqq 0, \; x \in R^n$$

where $G(t, r)$ is piecewise continuous in $t \geqq 0$, $r \geqq 0$ and non-decreasing (for fixed t) with respect to r,

$$3° \qquad \gamma + K_1 \int_0^t G(s, g(s))\, ds + K_2 \int_t^\infty G(s, g(s))\, ds \leqq g(t) \tag{3}$$

for certain function $g(t) \geqq 0$ continuous for $t \geqq 0$ and arbitrary constant $\gamma > 0$,

4° *$z(t)$ is bounded continuous for $t \geqq 0$, $(z(t) \in C(R_+))$, then equation (7) has at least one solution $x(t)$ continuous for $t \geqq 0$ satisfying $\|x(t)\| \leqq g(t)$, $t \geqq 0$.*

PROOF. Let $E = C_c(R_+, R^n)$ be chosen, where C_c means the space of the functions defined and continuous on $R_+ : t \geqq 0$ with their values in R^n. The topology of this space is the uniform convergence on every compact set of R_+. C_c is a complete, locally convex, linear topological space.

[1] In the sequel this will be assumed.

[2] As norm of $x = (x_1, \ldots, x_n)$ and of $A = (a_{ik})$ it will be taken $\|x\| = \sum_i |x|$ and $\|A\| = \sum_{i,k} |a_{ik}|$, respectively.

[3] Condition 3° implies the existence of the second integral.

The operator $Ux(t)$ defined by (8) is a compact one (in the topology of C_c) on the subset $A \subset C_c$ defined by

$$A = \{x \in C_c : \|x(t)\| \leqq g(t),\ t \geqq 0\}. \tag{9}$$

Clearly, A is closed convex and bounded (in the topology of C_c, i. e. in every compact set of R_+).

The existence of Ux for $x \in A$ is consequence of 3° and 2°. The continuity of Ux on A means that

$$x_n, x \in A,\ x_n \to x \quad \text{implies} \quad Ux_n \to Ux$$

for every finite interval $a \leqq t \leqq b$. But (8) gives

$$Ux - Ux_n = \int_0^t k_1(t, s)\left[f(s, x(s)) - f(s, x_n(s))\right] ds +$$

$$+ \int_t^\infty k_2(t, s)\left[f(s, x(s)) - f(s, x_n(s))\right] ds$$

whence

$$\|Ux - Ux_n\| \leqq K_1 \int_0^t \|f(s, x(s)) - f(s, x_n(s))\| ds + \tag{10}$$

$$+ K_2 \int_t^T \|f(s, x(s)) - f(s, x_n(s))\| ds + 2K_2 \int_T^\infty G(s, g(s))\, ds, \quad T \geqq t.$$

Choosing T so large that the last term is less than $\varepsilon/2$ (where $\varepsilon > 0$) and $N > 0$ so that the first two integrals together in (10) are less than $\varepsilon/2$ for $n > N$ (on account of the uniform continuity of $f(t, x)$ in every finite and closed domain), we have

$$\|Ux - Ux_n\| \leqq \varepsilon,\ n > N.$$

Being A bounded and U continuous the set $U(A)$ is (uniformly) bounded and by 3° $U(A) \subset A$ if $\|z(t)\| \leqq \gamma$, $t \geqq 0$. To prove the compactness of $U(A)$ it remains to show the equicontinuity of the functions $Ux(t)$.

Concerning the difference

$$Ux(t) - Ux(\tau) = z(t) - z(\tau) + \int_0^t [k_1(t, s) - k_1(\tau, s)] f(s, x(s))\, ds +$$

$$+ \int_t^\infty [k_2(t, s) - k_2(\tau, s)] f(s, x(s))\, ds + \int_\tau^t [k_1(\tau, s) - k_2(\tau, s)] f(s, x(s))\, ds$$

we have

$$\|Ux(t) - Ux(\tau)\| \leqq \|z(t) - z(\tau)\| + \int_0^t \|k_1(t, s) - k_1(\tau, s)\| G(s, g(s))\, ds + \tag{11}$$

$$+ \int_t^\infty \|k_2(t, s) - k_2(\tau, s)\| G(s, g(s))\, ds + \left| \int_\tau^t \|k_1(\tau, s) - k_2(\tau, s)\| G(s, g(s))\, ds \right|.$$

Let $a>0$ be a fixed number, $0\leqq t, \tau\leqq a$, $G(s, g(s))\leqq G_a$ in $0\leqq s\leqq a$. Then, on account of the uniform continuity of the functions continuous on a finite closed domain, to every $\varepsilon>0$ there is a $\delta_1>0$ such that

$$\|z(t)-z(\tau)\| \leqq \frac{\varepsilon}{5}, \quad \|k_1(t, s)-k_1(\tau, s)\| \leqq \frac{\varepsilon}{5aG_a}$$

provided that $|t-\tau|\leqq\delta_1$, $0\leqq s\leqq a$ and

$$\int_\tau^t \|k_1(\tau, s)-k_2(\tau, s)\| G(s, g(s))\, ds \leqq (K_1+K_2)G_a|t-\tau| \leqq \frac{\varepsilon}{5}$$

if $|t-\tau|\leqq\delta_2=\dfrac{\varepsilon}{5(K_1+K_2)G_a}$. Then the first, second and the fourth terms on the right of (11) are less than $\varepsilon/5$. In the third term we write $\int_t^\infty = \int_t^T + \int_T^\infty$ $(T\geqq a)$. The value of T may be chosen that

$$\int_T^\infty \|k_2(t, s)-k_2(\tau, s)\| G(s, g(s))\, ds \leqq (2K_2)\int_T^\infty G(s, g(s))\, ds \leqq \frac{\varepsilon}{5}.$$

To this T a number $\delta_3>0$ can be found that

$$\int_t^T \|k_2(t, s)-k_2(\tau, s)\| G(s, g(s))\, ds \leqq \frac{\varepsilon}{5}$$

be fulfilled provided $|t-\tau|\leqq\delta_3$, viz. $k_2(t, s)$ is uniformly continuous in $0\leqq t,\ \tau\leqq T$. Taking all these into account, the right member of (11) will be less than ε if $\|t-\tau\|\leqq$ $\leqq\min(\delta_1, \delta_2, \delta_3)$, which is equivalent to the stated equicontinuity. Therefore the closure of $U(A)$ in C_c is compact.

Let us apply Theorem 1 to equation (6). Now

$$\|Y_1(t)\|\leqq c_1 e^{-\alpha t},\ t\geqq 0,\ \|Y_2(t)\|\leqq c_2,\ t\gtreqless 0,\ \alpha>0$$

Specify now $G(t, r)$ as $G=h(t)\omega(r)$ where $\omega(r)\geqq 0$ is piecewise continuous, non-decreasing and

$$\text{(C)} \qquad \int_0^u \frac{dr}{\omega(r)} = \infty (u>0), \quad \int^\infty \frac{dr}{\omega(r)} = \infty, \quad \int_0^\infty h(t)\, dt < \infty$$

then condition 3° reads as

$$\text{(12)} \qquad \gamma + c_1\int_0^t h(s)\omega(g(s))\, ds + c_2\int_t^\infty h(s)\omega(g(s))\, ds \leqq g(t).$$

It will be shown that (12) has a solution $g(t)\in C$ for every $\gamma>0$. Letting

$$\Omega(u)=\int_{u_0}^{u}\frac{dr}{\omega(r)}\ (u_0>0),\quad \varrho_i=c_i q,\quad i=1,2,\quad q=\int_0^{\infty}h(t)\,dt$$

and defining the number $a>0$ — if possible — by the equation

$$\text{(D)}\qquad \Omega\left(\gamma+\frac{\varrho_2}{\varrho_1}a\right)-\Omega(\gamma+a)=\int_{\gamma+a}^{\gamma+\frac{\varrho_2}{\varrho_1}a}\frac{dr}{\omega(r)}=\varrho_2-\varrho_1$$

$\left(\text{that exists certainly for } \omega(r)\equiv r.\ \text{Its value is } a=\gamma\varrho_1\,\frac{e^{\varrho_1}-e^{\varrho_2}}{\varrho_1 e^{\varrho_2}-\varrho_2 e^{\varrho_1}}\right)$ the function

$$\text{(13)}\qquad g(t)=\Omega^{-1}\left(\Omega(\gamma+a)+(c_2-c_1)\int_t^{\infty}h(s)\,ds\right)$$

satisfies (12) with the sign $=$, provided that $c_1\neq c_2$. This assertion may be verified (like a similar one in [1]) in the following way. By (13)

$$\Omega\big(g(t)\big)=\Omega(\gamma+a)+(c_2-c_1)\int_t^{\infty}h(s)\,ds$$

$$\frac{d\Omega\big(g(t)\big)}{dt}=\frac{g'(t)}{\omega\big(g(t)\big)}=-(c_2-c_1)h(t)\quad\text{or}\quad g'(t)=-(c_2-c_1)h(t)\omega\big(g(t)\big).$$

Therefore

$$I(t)\equiv\gamma+c_1\int_0^t h(s)\omega\big(g(s)\big)\,ds+c_2\int_t^{\infty}h(s)\omega\big(g(s)\big)\,ds=$$

$$=\gamma-\frac{c_1}{c_2-c_1}\int_0^t g'(s)\,ds-\frac{c_2}{c_2-c_1}\int_t^{\infty}g'(s)\,ds=\gamma+g(t)+\frac{c_1 g(0)-c_2 g(\infty)}{c_2-c_1}.$$

Here we have by (D)

$$g(0)=\Omega^{-1}\big(\Omega(\gamma+a)+(c_2-c_1)g\big)=\gamma+\frac{\varrho_2}{\varrho_1}a$$

$$g(\infty)=\gamma+a$$

thus $I(t)=g(t)$ as stated. — For $\omega(r)\equiv r$

$$g(t)=\gamma e^{\varrho_1}\,\frac{\varrho_1-\varrho_2}{\varrho_1 e^{\varrho_2}-\varrho_2 e^{\varrho_1}}\exp\left[(c_2-c_1)\int_t^{\infty}h(s)\,ds\right].$$

In the case $c_1 = c_2$, $g(t) = \gamma + a$ (where a satisfies $a = \varrho\omega(\gamma + a)$ if possible) and for $\omega(r) \equiv r$, $g(t) = \dfrac{\gamma}{1-\varrho}$ provided $\varrho < 1$, where $\varrho = \varrho_1 = \varrho_2$ (s. [1]).

The assumptions $\Omega(\infty) = \infty$, $\Omega(0) = -\infty$, i.e. condition (C) are necessitated by the circumstance that without them the argumentum of Ω^{-1} in (12) can leave the domain of Ω^{-1}. The foregoing may be summarized as

THEOREM 1′. *If in equation* (6)

1° $z(t) \in C(R_+, R^n)$ *i.e.* $z(t)$ *is continuous and bounded for* $t \in R_+$,

2° $Y_1(t)$ *and* $Y_2(t)$ *are like in the Introduction*,

3° $f(t, x)$ *is continuous and* $\| f(t, x) \| \leqq h(t)\omega(\|x\|)$, $t \in R_+$, $x \in R^n$,

4° $h(t)$ *and* $\omega(r)$ *satisfy conditions* (C) *and* (D),

then (6) *has a solution (or several solutions)* $x \in A$ *where* $g(t)$ *is given by* (13). *As a consequence* $x(t) - z(t) \to 0$, $t \to \infty$ (s. [1]).

REFERENCES

[1] BIHARI, I.: The asymptotic behaviour of a system of nonlinear differential equations, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **8** A (1963) 475—88.

[2] CORDUNEANU, C.: Problèmes globaux dans la théorie des équations integrales de Volterra, *Ann. Mat. Pura Appl.* **4** (67) (1965) 349—64.

[3] STOKES, A.: The applications of a fixed point theorem to a variety of non-linear stability problems, *Contributions to differential equations, V*, Princeton Univ. Press. Princeton, 1960.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDEPEST

(Received February 20, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 7—21.

ON THE DETERMINANT OF (0,1) MATRICES

by

J. KOMLÓS

I. Introduction

a) In the present paper we consider $n \times n$ matrices with elements 0,1 and our purpose is to investigate the number of all non-singular ones. We shall prove that the singular matrices form a negligible percent asymptotically. More precisely, we shall prove the following

THEOREM

Let A_n denote the number of $n \times n$ matrices with elements 0, 1 *having determinant* 0, *then*

$$\lim_{n=+\infty} \frac{A_n}{2^{n^2}} = 0.$$

b) In other words let us choose at random a matrix from the set of $n \times n$ (0, 1) matrices such that all matrices have the same probability (2^{-n^2}). If a_n means the probability of the event that the determinant of the chosen matrix equals 0, then $\lim_{n=+\infty} a_n = 0$. It is easy to see that the following fact is equivalent to our theorem:

If $\varepsilon_{i,j}$ are independent random variables which take the values 0 and 1 with probabilities $\frac{1}{2}, \frac{1}{2}$ and

$$p_n = \mathbf{P}\left(\begin{vmatrix} \varepsilon_{1,1} & \varepsilon_{1,2} \dots \varepsilon_{1,n} \\ \varepsilon_{2,1} & \varepsilon_{2,2} \dots \varepsilon_{2,n} \\ \dots\dots\dots\dots\dots \\ \varepsilon_{n,1} & \varepsilon_{n,2} \dots \varepsilon_{n,n} \end{vmatrix} = 0\right)$$

then

$$\lim_{n=+\infty} p_n = 0.$$

We shall use all versions at the same time. In the section **VI.** we deal with a generalization of this problem in the case of infinite matrices.

c) The proof goes as follows: We show that the probability of the event, that the rank of an $n \times n$ (0, 1) matrix is $k+2$, where k denotes the rank of the $(n-1) \times (n-1)$ matrix, consisting of its first $n-1$ rows and columns, or is equal to n, tends to 1 if $n \to \infty$.

Using this fact we prove that

$$\liminf_{n=+\infty} \frac{A_n}{2^{n^2}} = 0.$$

Having proved this, we prove the convergence of the sequence $A_n/2^{n^2}$.

Before the proof of the theorem we give some definitions and lemmas.

II. Definitions and Lemmas

a) Let $A_{n,k}$ denote the number of $n \times n$ (0, 1) matrices whose rank is equal to k. Clearly

$$A_n = \sum_{k=1}^{n-1} A_{n,k} = 2^{n^2} - A_{n,n}.$$

Then we have to prove that

$$\lim_{n=+\infty} \frac{A_{n,n}}{2^{n^2}} = 1.$$

First we give a known lemma.

LEMMA 1. *Let $a_1, a_2, \ldots, a_n$ be real numbers different from 0 and c an arbitrary real number, then at most $\binom{n}{\left[\frac{n}{2}\right]}$ among the sums $\sum_{i=1}^{n} \varepsilon_i a_i$ (ε_i is equal to 0 or 1) are equal to c.*

PROOF. Let us consider instead of the numbers $\sum_{i=1}^{n} \varepsilon_i a_i$ the sums $2 \cdot \sum_{i=1}^{n} \varepsilon_i a_i - \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \varphi_i a_i$, where $\varphi_i = 2\varepsilon_i - 1$, then φ_i is equal to 1 or -1 if ε_i is equal to 1 or 0, respectively. The sum $\sum_{i=1}^{n} \varepsilon_i a_i$ equals c if the sum $\sum_{i=1}^{n} \varphi_i a_i$ equals $d = 2c - \sum_{i=1}^{n} a_i$. Then we can reformulate the lemma so that the numbers ε_i are equal to 1 or -1. In this case we can suppose without violating the generality, that the numbers $a_1, a_2, \ldots, a_n$ are all positive.

Then it is enough to prove the following: if $a_1, a_2, \ldots, a_n$ are positive numbers and d is an arbitrary real number, then at most $\binom{n}{\left[\frac{n}{2}\right]}$ among the numbers $\sum_{i=1}^{n} \varepsilon_i a_i$ (ε_i equals 1 or -1) are equal to d.

Let us correspond for every sum $\sum_{i=1}^{n} \varepsilon_i a_i$ the set of those natural numbers i for which $\varepsilon_i = 1$ holds. If for two different sums $\sum_{i=1}^{n} \varepsilon_i a_i = \sum_{i=1}^{n} \varepsilon_i' a_i$, then the corresponding sets of the two sums cannot contain each other.

The Sperner-theorem implies that the number of sums equal to any constant is at most $\binom{n}{\left[\frac{n}{2}\right]}$.

Clearly we can formulate the lemma as follows: if $a_1, a_2, \ldots, a_m$ are real numbers, among which n are different from 0 and c is an arbitrary real number

then among the numbers $\sum_{i=1}^{n} \varepsilon_i a_i$ (ε_i equals 0 or 1) at most $\binom{n}{\left[\frac{n}{2}\right]} 2^{m-n} < \frac{2^m}{\sqrt{n}}$

are equal to c.

b)

DEFINITIONS.

A system of k linearly independent row (resp. column) vectors of a matrix of rank k is called a row (resp. column) basis of the matrix.

We shall use that any row (resp. column) vector is a uniquely determined linear combination of the vectors of any fixed row (resp. column) basis.

1) *The degree of a row (resp. column) vector with respect to a given row(resp. column) basis, is the number of those elements of the row (resp. column) basis, which have coefficients different from* 0 *in the above mentioned linear combination.*

2) *The degree of a row (resp. column) vector is the largest one among the degrees of this row (resp. column) vector with respect to all possible row (resp. column) basises.*

3) *The row (resp. column) degree of a matrix is the largest one among the degrees of its row (resp. column) vectors.*

LEMMA 2. *If the row-degree of an $m \times n$ (0, 1) matrix is l and its rank is k, then we can add to the matrix a column vector (with components* 0, 1*) so that the rank of the obtained $m \times (n+1)$ matrix is k again, at most $\frac{2 \cdot 2^m}{\sqrt{l}}$ different ways.*

PROOF. For the sake of simplicity let us suppose that the first k row vectors form the basis, with respect to which the degree of the t-th row vector is equal to l.

Let us denote the i-th row vector by $\mathbf{a}_i$, the j-th column vector by $\mathbf{b}_j$ and the additional (the $(n+1)$-th) column vector by $\mathbf{b}_{n+1}$, i. e.

$$\mathbf{a}_i = (a_{i,1}; a_{i,2}; \ldots; a_{i,n}),$$

$$\mathbf{b}_j = \begin{pmatrix} b_{1,j} \\ b_{2,j} \\ \vdots \\ b_{m,j} \end{pmatrix}, \quad \mathbf{b}_{n+1} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

The row vectors of the enlarged matrix are

$$\mathbf{a}_i' = (a_{i,1}; a_{i,2}; \ldots; a_{i,n}; b_i).$$

So we have $\mathbf{a}_t = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \ldots + c_k\mathbf{a}_k$ where among the constants c_i l are different from 0.

If the degree of the new $(m \times (n+1))$ matrix is also k then (because the maximal numbers of linearly independent row and column vectors are equal to each other and clearly $\mathbf{a}_1', \mathbf{a}_2', \ldots, \mathbf{a}_k'$ are also linearly independent)

$$\mathbf{a}_t' = c_1\mathbf{a}_1' + c_2\mathbf{a}_2' + \ldots + c_k\mathbf{a}_k'$$

hence

$$b_t = c_1b_1 + c_2b_2 + \ldots + c_kb_k.$$

But b_i is equal to 0 or 1 and among the numbers c_i l are different from 0, so by Lemma 1 we can choose the vector $(b_1, b_2, \ldots, b_k)$ at most $\frac{2^k}{\sqrt{l}}$ different ways such that $b_t = 0$ holds; similarly we can choose $(b_1, b_2, \ldots, b_k)$ at most $\frac{2^k}{\sqrt{l}}$ different ways such that $b_t = 1$ holds. That is, we have at most $\frac{2 \cdot 2^m}{\sqrt{l}}$ possibilities to choose the vector $\mathbf{b}_{n+1}$. Q. e.d.

Similarly, if the column-degree of a matrix is l, then we can construct to the matrix a row vector at most $\frac{2 \cdot 2^n}{\sqrt{l}}$ different ways such that the maximal numbers of linearly independent vectors of both matrices are equal to each other.

c)

LEMMA 3. *By k m-dimensional vectors (with elements 0, 1) we can construct at most 2^{2^k} different vectors (with components 0, 1) with linear combinations.*

PROOF. Let us consider a $k \times m$ matrix with row vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$. It contains at most 2^k different column vectors (because it has only 0 or 1 components). If the i_1-th, i_2-th, ..., i_t-th column vectors are the different ones ($t \leqq 2^k$), so any of the others is equal to one of these, then in the linear combinations of the row vectors the i_1-th, i_2-th, ..., i_t-th components can arbitrarily vary. Then among the linear combinations, whose components are 0, 1, at most $2^t \leqq 2^{2^k}$ can be different. Q. e. d.

LEMMA 4. *There exists a natural number m_0 so that the number of those $m \times n$ (0, 1) matrices whose row-degree is at most* $\log m$ *but not equal to* 1, *is less than $2^{n(m-1)} \cdot 2^{m^{4/5}}$ if $m > m_0$.*

PROOF. Let us denote by D_l the number of those $m \times n$ (0, 1) matrices, whose row-degree is l and by $D_{l,i}$; the number of those (0, 1) matrices in which the i-th row vector has degree l. Then

$$D_l \leqq \sum_{i=1}^{m} D_{l,i} = m \cdot D_{l,m}$$

(because evidently $D_{l,1} = D_{l,2} = \ldots = D_{l,m}$).

We shall prove that

$$D_{l,m} < m^l \cdot 2^{2^l} \cdot 2^{n(m-1)} \qquad (l \geqq 2),$$

what proves our Lemma because the number of those matrices whose row-degree is at most $\log m$ but is not equal to 1 is

$$\sum_{l=2}^{[\log m]} D_l \leqq 2^{n(m-1)} \sum_{l=2}^{[\log m]} m \cdot m^l \cdot 2^{2^l} < 2^{n(m-1)} \cdot \log m \cdot m \cdot m^{\log m} \cdot 2^{2^{\log m}} < 2^{n(m-1)} \cdot 2^{m^{4/5}}$$

if $m > m_0$ for some suitable natural number m_0.

If we fill in the first $m-1$ rows of the matrix arbitrarily (it can be done by $2^{n(m-1)}$ different ways) we can construct the last row using a row-basis consisting of the first $m-1$ rows by a linear combination (because $l \geqq 2$) but actually we use only l rows of the row-basis, because the coefficients of the other rows are equal to 0. We have $\binom{m-1}{l} < m^l$ possibilities to choose the l vectors and by l vectors we can construct at most 2^{2^l} vectors as linear combinations according to Lemma3, that is

$$D_{l,m} < m^l \cdot 2^{2^l} \cdot 2^{n(m-1)} \qquad (l \geqq 2).$$

Q. e. d.

Similarly the number of $m \times n$ (0, 1) matrices whose column-degree is at most $\log n$ but is not equal to 1, is less than $2^{m(n-1)} \cdot 2^{n^{4/5}}$, if $n > m_0$.

If the row-(resp. column) degree of an $m \times n$ matrix is equal to 1, then we have two possibilities: either there are two rows(resp. columns) which are equivalent (the number of such matrices is less than $m^2 \cdot 2^{(m-1)n}$(resp. $n^2 \cdot 2^{m(n-1)}$), or the rank of the matrix is m (resp. n) — these are the good cases for us.

d) Let us consider an $n \times n$ (0, 1) matrix ($n > m_0$).

A) If its rank is n, then any additional column vector is linearly dependent of the column vectors of the matrix.

B) 1. If its rank is $k < n$ and its row-degree is $l > \log n$ then by Lemma 2 we have at most $\frac{2 \cdot 2^n}{\sqrt{l}} < \frac{2 \cdot 2^n}{\sqrt{\log n}}$ possibilities to add a column vector so that the rank of the obtained $n \times (n+1)$ is also k.

B) 2. The number of those $n \times (n+1)$ (0, 1) matrices for which the row-degree of the $n \times n$ matrix consisting of its first n columns is less than $\log n$ but not equal to 1 — by Lemma 4 — is less than $2^{n(n+1)} \cdot \frac{2^{n^{4/5}}}{2^n}$.

B) 3. If an $n \times (n+1)$ matrix has the property that the row-degree of the $n \times n$ matrix consisting of its first n columns is equal to 1, then (because $k < n$) in the latter matrix there exist two rows which are equivalent. So the number of these matrices is less than $2^{n(n+1)} \cdot \frac{n^2}{2^n}$.

Let **B** denote the set of matrices of the types B)2. and B)3. The number of elements of **B** is less than

$$2^{n(n+1)} \left(\frac{2^{n^{4/5}}}{2^n} + \frac{n^2}{2^n} \right) < 2^{n(n+1)} \cdot \frac{1}{2^{n/2}}$$

if $n > n_0 \geqq m_0$ for some suitable natural number n_0.

By a similar way we can prove that if we enlarge the obtained $n \times (n+1)$ matrix by a row vector and if the matrix is not an element of the set **B**, then the probability of the event, that the rank of the new matrix is larger than the rank of the first matrix is at least

$$1 - \frac{2}{\sqrt{\log n}}.$$

III.

a) So we have proved the following

LEMMA 5. *Let us consider an arbitrary* $n \times n$ (0, 1) *matrix which is not element of the set* **B**. *Let us enlarge the matrix by a column vector (with components* 0, 1*) and let us add to the new matrix a row vector in all possible ways. So we obtain* 2^{2n+1} $(n+1)\times(n+1)$ *matrix.*

If the rank of the first matrix is $k<n$, *then the rank of the new matrices are* $k+2$ *except for at most* $\frac{2}{\sqrt{\log n}}\cdot 2^{2n+1}$ *matrices, and if the rank of the first matrix is* $k=n$, *then the rank of the new matrices are* $n+1$ *except for at most* $\frac{2}{\sqrt{\log n}}\cdot 2^{2n+1}$ *matrices.*

b) Using Lemma 5 we obtain

LEMMA 6. *There exists a sequence* $n_1, n_2, \ldots, n_k, \ldots$ *of natural numbers such that*

$$A_{n_k,n_k} > 2^{n_k^2}\left(1-\frac{6}{\sqrt{\log n_k}}\right) \qquad (k = 1, 2, \ldots),$$

where $A_{m,r}$ *denotes the number of* $m\times m$ (0, 1) *matrices whose ranks are equal to* r.

By other words

$$\liminf_{n=+\infty} p_n = \liminf_{n=+\infty} \frac{A_n}{2^{n^2}} = \limsup\left(1-\frac{A_{n,n}}{2^{n^2}}\right) = 1-1 = 0.$$

PROOF. Let us put $S_n = \sum_{k=0}^{n} A_{n,k}\cdot k$ and $f(n) = \frac{S_n}{2^{n^2}}$. The inequality $S_n < \sum_{k=0}^{n} A_{n,k}\cdot n = n\cdot 2^{n^2}$ implies that $f(n)<n$. Let $\bar{A}_{n,k}$ denote the number of those $n\times n$ matrices whose ranks are k and which are not elements of the set **B** and $B_{n,k} = A_{n,k} - \bar{A}_{n,k}$. We can obtain all $(n+1)\times(n+1)$ matrices so that we enlarge the $n\times n$ matrices by a column vector to the right and after it by a row vector upwards in all possible ways.

So we can obtain from the $n\times n$ matrices of number $\bar{A}_{n,k}$ and of rank k new $(n\times 1)+(n+1)$ matrices the number of which is $\bar{A}_{n,k}2^{2n+1}$ and among them $x_{n,k}\bar{A}_{n,k}2^{2n+1}$ have rank smaller than $\min(k+2, n+1)$. By Lemma 5.

$$x_{n,k} < \frac{2}{\sqrt{\log n}} \qquad (k = 0, 1, 2, \ldots, n).$$

c) So we have

$$S_{n+1} = \sum_{k=0}^{n+1} k \cdot A_{n+1,k} \geqq 2^{2n+1} \sum_{k=0}^{n-1} \bar{A}_{n,k}(1-x_{n,k})(k+2) +$$

$$+ 2^{2n+1} \sum_{k=0}^{n-1} \bar{A}_{n,k} \cdot x_{n,k} \cdot k + 2^{2n+1} \bar{A}_{n,n}(1-x_{n,n})(n+1) + 2^{2n+1} \bar{A}_{n,n} \cdot x_{n,n} \cdot n =$$

$$= 2^{2n+1} \left(\sum_{k=0}^{n-1} \bar{A}_{n,k}(k+2) + \bar{A}_{n,n}(n+1) \right) - 2^{2n+1} \left(2 \cdot \sum_{k=0}^{n-1} \bar{A}_{n,k} \cdot x_{n,k} + \bar{A}_{n,n} x_{n,n} \right) \geqq$$

$$\geqq 2^{2n+1} \left(\sum_{k=0}^{n-1} \bar{A}_{n,k}(k+2) + \bar{A}_{n,n}(n+1) \right) - 2^{2n+1} \left(2 \cdot \sum_{k=0}^{n-1} \bar{A}_{n,k} + \bar{A}_{n,n} \right) \frac{2}{\sqrt{\log n}} =$$

$$= 2^{2n+1} \left(\sum_{k=0}^{n} \bar{A}_{n,k}(k+2) - \bar{A}_{n,n} \right) - 2^{2n+1} \left(2 \cdot \sum_{k=0}^{n} \bar{A}_{n,k} - \bar{A}_{n,n} \right) \frac{2}{\sqrt{\log n}} =$$

$$= 2^{2n+1} \sum_{k=0}^{n} \bar{A}_{n,k} \cdot k + 2 \cdot 2^{(n+1)^2} \left(1 - \frac{2}{\sqrt{\log n}} \right) - 2^{2n+1} \cdot \bar{A}_{n,n} \left(1 - \frac{2}{\sqrt{\log n}} \right) -$$

$$- 2 \cdot 2^{2n+1} \left(1 - \frac{2}{\sqrt{\log n}} \right) \sum_{k=0}^{n} B_{n,k} = 2^{2n+1} \sum_{k=0}^{n} A_{n,k} \cdot k + 2 \cdot 2^{(n+1)^2} \left(1 - \frac{2}{\sqrt{\log n}} \right) -$$

$$- 2^{2n+1} \cdot A_{n,n} \left(1 - \frac{2}{\sqrt{\log n}} \right) - 2 \cdot 2^{2n+1} \left(1 - \frac{2}{\sqrt{\log n}} \right) \sum_{k=0}^{n} B_{n,k} -$$

$$- 2^{2n+1} \sum_{k=0}^{n} B_{n,k} \cdot k + 2^{2n+1} \cdot B_{n,n} \left(1 - \frac{2}{\sqrt{\log n}} \right) \geqq$$

$$\geqq 2^{2n+1} \cdot S_n + 2 \cdot 2^{(n+1)^2} \left(1 - \frac{2}{\sqrt{\log n}} \right) - 2^{2n+1} \cdot A_{n,n} \left(1 - \frac{2}{\sqrt{\log n}} \right) -$$

$$- 2^{2n+1} \cdot 2^{n^2} \frac{n+2}{2^{n/2}} \geqq 2^{2n+1} \cdot S_n + 2 \cdot 2^{(n+1)^2} \left(1 - \frac{3}{\sqrt{\log n}} \right) - 2^{2n+1} \cdot A_{n,n} \left(1 - \frac{2}{\sqrt{\log n}} \right).$$

d) Dividing by $2^{(n+1)^2}$ we get

$$f(n+1) \geqq f(n) + 2 \left(1 - \frac{3}{\sqrt{\log n}} \right) - \frac{A_{n,n}}{2^{n^2}} \left(1 - \frac{2}{\sqrt{\log n}} \right).$$

If we suppose that there exists a number N_0 such that

$$A_{n,n} < 2^{n^2} \left(1 - \frac{6}{\sqrt{\log n}} \right)$$

holds for all $n \geqq N_0$ then we have

$$f(n+1) \geqq f(n)+2\left(1-\frac{3}{\sqrt{\log n}}\right)-\left(1-\frac{7}{\sqrt{\log n}}\right)$$

that is

$$f(n+1) \geqq f(n)+\left(1+\frac{1}{\sqrt{\log n}}\right) \tag{1}$$

for all $n \geqq N_0$. But $\sum_{k=N_0}^{\infty} \frac{1}{\sqrt{\log k}} = +\infty$ therefore using Relation (1) $(n-N_0)$ times we obtain

$$f(n+1) \geqq f(N_0)+(n-N_0)+\sum_{k=N_0}^{n} \frac{1}{\sqrt{\log k}}.$$

for all $n \geqq N_0$. If N is so large that $\sum_{k=N_0}^{N} \frac{1}{\sqrt{\log k}} > N_0+1$ holds, then we have $f(N+1)>N+1$ which is a contradiction. Q.e.d.

IV.

a) LEMMA 7.

Let $f(x, y)$ be a function defined for all pairs $x \geqq y$ of natural numbers with the following properties:

There exists a natural number n and a real number $0<c<1$ such that

1° $f(x, y) \geqq 0$
2° $f(x, x) = 1$
3° $f(x, y+1) \geqq f(x, y)$
4° $f(n, n-1) < c$
5° $f(m+1, k) \leqq cf(m, k)+(1-c)f(m, k-2)+d_m$

for all $m \geqq n$ and $0 \leqq k \leqq m$, where $\{d_m\}$ is a sequence of positive numbers.

We show that these properties imply that

$$f(m, m-1) < 2c+\sum_{s=n}^{\infty} d_s \tag{2}$$

or all $m \geqq n$.

b) By a double application of 5° we get

$$f(m+2, k) \leqq c^2 f(m, k)+2c(1-c)f(m, k-2)+(1-c)^2 f(m, k-4)+d_m+d_{m+1} \tag{3}$$

$$\begin{pmatrix} m \geqq n \\ 0 \leqq k \leqq m \end{pmatrix}$$

and this inequality implies (as $f(m, k-4) \leqq f(m, k-2)$):

$$(4) \qquad f(m+2, k) \leqq c^2 f(m, k) + (1-c^2) f(m, k-2) + d_m + d_{m+1}.$$

$$\begin{pmatrix} m \geqq n \\ 0 \leqq k \leqq m \end{pmatrix}$$

The relation

$$f(N+1, k) \leqq c f(N, k) + (1-c) f(N, k-2) + d_N \leqq$$
$$\leqq f(N, k)[c + (1-c)] + d_N = f(N, k) + d_N$$

$$\begin{pmatrix} N \geqq n \\ 0 \leqq k \leqq N \end{pmatrix}$$

and Relation 1° $\left(f(n, n-1) < c\right)$ show that

$$(5) \qquad f(m, k) < c + \sum_{s=n}^{m-1} d_s \quad \text{for all} \quad k \leqq n-1 \qquad (m \geqq n).$$

Now we prove by induction that the following inequality holds:

$$(6) \qquad f(n+t, n-2+t-i) \leqq c + \sum_{s=0}^{\left[\frac{t-i}{2}\right]-1} \binom{i+s}{s} c^{i+s+2} + \sum_{s=n}^{n+t-1} d_s$$

$$(t \geqq 2, \ i \geqq 0).$$

If $i > t-2$, then we have to prove that $f(n+t, n-(i-t+2)) \leqq c + \sum_{s=n}^{n+t-1} d_s$; but this is an immediate consequence of (5). Let us suppose that

$$i \leqq t-2.$$

c) In the case $t=2$ (and so $i=0$) the inequality is

$$f(n+2, n) \leqq c + c^2 + d_n + d_{n+1}.$$

By (4) we have

$$f(n+2, n) \leqq c^2 f(n, n) + (1-c^2) f(n, n-2) + d_n + d_{n+1} \leqq$$
$$\leqq c^2 + (1-c^2) c + d_n + d_{n+1} < c + c^2 + d_n + d_{n+1}.$$

In the case $t=3$ the inequality is (for $i=1$ or $i=0$)

$$f(n+3, n) \leqq c + c^3 + d_n + d_{n+1} + d_{n+2}$$
$$f(n+3, n+1) \leqq c + c^2 + d_n + d_{n+1} + d_{n+2}.$$

Using Relation (4):

$$f(n+3, n) \leqq c^2 f(n+1, n) + (1-c^2) f(n+1, n-2) + d_{n+1} + d_{n+2} \leqq$$
$$\leqq c^2 [c f(n, n) + (1-c) f(n, n-2) + d_n] + (1-c^2)(c + d_n) + d_{n+1} + d_{n+2} \leqq$$
$$\leqq c^3 + c^3 (1-c) + c(1-c^2) + d_n + d_{n+1} + d_{n+2} <$$
$$< c + c^3 + d_n + d_{n+1} + d_{n+2}$$

or similarly:

$$f(n+3, n+1) \leqq c^2 f(n+1, n+1) + (1-c^2) f(n+1, n-1) + d_{n+1} + d_{n+2} \leqq$$
$$\leqq c^2 + (1-c^2)(c+d_n) + d_{n+1} + d_{n+2} <$$
$$< c + c^2 + d_n + d_{n+1} + d_{n+2}.$$

That is the inequality is proved in the cases $t=2$ and $t=3$.

d) Let us suppose that the inequality is proved for $t=T$ and let us prove it for $t=T+2$. Denote $\left[\frac{T-i}{2}\right] = w$. Applying (3) we get, if $i \geqq 2$ $\left(\binom{n}{k} = 0 \text{ per. def. if } k>n \text{ or } k<0\right)$

$$f(n+T+2, n-2+(T+2)-i) = f(n+T+2, n+T-i) \leqq c^2 f(n+T, n+T-i) +$$
$$+ 2c(1-c) f(n+T, n-2+T-i) + (1-c)^2 f(n+T, n-4+T-i) + d_{n+T} + d_{n+T+1} \leqq$$
$$\leqq c^2 \left(c + \sum_{s=0}^{w} \binom{i+s-2}{s} c^{i+s} + \sum_{s=n}^{n+T-1} d_s\right) +$$
$$+ 2c(1-c) \left(c + \sum_{s=0}^{w-1} \binom{i+s}{s} c^{i+s+2} + \sum_{s=n}^{n+T-1} d_s\right) +$$
$$+ (1-c)^2 \left(c + \sum_{s=0}^{w-2} \binom{i+s+2}{s} c^{i+s+4} + \sum_{s=n}^{n+T-1} d_s\right) + d_{n+T} + d_{n+T+1} =$$
$$= c + \sum_{s=n}^{n+T+1} d_s + \sum_{s=0}^{w} \binom{i+s-2}{s} c^{i+s+2} + 2 \sum_{s=1}^{w} \binom{i+s-1}{s-1} c^{i+s+2} -$$
$$- 2 \sum_{s=2}^{w+1} \binom{i+s-2}{s-2} c^{i+s+2} + \sum_{s=2}^{w} \binom{i+s}{s-2} c^{i+s+2} - 2 \sum_{s=3}^{w+1} \binom{i+s-1}{s-3} c^{i+s+2} +$$
$$+ \sum_{s=4}^{w+2} \binom{i+s-2}{s-4} c^{i+s+2} = S.$$

Using the following identity

$$\binom{i+s-2}{s} + 2\binom{i+s-1}{s-1} - 2\binom{i+s-2}{s-2} + \binom{i+s}{s-2} - 2\binom{i+s-1}{s-3} + \binom{i+s-2}{s-4} = \binom{i+s}{s}$$

(this identity holds for $s \geqq 1$, $i \geqq 1$)

one can see, that

$$S = c + \sum_{s=n}^{n+T+1} d_s + \sum_{s=0}^{w} \binom{i+s}{s} c^{i+s+2} - 2\binom{i+w-1}{w-1} c^{i+w+3} -$$
$$- 2\binom{i+w}{w-2} c^{i+w+3} + \binom{i+w}{w-2} c^{i+w+4} + \binom{i+w-1}{w-3} c^{i+w+3},$$

and as

$$\binom{i+w}{w-2}+\binom{i+w-1}{w-3} \leqq 2\binom{i+w}{w-2},$$

we get the relation

$$f(n+(T+2), n-2+(T+2)-i) \leqq S \leqq c+\sum_{s=0}^{\left[\frac{T-i}{2}\right]}\binom{i+s}{s}c^{i+s+2}+\sum_{s=n}^{n+T+1} d_s$$

what we had to prove.

If $i=0$ or $i=1$, then the estimate

$$f(n+T, n+T-i) \leqq c+\sum_{s=0}^{w}\binom{i+s-2}{s}c^{i+s}+\sum_{s=n}^{n+T-1} d_s$$

and also the identity was false. Instead of this estimate we write $f(n+T, n+T-i) \leqq 1$, and so we get for S the same formula as above.

e) Let us apply the proved inequality in the case $i=0$.

$$f(n+t, n+t-2) \leqq c+\sum_{s=0}^{\left[\frac{t}{2}\right]-1} c^{s+2}+\sum_{s=n}^{n+t-1} d_s \qquad (t \geqq 2).$$

Hence

$$f(n+t+1, n+t) \leqq cf(n+t, n+t)+(1-c)f(n+t, n+t-2)+d_{n+t} \leqq$$

$$\leqq c+(1-c)\left(c+\sum_{s=0}^{\left[\frac{t}{2}\right]-1} c^{s+2}+\sum_{s=n}^{n+t-1} d_s\right)+d_{n+t} < c+(1-c)\left(c+\sum_{s=0}^{\infty} c^{s+2}\right)+$$

$$+\sum_{s=n}^{\infty} d_s = c+c-c^2+(1-c)\frac{c^2}{1-c}+\sum_{s=0}^{\infty} d_s = 2c+\sum_{s=n}^{\infty} d_s$$

for all $t \geqq 2$.

But

$$f(n+2, n+1) \leqq cf(n+1, n+1)+(1-c)f(n+1, n-1)+d_{n+1} \leqq$$

$$\leqq c+(1-c)(c+d_n)+d_{n+1} < 2c+\sum_{s=n}^{\infty} d_s$$

and

$$f(n+1, n) \leqq cf(n, n)+(1-c)f(n, n-2)+d_n \leqq c+(1-c)c+d_n < 2c+\sum_{s=n}^{\infty} d_s,$$

hence we proved that for all $m \geqq n$

$$f(m, m-1) < 2c+\sum_{s=n}^{\infty} d_s$$

holds. Q. e.d.

V.

Now we can already prove the theorem:

Let ε be an arbitrary positive number. Let the integer N be so large that for the N-th element of the sequence n_k (defined in Lemma 6)

$$\frac{13}{\sqrt{\log n_N}} < \varepsilon.$$

Let us put

$$f(m, k) = \sum_{i=0}^{k} \frac{A_{m,i}}{2^{m^2}},$$

$$c = \frac{6}{\sqrt{\log n_N}},$$

$$n = n_N,$$

$$d_m = \frac{1}{2^{m/2}}.$$

It is easy to see that for the function $f(m, k)$ 1°—2°—3° hold.

The fulfilment of 4° follows from the definition of the sequence $\{n_k\}$ (in lemma 6). Let us prove that 5° holds.

From the $\bar{A}_{m,k-1}$ matrices of rank $k-1$ except for at most $c \cdot \bar{A}_{m,k-1} \cdot 2^{2m+1}$ ones, and from the $\bar{A}_{m,k}$ matrices of rank k except for at most $c \cdot \bar{A}_{m,k} \cdot 2^{2m+1}$ ones we get such matrices, which have at least $k+1$ as rank. So we have

$$2^{(m+1)^2} f(m+1, k) \leqq \sum_{i=0}^{k-2} A_{m,i} \cdot 2^{2m+1} + c \cdot 2^{2m+1} (\bar{A}_{m,k-1} + \bar{A}_{m,k}) +$$

$$+ 2^{2m+1} \cdot d_m \cdot 2^{m^2} \leqq \sum_{i=0}^{k-2} A_{m,i} \cdot 2^{2m+1} + c \cdot 2^{2m+1} (A_{m,k-1} + A_{m,k}) + d_m \cdot 2^{(m+1)^2} =$$

$$= f(m, k-2) \cdot 2^{(m+1)^2} + c \cdot 2^{(m+1)^2} \big(f(m, k) - f(m, k-2)\big) + d_m \cdot 2^{(m+1)^2} =$$

$$= 2^{(m+1)^2} [cf(m, k) + (1-c) f(m, k-2) + d_m].$$

Dividing by $2^{(m+1)^2}$ we obtain

$$f(m+1, k) \leqq cf(m, k) + (1-c) f(m, k-2) + d_m,$$

that is 5° holds.

By lemma 7 we get:

$$f(m, m-1) < 2c + \sum_{s=n}^{\infty} d_s$$

for all $m \geqq n$. But

$$\sum_{s=n}^{\infty} d_s = \sum_{s=n}^{\infty} \frac{1}{2^{s/2}} = \frac{4}{2^{n/2}} < \frac{1}{\sqrt{\log n}}$$

that is

$$f(m, m-1) < \frac{12}{\sqrt{\log n_N}} + \frac{1}{\sqrt{\log n_N}} < \varepsilon$$

for all $m \geqq n_N$ or in other terms

$$A_{m,m} > 2^{m^2}(1-\varepsilon) \quad \text{for all} \quad m \geqq n_N,$$

what proves our theorem.

VI.

a)

Professor EGYED asked whether the following generalization of this theorem is true:

Let us consider the matrices:

$$\begin{matrix} a_{1,1} & a_{1,2} \dots a_{1,k} \dots \\ a_{2,1} & a_{2,2} \dots a_{2,k} \dots \\ & \cdot \quad \cdot \quad \cdot \\ a_{i,1} & a_{i,2} \dots a_{i,k} \dots \\ \vdots & \vdots \qquad \vdots \end{matrix}$$

where the elements $a_{i,k}$ equal to 0 or 1. The set of those matrices in which the rows or the columns are not "linearly independent", has a measure 0.

First we have to agree in that what is the meaning of "linearly independent" in this case.

Let $a_{i,k}$ $(i=1, 2, \dots;\ k=1, 2, \dots)$ be mutually independent random variables which take on the values 0, 1 with probabilities $\frac{1}{2}, \frac{1}{2}$. Let us form by these random variables the above matrix.

We make use of two definitions of the linear dependence of the rows of a matrix.

The rows of a matrix are *finitely linearly dependent*, if there exists a natural number i, some natural numbers (finitely many) $i_1 < i_2 < \dots < i_s$ and real numbers $\alpha_1, \alpha_2, \dots, \alpha_s$ with the properties:

$$i_\nu \neq i \quad \text{for} \quad \nu = 1, 2, \dots, s$$

and

$$a_{i,k} = \sum_{\nu=1}^{s} \alpha_\nu a_{i_\nu,k} \quad \text{for} \quad k = 1, 2, \dots . \tag{7}$$

The rows of a matrix are *infinitely linearly dependent*, if there exists a natural number i and real numbers $\alpha_1, \alpha_2, \dots, \alpha_{i-1}, \alpha_i = 0, \alpha_{i+1}, \dots$ such that

$$a_{i,k} = \sum_{\nu=1}^{\infty} \alpha_\nu a_{\nu,k} \quad \text{for} \quad k = 1, 2, \dots . \tag{8}$$

Let A denote the event that the rows of a random matrix are finitely linearly dependent and B the event that they are infinitely linearly dependent.

Making use of these definitions we can formulate the question as follows:

What are the probabilities $\mathbf{P}(A)$ and $\mathbf{P}(B)$ equal to?

b) *The answer is:*

(9) $$\mathbf{P}(A)=0,$$

(10) $$\mathbf{P}(B)=1.$$

The proofs of these relations are simple.

Proof of (9):

Let A_t denote the event that

$$a_{i_1,t} = a_{i_2,t} = \ldots = a_{i_s,t} = 0 \quad \text{and} \quad a_{i,t} = 1.$$

Clearly $A_1, A_2, \ldots, A_t, \ldots$ are mutually independent and $0<\mathbf{P}(A_1)=\mathbf{P}(A_2)=\ldots=$ $=\mathbf{P}(A_t)=\ldots$. One can see from the relation (7) that $A_t \cap A = \varnothing$, so $\left(\bigcup_{t=1}^{\infty} A_t\right) \cap A = \varnothing$ and thus $\overline{\bigcup_{t=1}^{\infty} A_t} = \bigcap_{t=1}^{\infty} \bar{A}_t \supset A$. This fact implies that

$$\mathbf{P}\left(\bigcap_{t=1}^{\infty} \bar{A}_t\right) \geqq \mathbf{P}(A).$$

But we have $\mathbf{P}\left(\bigcap_{t=1}^{\infty} \bar{A}_t\right) = \prod_{t=1}^{\infty} \mathbf{P}(\bar{A}_t)$ because the events $\bar{A}_t$ are independent.

As $\mathbf{P}(\bar{A}_t)<1$ and

$$\mathbf{P}(\bar{A}_1) = \mathbf{P}(\bar{A}_2) = \ldots = \mathbf{P}(\bar{A}_t) = \ldots,$$

we get $\prod_{t=1}^{\infty} \mathbf{P}(\bar{A}_t)=0$ whence $\mathbf{P}(A)=0$.

Proof of (10):

Let B_t denote the event $(t=1, 2, \ldots)$ that there exists a natural number i_t (different from $i_1, i_2, \ldots, i_{t-1}$) such that

$$a_{i_t,1} = a_{i_t,2} = \ldots = a_{i_t,t-1} = 0 \quad \text{and} \quad a_{i_t,t} = 1.$$

Clearly $\mathbf{P}(B_t)=1$, therefore $\mathbf{P}\left(\bigcap_{t=1}^{\infty} B_t\right) = 1$.

That is, a random matrix contains a triangular matrix, in which all diagonal elements are equal to 1, with probability 1. Clearly the matrix also contains an i-th row vector which is different from the rows of the triangular matrix, with probability 1. We show that such a row of the matrix is an infinite linear combination of the i_1-th, i_2-th, ..., i_t-th, ... rows.

Put $\alpha_{i_1}=a_{i,1}$ and define the numbers α_{i_k} successively as

$$\alpha_{i_k} = a_{i,k} - \sum_{\nu=1}^{k-1} \alpha_{i_\nu} a_{i_\nu,k}.$$

If t is not one of the numbers i_k then let $\alpha_t=0$.

Since

$$a_{i_k,k}=1,$$

$$a_{i_\nu,k}=0 \quad \text{for} \quad \nu>k$$

and

$$\alpha_\nu=0 \quad \text{if} \quad \nu\neq i_t,$$

we have

$$a_{i,k}=\alpha_{i_k}+\sum_{\nu=1}^{k-1}\alpha_{i_\nu}a_{i_\nu,k}=\sum_{\nu=1}^{k}\alpha_{i_\nu}a_{i_\nu,k}=\sum_{\nu=1}^{\infty}\alpha_{i_\nu}a_{i_\nu,k}=\sum_{\mu=1}^{\infty}\alpha_\mu a_{\mu,k},$$

that is (10) holds.

The condition (8) says that

$$\lim_{N=+\infty}\sum_{\nu=1}^{N}\alpha_\nu a_{\nu,k}=a_{i,k} \quad \text{for} \quad k=1,2,\ldots.$$

If we substitute this condition by the condition

$$\lim_{N=+\infty}\sum_{\nu=1}^{N}\alpha_\nu a_{\nu,k}=a_{i,k} \text{ uniformly in } k,$$

then the probability in question is equal to 0.

c)

In the proof of (10) we actually proved that the rows (and clearly the columns too) of a random matrix contain an infinite basis with probability 1.

(A subset of a set of vectors is called) "basis infinitely", if any element of the set can uniquely be represented by an infinite linear combination of the elements of the subset.

A subset of a set of vectors is called to be "basis finitely" (it can contain infinitely many vectors) if any element of the set can uniquely be represented by a linear combination of finitely many vectors of the elements of the subset).

I do not know whether there exists a set of vectors (with countably many components) containing no "basises infinitely".

(Finitely many vectors ever contain basis. It is easy to see that a set of countably many vectors also contain at least one "basis finitely" and we proved above that this basis is the whole set with probability 1.)

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received March 17, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 23—28.

ON A PROBLEM OF GRAPH THEORY

by

G. KATONA and E. SZEMERÉDI

1. Introduction

We say that a directed graph has the diameter 2 if any two vertices are connected by a directed path of length at most 2. Let $V(G)$ denote the number of vertices of a graph G. Let $E(G)$ denote the number of edges of G and $D(G)$ the diameter of G.

Put

$$F(n) = \min_{\substack{V(G)=n \\ D(G)=2}} E(G).$$

P. ERDŐS, A. RÉNYI and V. T. SÓS [1] proposed the question of determining the value of $F(n)$. The problem has the following interpretation. There are n airports. Any (ordered) pair A, B of these airports is connected by at most one (directed) flight from A to B. How many directed connections have to be established to assure the possibility to fly from every airport to any other by changing the plane at most once?

It was noticed also by P. ERDŐS, A. RÉNYI and V. T. SÓS that we can reduce this problem to the following one. A (non-directed) graph is called to be a complete even graph if we can split its vertices into two disjoint subsets, so that two vertices are connected if and only if they are in different subsets. At least how large is the sum of the numbers of the vertices of even complete graphs covering every edge of a complete graph having n vertices? We are going to prove in 2 of this paper that this number is at least $n \log n$. (Now and in what follows $\log n$ denotes $\overset{2}{\log} n$). In 3 we deduce from this result of 2 estimates for $F(n)$. In 4 we consider the sum of the numbers of the vertices of complete even graphs covering an arbitrary given graph.

2. On Covering of a Complete Graph by Complete Even Graphs

Let A and B be two finite disjoint sets. Let (A, B) denote the complete even graph in which x and y are connected if and only if $x \in A$ and $y \in B$, or if $x \in B$ and $y \in A$. We denote the number of elements of a set A by $|A|$. We say that the G graph is covered by the family of complete even graphs $(A_i B_i)$ $1 \leqq i \leqq m$ if any edge of G is the edge of some of the complete even graphs $(A_i B_i)$. The complete graph having n vertices is denoted by $\langle n \rangle$.

THEOREM 1. *If $\langle n\rangle$ is covered by the family of the complete even graphs (A_iB_i) $1\leqq i\leqq m$ then*

$$\sum_{i=1}^{m}|A_i|+|B_i|\geqq n\log n. \tag{1}$$

PROOF OF THEOREM 1. Let us denote the vertices of $\langle n\rangle$ by $x_1, x_2, \ldots, x_n$. We construct a matrix M of m rows and n columns each element of which is equal to one of the numbers 0, 1, 2.

We put $M=(a_{ij})$ where

$$a_{ij}=\begin{cases}0 & \text{if } x_j\in A_i\\ 1 & \text{if } x_j\in B_i\\ 2 & \text{if } x_j\notin A_i \quad x_j\notin B_i.\end{cases}$$

We denote the number of zeros and ones being in the j-th column by h_j. In case the family of complete even graphs (A_iB_i) $1\leqq i\leqq m$ cover $\langle n\rangle$ then there is to any of the pairs (x_k, x_l) an (A_iB_i), so that one of them is an element of A_i, and the other is an element of B_i. Concerning M this means that to any two different columns there is a row, so that in this row in one of the two columns there stands 0 and in the other 1. By other words to each j and k $(j\neq k)$ there is at least one i such that either $a_{ij}=0$ and $a_{ik}=1$ or $a_{ij}=1$ and $a_{ik}=0$. Thus the k-th column and l-th column are different if $k\neq l$, and they remain different even if we replace any 2 by either 0 or 1. There are in the k-th column $m-h_k$ elements equal to 2, so we obtain from the k-th column 2^{m-h_k} different columns, if we replace 2 whereever it occurs either by 0 or by 1. As, however, the number of columns having length m consisting of either 0 or 1 is 2^m, we get the inequality

$$\sum_{k=1}^{n}2^{m-h_k}\leqq 2^m \tag{2}$$

that is

$$\sum_{k=1}^{n}\frac{1}{2^{h_k}}\leqq 1. \tag{3}$$

The inequality between the geometrical and the arithmetical means and (3) imply

$$\sqrt[n]{\frac{1}{2^{\sum_{k=1}^{m}h_k}}}\leqq\sum_{k=1}^{n}\frac{1}{n}\frac{1}{2^{h_k}}\leqq\frac{1}{n}$$

wherefrom

$$\sum_{k=1}^{n}h_k\geqq n\log n. \tag{4}$$

However the total number of zeros and ones in the matrix M is equal to $\sum_{k=1}^{n}h_k$, and it is equal to $\sum_{i=1}^{m}|A_i|+|B_i|$ as well.

Otherwise it is obvious that we can always find such a family (A_iB_i) $1 \leqq i \leqq m$ which coverns $\langle n \rangle$ and satisfies

$$\sum_{i=1}^{m} |A_i| + |B_i| = n\{\log n\}$$

where $\{x\}$ denotes the smallest integer which is greater than x or equal to x. Let us construct the family in the form of a matrix. Let the j-th column consist of the sequence of digits of the number $j-1$ in the binary system. Since the number of digits of any $k \leqq n-1$ in the binary system is $\leqq [\log (n-1)] + 1 = \{\log n\}$ the number of the rows of the matrix will be $\{\log n\}$. M will entirely consist of 0-s and 1-s. Obviously the columns of M are all different and M has $n\{\log n\}$ elements. In this example was no 2 in the matrix. This corresponds to the case when the set is divided into two disjoint subsets by (A_iB_i). To cover $\langle n \rangle$ $\log n$ such pairs (A_iB_i) are necessary, as it was proved in [2]. As then $|A_i| + |B_i| = n$ the total number of vertices of the covering graphs is in this case $n\{\log n\}$. This fact led to conjecturing the result of Theorem 1.

3. Lower and Upper Bounds for $F(n)$

Now let us consider a directed graph of diameter 2, having n vertices $x_1x_2 \ldots x_n$ and between any two different vertices there is at most one directed edge. Let us consider the vertex x_j. Let the set A_j consist of those vertices to which a directed edge leads from x_j, and let the set B_j consist of x_j and of those vertices wherefrom a directed edge leads into x_j. Let x_k and x_l be any two different verices then, according to our assumption, either a directed edge or a directed path of length two leads from x_k into x_l. In the first case $x_k \in B_k$ and $x_l \in A_k$ while in the second case, if the directed path of length 2 from x_k to x_l goes through x_j, then $x_k \in B_j$ and $x_l \in A_j$. That means that $\langle n \rangle$ is covered by the family of (A_i, B_i) $1 \leqq j \leqq n$ complete even graphs. Theorem 1 implies

$$\sum_{i=1}^{n} |A_i| + |B_i| \geqq n \log n.$$

Since the number of the edges starting from x_j is $|A_j|$ and the number of edges ending in x_j is $|B_j| - 1$

$$E(G) = \sum_{j=1}^{n} \frac{|A_j| + |B_j| - 1}{2} \geqq \frac{n \cdot \log n}{2} - \frac{n}{2}.$$

Thus we obtained the following theorem:

THEOREM 2. *Let G be a directed graph of diameter two between any two vertices of which there is at most one directed edge and the number of its edges is $E(G)$, then*

$$E(G) \geqq \frac{n}{2} \log \frac{n}{2}. \tag{5}$$

Thus for the function $F(n)$ defined in § 1 we obtain the inequality

$$F(n) \geqq \frac{n}{2} \log \frac{n}{2}. \tag{6}$$

On the other hand, it follows from the remark at the end of § 1 that $F(n) \leqq n\{\log n\}$.

4. The Covering of Arbitrary Graphs by Complete Even Graphs

THEOREM 3. *Let G be a nondirected graph having n vertices $x_1, x_2 \dots x_n$ and let the degree of x_j be f_j and let G be covered by the family of complete even graphs, $(A_i B_i)$ $1 \leqq i \leqq m$ then*

$$\sum_{i=1}^{m} |A_i| + |B_i| \geqq \sum_{i=1}^{n} \log \frac{n}{n-f_i}. \tag{7}$$

PROOF OF THEOREM 3. The following lemma is necessary for the proof.

LEMMA: Let H be a set of k elements, let $H_1 \dots H_n$ be subsets of H and suppose that from any fixed i the number of j-s for which $H_i \cap H_j \neq 0$ is at least g_i. Then

$$\sum_{i=1}^{n} \frac{|H_i|}{g_i} \leqq k. \tag{8}$$

PROOF OF THE LEMMA. It is easy to see that

$$\sum_{i=1}^{n} \frac{|H_i|}{g_i} = \sum_{i=1}^{n} \sum_{x \in H_i} \frac{1}{g_i} = \sum_{x \in H} \sum_{\{i:\, x \in H_i\}} \frac{1}{g_i}. \tag{9}$$

To any fixed x, however, if $x \in H_j$ we have $|\{i: x \in H_i\}| \leqq g_j$, because H_j has common elements at least with $|\{i: x \in H_i\}|$ sets H_i. Therefore $\sum_{\{i:\, x \in H_i\}} \frac{1}{g_i} \leqq 1$ and thus from (9) $\sum_{i=1}^{n} \frac{|H_i|}{g_i} \leqq k$.

Let us return to the proof of the theorem. Similarly as in the proof of Theorem 1 we construct a matrix M of m rows and n columns, whose elements a_{ij} are defined as follows:

$$a_{ij} = \begin{cases} 0 & \text{if} \quad x_j \in A_i \\ 1 & \text{if} \quad x_j \in B_i \\ 2 & \text{if} \quad x_j \notin A_i \quad x_j \notin B_i \end{cases}$$

where $x_1, x_2 \dots x_n$ are the vertices of the graph G. Let H_j be the set of those columns which we can get from the j-th column of M, so that we write instead of any 2 occurring there independently either 0 or 1. Further, let us examine at most with how many H_i has a fixed H_j a common element. If in a graph G between x_j and x_i there is an edge, then there is an (A_l, B_l) such that $x_i \in A_l$ $x_j \in B_l$ or $x_i \in B_l$ $x_j \in A_l$. That is there may be found such l-th row to the j-th and i-th column in which the j-th element is 0 and the i-th element is 1 or vice versa. However, if we write into the j-th and i-th columns in the place of 2 either 0 or 1, the columns remain different, that is $H_i \cap H_j = 0$. Since the degree of x_j is f_j, at most $n - f_j$ is the number of those H_i for which $H_i \cap H_j \neq 0$. We can get from the lemma

$$\sum_{i=1}^{n} \frac{|H_i|}{n-f_i} \leqq 2^m$$

If h_i denotes the sum of the number of 0-s and 1-s in the i-th column we get

$$\sum_{i=1}^{n} \frac{2^{m-h_i}}{n-f_i} \leqq 2^m$$

that is

$$\sum_{i=1}^{n} \frac{1}{2^{h_i}(n-f_i)} \leqq 1. \tag{10}$$

The inequality between the geometrical and arithmetical means and (10) imply

$$\sqrt[n]{\frac{1}{2^{\sum_{i=1}^{n} h_i} \prod_{i=1}^{n} n-f_i}} \leqq \sum_{i=1}^{n} \frac{1}{n} \frac{1}{2^{h_i}(n-f_i)} \leqq \frac{1}{n}$$

wherefrom, after some calculations using the equality

$$\sum_{i=1}^{m} |A_i| + |B_i| = \sum_{i=1}^{n} h_i$$

we get

$$\sum_{i=1}^{m} |A_i| + |B_i| \geqq \sum_{i=1}^{n} \log \frac{n}{n-f_i} \tag{11}$$

what was to be proved.

Remarks

It is easy to see that Theorem 1 and Theorem 3 can be generalized in such a way that we use as covering graphs (instead of even graphs) graphs of the form $(A_1, A_2, \ldots A_s)$ which are defined as follows: The sets of vertices $A_1, A_2, \ldots A_s$ are disjoint and two vertices are connected by an edge if and only if they do not belong to the same set A_i. In this case the result, according to Theorem 3 is:

$$\sum_{i=1}^{m} |A_1^i| + |A_2^i| \ldots |A_s^i| \geqq \sum_{i=1}^{n} \overset{s}{\log} \frac{n}{n-f_i}.$$

A good estimate can be obtained by Theorem 3 if f_i is large. The case is of special interest when $f_i = f$ $(1 \leqq i \leqq n)$. In this case the right-hand side of (11) takes the form of $n \log \frac{n}{n-f}$. If, for instance, $f = n - c$ then $\sum_{i=1}^{m} |A_i| + |B_i| \geqq n \log \frac{n}{c}$. That is it does not differ essentially from the case of a complete graph. Taking the other interesting case when $f = c \cdot n$ then the right-hand side of (11) is $n \log \frac{1}{1-c}$.

REFERENCES

[1] ERDŐS, P., RÉNYI, A. and SÓS, V. T.: On a problem of graph theory, *Studia Sci. Math. Hung.* **1** (1966) 215—235.
[2] RÉNYI, A.: On random generating elements of a finite Boolean algebra, *Acta Sci. Math.* (Szeged) **22** (1961) 75—81.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received March 20, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 29—35.

THE PRINCIPLE OF CONSERVATION OF ENTROPY IN A NOISELESS CHANNEL

by

G. KATONA and G. TUSNÁDY

Introduction

The main aim of this paper is to formulate precisely and to prove the following statement:

If we have an information source, more precisely, a sequence of random variables $\xi_1, \xi_2, \ldots$ with entropy $H(\mathcal{X})$ and we code this sequence in a uniquely decodable manner, the obtained sequence $\mathcal{Y}$ has the entropy

$$H(\mathcal{Y}) = \frac{H(\mathcal{X})}{L}, \tag{1}$$

where L is the average length of the codes.

The intuitive meaning of (1) is clear: it expresses the principle of conservation of information, when the coding is uniquely decodable (and no noise is present). In spite of this according to our best knowledge (1) has not been proved in full generality up to now.

If we have finite number of code signals $y_1, \ldots, y_m$, the maximum of $H(\mathcal{Y})$ is $\log m$, where log denotes the logarithm with respect to the base 2. It follows from (1) that

$$\frac{H(\mathcal{X})}{\log m} \leqq L.$$

This is a well known theorem of SHANNON, and according to our best knowledge only this consequence of (1) was proved (e. g. [1]).

In the case when the coding is not necessarily uniquely decodable instead of (1) we prove the inequality

$$H(\mathcal{Y}) \leqq \frac{H(\mathcal{X})}{L},$$

which has also an intuitive meaning.

Precise Formulation

Let $X = \{x_1, \ldots, x_n\}$ be the set of possible signals (the alphabet) of the information source, and let X^∞ be the set of all infinite sequences formed from the letters $x_1, \ldots, x_n$. If $1 \leqq i_1 \leqq n, \ldots, 1 \leqq i_k \leqq n$, we denote by $[x_{i_1}, \ldots, x_{i_k}]$ the set of all sequences having $x_{i_1}, \ldots, x_{i_k}$ on the first k places. We call such subsets of X^∞ cylinder sets. Let $\mathfrak{A}_X$ denote the σ-field generated by the cylinder sets. The measure space $\mathcal{X} = (X^\infty, \mathfrak{A}_X, p_X)$

is called an information source, if p_X is a probability measure on $\mathfrak{A}_X$. This space defines an other space on the sequences of length k: $\mathcal{X}^k=(X^k, \mathfrak{A}_X^k, p_X^k)$, where X^k is the space of the sequences $(x_{i_1}, \ldots, x_{i_k})$. The average information contained in the first k signals of the information source is

$$H(\mathcal{X}^k) = -\sum_{\substack{1\leqq i_1\leqq n\\ \vdots\\ 1\leqq i_k\leqq n}} p_X^k(x_{i_1}, \ldots, x_{i_k}) \log p_X^k(x_{i_1}, \ldots, x_{i_k}) =$$

$$= -\sum_{\substack{1\leqq i_1\leqq n\\ \vdots\\ 1\leqq i_k\leqq n}} p_X[x_{i_1}, \ldots, x_{i_k}] \log p_X[x_{i_1}, \ldots, x_{i_k}].$$

Finally, the definition of the entropy of $\mathcal{X}$ is

$$H(\mathcal{X}) = \lim_{k\to\infty} \frac{H(\mathcal{X}^k)}{k}$$

(the average information content of one signal), if this limit exists.

Consider now the definition of the coding. Let $Y=\{y_1, \ldots, y_m\}$ be the set of possible code signals. Let $c(x_i)$ $(1\leqq i\leqq n)$ be a finite, non empty sequence formed from elements of Y. We call this function coding. Thus we may associate to every $(x_{i_1}, \ldots, x_{i_k})$, resp. $(x_{j_1}, x_{j_2}, \ldots)$ a sequence of y_i's. Let us write successively the codes $c(x_{i_1})$, ..., ..., $c(x_{i_k})$, resp. $c(x_{j_1})$, $c(x_{j_2})$, ... Denote by $c(x_{i_1}, \ldots, x_{i_k})$, resp. $d(x_{j_1}, x_{j_2}, \ldots)$ the resulting sequence. Let Y^∞ be the set of all infinite y-sequences. Thus the function $d(x_{j_1}, x_{j_2}, \ldots)$ transforms the set X^∞ into a subset Y^* of Y^∞. Let $\mathfrak{A}_Y$ be the σ-field in Y^* generated by the mapping $d(x_{j_1}, x_{j_2}, \ldots)$ of $\mathfrak{A}_X$ and let us define the measure P_Y on $\mathfrak{A}_Y$ by putting:

$$p_Y(A)=p_X(d^{-1}(A)) \quad A\in\mathfrak{A}_Y$$

where $d^{-1}(A)$ denotes the inverse image of A. $\mathcal{Y}=(Y^*, \mathfrak{A}_Y, p_Y)$ is the space of coded sequences. As above $[y_{i_1}, \ldots, y_{i_k}]$ denotes the cylinder set consisting of all sequences in Y^∞ of which the first k terms are $y_{i_1}, \ldots, y_{i_k}$.

LEMMA 1. $[y_{i_1}, \ldots, y_{i_k}]\cap Y^*\in\mathfrak{A}_Y$.

PROOF. We have to prove that the set of all sequences $(x_{j_1}, x_{j_2}, \ldots)$ having the image in $[y_{i_1}, \ldots, y_{i_k}]\cap Y^*$ is in $\mathfrak{A}_X$. We say that $(y_{i_1}, \ldots, y_{i_k})$ is a segment of $(y_{l_1}, \ldots, y_{l_s})$ if $k\leqq s$, and $y_{i_1}=y_{l_1}, \ldots, y_{i_k}=y_{l_k}$. Obviously, the image of $(x_{j_1}, x_{j_2}, \ldots)$ is in $[y_{i_1}, \ldots, y_{i_k}]\cap Y^*$ if and only if $(y_{i_1}, \ldots, y_{i_k})$ is a segment of $c(x_{j_1}, \ldots, x_{j_k})$. Thus

$$d^{-1}([y_{i_1}, \ldots, y_{i_k}]\cap Y^*)=\cup[x_{j_1}, \ldots, x_{j_k}], \tag{2}$$

where union runs over sequences $j_1, \ldots, j_k$ for which $(y_{i_1}, \ldots, y_{i_k})$ is a segment of $c(x_{j_1}, \ldots, x_{j_k})$. However the right side of (2) is a union of cylinder sets in $\mathfrak{A}_X$ which proves the Lemma.

Denote by Y^k the set of sequences $(y_{i_1}, \ldots, y_{i_k})$ for which the cylinder set $[y_{i_1}, \ldots, y_{i_k}]$ is in Y^*. Let $\mathfrak{A}_Y^k$ be the σ-field of all subsets of Y^k. By Lemma 1 we can define

$$p_Y^k(y_{i_1}, \ldots, y_{i_k})=p_Y([y_{i_1}, \ldots, y_{i_k}]\cap Y^*).$$

The average information content of the first k signals of the coded sequence is

$$H(\mathscr{Y}^k) = - \sum_{\substack{1 \leqq y_{i_1} \leqq m \\ \vdots \\ 1 \leqq y_{i_k} \leqq m}} p_Y^k(y_{i_1}, \ldots, y_{i_k}) \log p_Y^k(y_{i_1}, \ldots, y_{i_k}),$$

where $\mathscr{Y}^k = (Y^k, \mathfrak{A}_Y^k, p_Y^k)$. Finally the definition of the entropy of $\mathscr{Y}$ is

$$H(\mathscr{Y}) = \lim_{k \to \infty} \frac{H(\mathscr{Y}^k)}{k}$$

(the average information content of one signal of the coded sequence), if this limit exists.

If l_i denotes the length of the sequence $c(x_i)$ $(1 \leqq i \leqq n)$ then the length of $c(x_{i_1}, \ldots, x_{i_k})$ is $\sum_{j=1}^{k} l_{i_j}$. Let L_k be a random variable in the probability space $\mathscr{X}^k$, which takes on the value $\sum_{j=1}^{k} l_{i_j}$ if we have the sequence $(x_{i_1}, \ldots, x_{i_k})$. We say that the average code length is L, if

$$\frac{L_k}{k} \Rightarrow L$$

for $k \to \infty$, where $\Rightarrow$ denotes convergence in probability.

A coding $c(x_i)$ $(1 \leqq i \leqq n)$ is called uniquely decodable, if

$$c(x_{i_1}, \ldots, x_{i_k}) = c(x_{j_1}, \ldots, x_{j_s})$$

holds only in the case $k = s$, $x_{i_1} = x_{j_1}, \ldots, x_{i_k} = x_{j_k}$.

We would like to point out that in the above sequence of definitions only the definitions of entropies, average code length and uniquely decodable coding are important, and the other ones are technical.

Finally, one more definition is necessary to the proof. Denote by Z^N the set of the sequences $(x_{i_1}, \ldots, x_{i_s})$ satisfying the conditions

$$\sum_{j=1}^{s} l_{i_j} \geqq N, \quad \sum_{j=1}^{s-1} l_{i_j} < N$$

Let $\mathfrak{A}_Z^N$ be the σ-field of all subsets of Z^N and put

$$p_Z^N(x_{i_1}, \ldots, x_{i_s}) = p_X[x_{i_1}, \ldots, x_{i_s}].$$

It is easy to see, that

$$\sum_{(x_{i_1}, \ldots, x_{i_s}) \in Z^N} p_Z^N(x_{i_1}, \ldots, x_{i_s}) = 1$$

thus $\mathscr{Z}^N = (Z^N, \mathfrak{A}_Z^N, p_Z^N)$ is a probability space, and

$$H(\mathscr{Z}^N) = - \sum_{(x_{i_1}, \ldots, x_{i_s}) \in Z^N} p_X(x_{i_1}, \ldots, x_{i_s}) \log p_X(x_{i_1}, \ldots, x_{i_s}).$$

Theorems and Proofs

To prove our main theorem we need a lemma.

LEMMA 2. *If L exists,*

$$\left(\frac{H(\mathscr{X}^k)}{L\cdot k}-\frac{H(\mathscr{Z}^N)}{N}\right)\to 0$$

provided that $k=\left[\frac{N}{L}\right]$ *and* $N\to\infty$.

PROOF. Let $H(\mathscr{Z}^N|\mathscr{X}^k)$ be the conditional entropy

$$H(\mathscr{Z}^N|\mathscr{X}^k)=-\sum_{\substack{u\in X^k\\ z\in Z^N}} p_X(u,z)\log p_X(z|u),$$

where $p_X(u,z)$ denotes the probability of the subsets of elements in X^∞, for which the first k elements are $x_{i_1},\ldots,x_{i_k}$ and the first s elements are $x_{j_1},\ldots,x_{j_s}$, if $u=(x_{i_1},\ldots,x_{i_k})$, $z=(x_{j_1},\ldots,x_{j_s})$. Further

$$p_X(z|u)=\frac{p_X(u,z)}{p_X(u)}.$$

Obviously, $p_X(u,z)\neq 0$ $(p_X(z|u)\neq 0)$ if and only of if one of the u and z is a segment of the other.

It is well known that (e. g. [1])

$$H(\mathscr{X}^k)+H(\mathscr{Z}^N|\mathscr{X}^k)=H(\mathscr{Z}^N)+H(\mathscr{X}^k|\mathscr{Z}^N), \tag{3}$$

so it is sufficient to show, that

$$H(\mathscr{Z}^N|\mathscr{X}^k)=o(N)$$

and

$$H(\mathscr{X}^k|\mathscr{Z}^N)=o(N)$$

if $k=\left[\frac{N}{L}\right]$. Namely, in this case

$$\frac{H(\mathscr{X}^k)}{N}-\frac{H(\mathscr{Z}^N)}{N}=\frac{H(\mathscr{X}^k|\mathscr{Z}^N)}{N}-\frac{H(\mathscr{Z}^N|\mathscr{X}^k)}{N}\to 0$$

follows from (3).

However

$$H(\mathscr{Z}^N|\mathscr{X}^k)=\sum_{u\in X^k} p_X(u)H(\mathscr{Z}^N|u) \tag{4}$$

where $H(\mathscr{Z}^N|u)$ denotes the entropy $-\sum_{z\in Z^N} p_X(z|u)\log p_X(z|u)$. Let $L_k(u)=l(u)$ be the number $\sum_{j=1}^{k} l_{i_j}$ if $u=(x_{i_1},\ldots,x_{i_k})$.

If $l(u)\geqq N$, $p_X(z|u)\neq 0$ can hold only if z is a segment of u (as u cannot be a proper segment of z only if $u=z$), but because of definition of Z^N there can be only one segment of u in Z^N. Thus $p_X(z|u)=1$ for a certain z, and

$$H(\mathscr{Z}^N|u)=0. \tag{5}$$

In the case $l(u) < N$, $p_X(z|u) \neq 0$ if and only if u is a segment of z. On the other hand $s-k \leqq N-l(u)$ since in the case $s-k=N-l(u)$ for the length $l(z)$ the inequality $l(z) \geqq l(u)+s-k=N$ holds. Obviously, we have only $n^{N-l(u)}$ such sequences z, that is,

$$H(\mathcal{Z}^N|u) \leqq \log n^{N-l(u)} = (N-l(u)) \log n \tag{6}$$

(as the maximum of the entropy of a distribution on a set of M elements is $\log M$).

Applying (4), (5) and (6) we have

$$H(\mathcal{Z}^N|\mathcal{X}^k) = \sum_{\substack{u \in X^k \\ l(u) < N}} p_X(u) H(\mathcal{Z}^N|u) \leqq \sum_{\substack{u \in X^k \\ l(u) < N}} p_X(u)(N-l(u)) \log n =$$

$$= \sum_{\substack{u \in X^k \\ N(1-\varepsilon) \leqq l(u) < N}} p_X(u)(N-l(u)) \log n + \sum_{\substack{u \in X^k \\ l(u) < (1-\varepsilon)N}} p_X(u)(N-l(u)) \log n \leqq$$

$$\leqq N\varepsilon \log n + p_X(u \in X^k,\ l(u) < (1-\varepsilon)N) N \log n.$$

Thus we have the inequality

$$\frac{H(\mathcal{Z}^N|\mathcal{X}^k)}{N} \leqq \varepsilon \log n + p_X(u \in X^k, l(u) < (1-\varepsilon)N) \log n. \tag{7}$$

Since $\frac{l(u)}{k} = \frac{L_k(u)}{k}$ converges stochastically to L, $p_X\left(u \in X^k, \left|\frac{l(u)}{k} - L\right| > \varepsilon\right)$ converges to zero if $k \to \infty$. It follows that on the right side of (7)

$$p_X(u \in X^k, l(u) < (1-\varepsilon)N) = p_X\left(u \in X^k, \frac{l(u)}{k} - L < (1-\varepsilon)\frac{N}{k} - L\right) \leqq$$

$$\leqq p_X\left(u \in X^k, \frac{l(u)}{k} - L < -L \cdot \frac{\varepsilon}{2}\right)$$

tends to zero for $N \to \infty$, because of $k = \left[\frac{N}{L}\right]$. Thus, if N is sufficiently large,

$$p_X(u \in X^k, l(u) < (1-\varepsilon)N) < \varepsilon$$

that is

$$\frac{H(\mathcal{Z}^N|\mathcal{X}^k)}{N} \leqq 2\varepsilon \log n$$

consequently

$$\lim_{N \to \infty} \frac{H(\mathcal{Z}^N|\mathcal{X}^k)}{N} = 0 \qquad \left(k = \left[\frac{N}{L}\right]\right).$$

We prove in similar way that

$$\lim_{N \to \infty} \frac{H(\mathcal{X}^k|\mathcal{Z}^N)}{N} = 0 \qquad \left(k = \left[\frac{N}{L}\right]\right).$$

Obviously

$$H(\mathcal{X}^k|\mathcal{Z}^N) = \sum_{z \in Z^N} p_X(z) H(\mathcal{X}^k|z),$$

where $H(\mathscr{X}^k|z)=0$ in the case $z=(x_{i_1}, \ldots, x_{i_s})$, $s \geqq k$. Further, if $s<k$, we have only n^{k-s} different u's with $p_X(u|z)\neq 0$, that is, $H(\mathscr{X}^k|z)\leqq(k-s)\log n$. Finally, as above

$$H(\mathscr{X}^k|\mathscr{Z}^N) \leqq \sum_{\substack{z\in Z^N \\ k(1-\varepsilon)\leqq s<k}} p_X(z)(k-s)\log n + \sum_{\substack{z\in Z^N \\ s<(1-\varepsilon)k}} p_X(z)(k-s)\log n \leqq$$

$$\leqq \varepsilon k \log n + k \log n\, p_X(z\in Z^N,\ s<(1-\varepsilon)k).$$

Here on the right side

$$p_X(z\in Z^N,\ s<(1-\varepsilon)k) = p_X(u\in X^{[(1-\varepsilon)k]},\ l(u)\geqq N) =$$

$$= p_X\left(u\in X^M,\ \frac{l(u)}{M}-L\geqq\frac{N}{(1-\varepsilon)k}-L\right) \leqq p_X\left(u\in X^M,\ \frac{l(u)}{M}-L\geqq\varepsilon L\right)$$

which converges to zero if $M=[(1-\varepsilon)k]\to\infty$. Thus $H(\mathscr{X}^k|\mathscr{Z}^N)\leqq o(N)$, indeed, which proves the Lemma.

THEOREM 1. *If the entropy $H(\mathscr{X})$ and the average code length L exist, and the coding is uniquely decodable, then $H(\mathscr{Y})$ exists, and*

$$H(\mathscr{Y}) = \frac{H(\mathscr{X})}{L}.$$

PROOF. If $H(\mathscr{X})$ exists, in Lemma 2 $\dfrac{H(\mathscr{X}^k)}{L\cdot k}$ tends to $\dfrac{H(\mathscr{X})}{L}$, and so $\dfrac{H(\mathscr{Z}^N)}{N}$ does, too. Thus, it is sufficient to show that

$$\lim_{N\to\infty}\frac{H(\mathscr{Z}^N)}{N} = \lim_{N\to\infty}\frac{H(\mathscr{Y}^N)}{N}.$$

We can write

$$H(\mathscr{Y}^N)+H(\mathscr{Z}^N|\mathscr{Y}^N) = H(\mathscr{Z}^N)+H(\mathscr{Y}^N|\mathscr{Z}^N) \tag{8}$$

where

$$H(\mathscr{Z}^N|\mathscr{Y}^N) = -\sum_{\substack{z\in Z^N \\ v\in Y^N}} p_X(v,z)\log p_X(z|v),$$

$$H(\mathscr{Y}^N|\mathscr{Z}^N) = -\sum_{\substack{z\in Z^N \\ v\in Y^N}} p_X(v,z)\log p_X(v|z),$$

$$p_X(z|v)=\frac{p_X(v,z)}{p_Y(v)}, \quad p_X(v|z)=\frac{p_X(v,z)}{p_X(z)},$$

and $p_X(v, z)$ is the probability of the set of sequences in X^∞, for which the first s elements are $x_{i_1}, \ldots, x_{i_s}$ and the first N elements of its code are $y_{j_1}, \ldots, y_{j_N}$, if $z=(x_{i_1}, \ldots, x_{i_s})$, $v=(y_{j_1}, \ldots, y_{j_N})$.

Obviously $p_X(v, z)\neq 0$ only if v is a segment of $c(z)$. Thus for given z there is only one v satisfying $p_X(v, z)\neq 0$, that is, $p_X(v, z)=1$. Applying this result we obtain

$$H(\mathscr{Y}^N|\mathscr{Z}^N) = \sum_{z\in Z^N} p_X(z)H(\mathscr{Y}^N|z) = 0 \tag{9}$$

because of $H(\mathscr{Y}^N|z)=0$. On the other hand

$$H(\mathscr{Z}^N|\mathscr{Y}^N) = \sum_{v\in Y^N} p_Y(v)H(\mathscr{Z}^N|v). \tag{10}$$

Let l be the maximum of the numbers $l_1, \ldots, l_n$. Because of definition of Z^N, $N \leqq l(z) < N + l$ holds. Thus for a fixed v, the sequences z satisfying $p_X(z|v) \neq 0$ are such that the first N elements of $c(z)$ are equal to v, and the other elements are arbitrary. The number of such $c(z)$'s is at most m^{l-1}. Since the coding is uniquely decodable i. e. to a given $c(z)$ there exists only one z, the number of different z is also at most m^{l-1}. From (10) we obtain

$$H(\mathcal{X}^N|\mathcal{Y}^N) \leqq \sum_{v \in Y^N} p_Y(v) \log m^{l-1} = \log m^{l-1},$$

which tends to zero divided by N, if $N \to \infty$. The proof is finished by (8).

If the coding is not necessarily uniquely decodable, we can not prove the existence of $H(\mathcal{Y})$. In this case let us put

$$\overline{H}(\mathcal{Y}) = \overline{\lim_{k \to \infty}} \frac{H(\mathcal{Y}^k)}{k}.$$

In this case from the above proof we get only $\overline{H}(\mathcal{Y}^N) \leqq H(\mathcal{X}^N)$ that is, the following theorem holds.

THEOREM 2. *If the entropy $H(\mathcal{X})$ and the average code length L exist, then*

$$\overline{H}(\mathcal{Y}) \leqq \frac{H(\mathcal{X})}{L}. \tag{11}$$

Further Questions

1. A natural question is th e following: under which assumption does the limit $\lim_{k \to \infty} \frac{H(\mathcal{Y}^k)}{k}$ exists in the not uniquely decodable case? Probably it is not difficult to answer this question if $\mathcal{X}$ is an information source, which produces independent signals.

2. It is easy to see, that for independent $\mathcal{X}$, and not uniquely decodable coding the strict inequality

$$\overline{H}(\mathcal{Y}) < \frac{H(\mathcal{X})}{L}$$

holds. In other words, in the independent case equality holds in (11) if and only if the coding is uniquely decodable. What is the necessary and sufficient condition, in general, of the equality in (11)?

We are greatly indebted to A. RÉNYI and I. CSISZÁR for several helpful comments and ideas.

REFERENCE

[1] FEINSTEIN, A.: *Foundation of information theory*, McGraw-Hill, New York, 1958.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received April 1, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 37—42.

ON THE ARRANGEMENT OF HOUSES IN A HOUSING ESTATE

by

L. FEJES TÓTH

On a large area we want to build as many houses as possible under the condition that the houses have congruent rectangular bases which are not allowed to get closer to one another than a prescribed distance. Calling the parallel domain of a rectangle a *site*, we face the problem of finding a densest packing of congruent sites.

It is known that the density of an arbitrary packing of congruent centro-symmetric convex discs cannot exceed the density of the densest lattice-packing of the discs[1]. This enables us to restrict ourselves to lattice-packings.

Without loss of the generality, we may suppose that the site s is a parallel domain at unit distance of a rectangle of sides a and b with $a \leqq b$. We will show that, depending on a, there are three types of densest lattice-packings which we define by certain properties of the centro-symmetric hexagon h of least area containing s. The corresponding packing arises by paving the plane by translated replicas of h and placing in each hexagon a translated replica of s.

We call the axis of symmetry of s parallel to the longer side of the rectangle the *axis* of s. If $a = b$, we agree to call only one of the axes of symmetry the axis of s.

We consider a centro-symmetric hexagon circumscribed about s in such a way that it has a pair of sides parallel to the axis of s and all of its side-midpoints lie on the boundary of s. Among the hexagons sharing all these properties we distinguish three types.

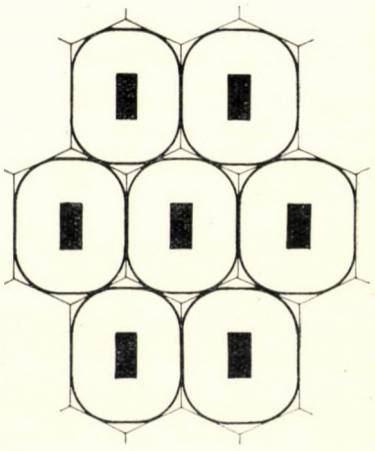


Fig. 1

Type 1. The hexagon has bilateral symmetry about the axis of s (Fig. 1).

Type 2. The hexagon has no bilateral symmetry about the axis of s and no side perpendicular to the axis of s (Fig. 2).

Type 3. The hexagon has two sides perpendicular to the axis of s (Fig. 3).

Our result reads as follows.

[1] See, e. g., L. Fejes Tóth, Regular Figures, Oxford 1964.

Let s be the parallel domain at unit distance of a rectangle whose shorter side has the length a. Then the centro-symmetric hexagon of least area containing s is of type 1, 2 *or* 3 *according as* $a \leqq 4-\sqrt{12}, 4-\sqrt{12} < a < 2-\sqrt{2}$ *or* $a \geqq 2-\sqrt{2}$, *respectively.*

As to the proof we must confess that it has a blemish: The comparison of the order of some functions with one variable had been carried out only by numerical computations (kindly performed by K. KÖNYVES TÓTH, using a computer giving 8 figures). Since this process seemed to be sufficiently reliable, I did not find it worth while to enter into tedious exact investigations.

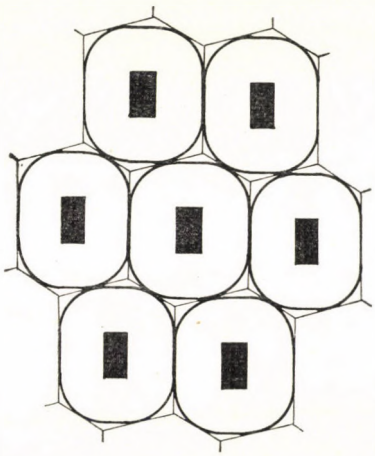


Fig. 2

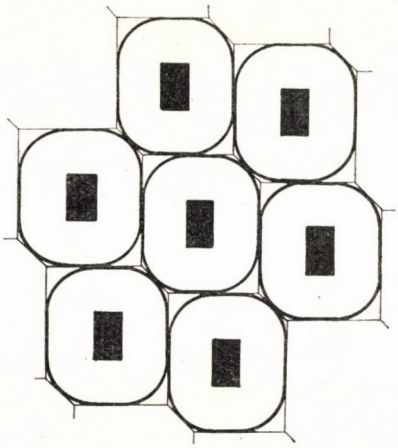


Fig. 3

First we consider the case that the rectangle is a square q with side-length $a=2x$. The fact that all side-midpoints of a minimal centro-symmetric hexagon lie on s is easily seen and we omit its proof. In order to survey all hexagons H with this property, we classify them as follows.

Class 0. H has no side parallel to a side of q.

Let A, B, C be consecutive vertices of q such that the unit circle with center B contains two side-midpoints S and T of H (Fig.4). Let LM and MN be the sides of H containing S and T, and let R and U be the adjacent side-midpoints. Since $LS=SM=MT=TN$, we have $LB=MB=NB$. Again, since in view of the central symmetry $RL=NU$, we have $AL=CN$. Thus the triangles ALB and CNB are congruent, in consequence of which H is symmetric about the diagonal of q passing through B.

Since $2ST=RU=2(1+\sqrt{2}x)$ and $ST<\sqrt{2}$, this case exists if and only if $2x<2-\sqrt{2}$. The half area of H equals

$$f_0(x)=(1+\sqrt{2}x)\left(\sqrt{3-\sqrt{8}x-2x^2}+\sqrt{8}x\right).$$

Class 1. H has exactly one pair of sides parallel to a side of q. Here we have two subclasses.

Subclass 11. H has an axis of symmetry.

Because of the central symmetry, H has two axis of symmetry, one of which must be parallel to those sides of H which are parallel to a side of q. This case can be realised only if $x<1$. The half area of H is

$$f_{11}(x) = (1+x)\left(\sqrt{3+2x-x^2}+2x\right).$$

Subclass 12. H has no axis of symmetry[2].

Let AB be a side of q not parallel to any side of H (Fig. 5). We imagine AB to be in a horisontal position and suppose that A and B are the upper vertices of q. Let L, M, N be the three upper vertices of H such that $AL=AM<BM=BN$.

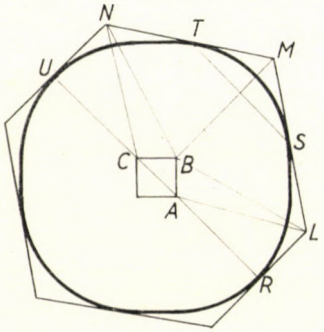


Fig. 4

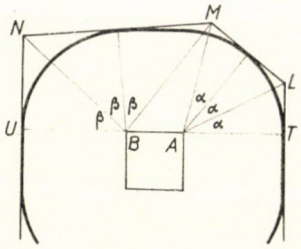


Fig. 5

Let the line AB intersect the vertical sides of H at T and U. Letting $\sphericalangle TAL=\alpha$ and $\sphericalangle UBN=\beta$, we have $AL=1/\cos\alpha$ and $BN=1/\cos\beta$. Hence, expressing the distance of M from the line TU both in terms of α and β, we find that

$$z(\alpha)=z(\beta), \tag{1}$$

where

$$z(\varphi) = \frac{\sin 3\varphi}{\cos\varphi}.$$

On the other hand, projecting the broken line AMB perpendicularly to the line AB, we obtain

$$2x = -\frac{\cos 3\alpha}{\cos\alpha} - \frac{\cos 3\beta}{\cos\beta},$$

whence, on account of

$$\frac{\cos 3\varphi}{\cos\varphi} = 1-4\sin^2\varphi,$$

$$x = 2(\sin^2\alpha+\sin^2\beta)-1. \tag{2}$$

[2] The possibility of such a hexagon was first pointed out by G. Hajós.

The assumption that $AL<BN$, along with the obvious conditions $\sin 3\alpha/\cos\alpha>1$ and $\tan\beta<1$, implies

$$\frac{\pi}{8}<\alpha<\beta<\frac{\pi}{4}. \tag{3}$$

In the interval $(\pi/8, \pi/4)$ the function $z(\varphi)$ has a single maximum defined by

$$z'\cos^2\varphi=8\cos^4\varphi-4\cos^2\varphi-1=0.$$

Denoting the root of this equation by γ, we have

$$\cos^2\gamma=\frac{1+\sqrt{3}}{4}.$$

In $(\pi/8, \gamma)$ the function $z(\varphi)$ increases and in $(\gamma, \pi/4)$ it decreases. Consequently, to any value of α such that $\pi/8<\alpha<\gamma$ the equation (1) associates a unique value of β such that $\gamma<\beta<\pi/4$. Thus, in view of (2), x may be considered as a function of α. Numerical computations show that x strictly decreases when α increases from $\pi/8$ to γ.

In the limiting case when $\alpha=\beta=\gamma$, we have

$$x=4\sin^2\gamma-1=3-4\cos^2\gamma=2-\sqrt{3}.$$

In the other limiting case, when $\alpha=\pi/8$, we have $\beta=\pi/4$ and $2x=2-\sqrt{2}$. Since these limiting cases do not belong to the subclass under consideration, we have $2-\sqrt{3}<x<1-\frac{\sqrt{2}}{2}$. The half area of H is equal to

$$f_{12}(x)=8(\sin^3\alpha\cos\alpha+\sin^3\beta\cos\beta)+2x(1+x),$$

where α and β are given by (1), (2) and (3).

Class 2. H has two pairs of sides parallel to the sides of q.

Now the third pair of sides must be perpendicular to a diagonal of q. This case can be realised only if $2x\geqq 2-\sqrt{2}$. The half area of H equals

$$f_2(x)=2x^2+4x+\sqrt{8}-1.$$

This completes the enumeration of the hexagons under consideration.

Now we must pick out of the functions f_0, f_{11}, f_{12} and f_2 the least one for various values of x. We denote the graphs of these functions by g_0, g_{11}, g_{12} and g_2.

For $0<x\leqq 2-\sqrt{3}$ we have to compare only f_0 and f_{11}. We have $f_0(0)=$ $=f_{11}(0)=\sqrt{3}$, $f_0'(0)=\sqrt{\frac{8}{3}+\sqrt{8}}\approx 4{,}46$ and $f_{11}'(0)=\frac{\sqrt{48}}{3}+2\approx 4{,}31$. Thus g_0 and g_{11} start from the same point, but g_0 has here a greater slope than g_{11}. The graph g_0 remains in the whole interval $(0, 2-\sqrt{3})$ above g_{11}. At $x=2-\sqrt{3}$ we still have $f_0=3{,}0426\ldots>f_{11}=3{,}0394\ldots$.

In the open interval $\left(2-\sqrt{3}, 1-\frac{\sqrt{2}}{2}\right)$ three functions compete, namely f_0, f_{11} and f_{12}. In the whole interval f_{12} turns out to be the smallest. At $x=2-\sqrt{3}\approx 0{,}26795$ we have $f_{11}=f_{12}$, and at the beginning f_{11} is a strong rival of f_{12}. At $x\approx 0{,}26806$ ($\alpha=33°30'$, $\beta\approx 35°1'28{,}9''$) we have $f_{12}\approx 3{,}0400064$ and $f_{11}\approx 3{,}0400065$. From here on the difference $f_{11}-f_{12}$ increases more rapidly. At $x=1-\frac{\sqrt{2}}{2}\approx 0{,}29289$ we have $f_{11}\approx 3{,}1761411$ and $f_{12}\approx 3{,}1715729$.

Yet, at the end of the interval, f_{12} has another rival. For, at a value of x near 0,28 the graph g_0 intersects g_{11} and at the end of the interval g_0 meets g_{12}. Nevertheless, g_0 remains above g_{12}. For instance, for $x\approx 0{,}29082$ ($\alpha=23°$, $\beta\approx$ $\approx 44°\ 35'\ 2{,}5''$) we have $f_{12}\approx 3{,}1608412$ and $f_0\approx 3{,}1608621$.

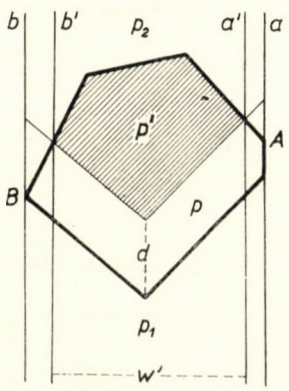


Fig. 6

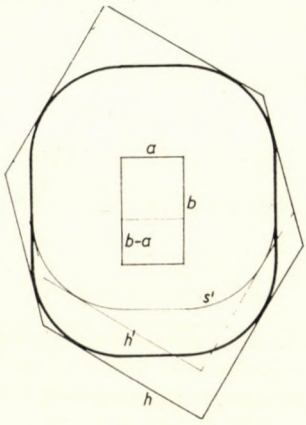


Fig. 7

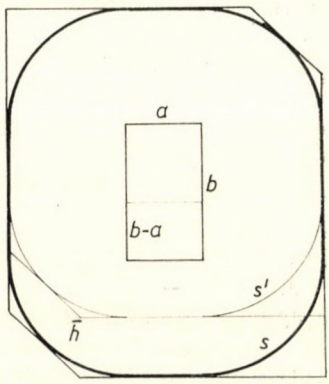


Fig. 8

For $x\geqq 1-\frac{\sqrt{2}}{2}$ the competition is less acute. At first there are two competitors, namely f_{11} and f_2. In view of $f_{12}\left(1-\frac{\sqrt{2}}{2}\right)=f_2\left(1-\frac{\sqrt{2}}{2}\right)<f_{11}\left(1-\frac{\sqrt{2}}{2}\right)$, g_2 starts from a lower point than g_{11} and remains throughout below g_{11}. At $x=1$ the function f_{11} gives up the game so that f_2 becomes unrivalled.

Recapitulating, the minimal centro-symmetric hexagon circumscribed about s belongs for $x\leqq 2-\sqrt{3}$ to class 11, for $2-\sqrt{3}<x<1-\frac{\sqrt{2}}{2}$ to class 12 and for $x\geqq 1-\frac{\sqrt{2}}{2}$ to class 2. Since these classes are identical with the above types 1, 2 and 3, respectively, our assertion is established for a square.

In order to settle the general case, we describe a transformation carrying a convex polygon p into a new one. Let a and b be two vertical supporting lines of p (Fig. 6). Let A and B be vertices of p lying on a and b, respectively. A and B decompose the boundary of p into an "upper" and a "lower" polygonal line $A\ldots B$ and $B\ldots A$. We consider p as being the intersection of the two convex pointsets

p_1 and p_2 bounded by a, $A...B$, b and b, $B...A$, a, respectively. Translating p_2 upwards through a distance d, we obtain a new polygon p'. We call this process *telescoping*.

We suppose p' not to be empty. Let a' and b' be the vertical supporting lines of p', and let w' be the distance of a' and b'. Then

$$t = t' + w'd + t'',$$

where t, t' and t'' are the areas of p, p' and the part of p outside the strip bounded by a' and b'.

Now let s be a site with a vertical axis arising from a rectangle with side-length a and b such that $a < b$. Let h be an arbitrary centro-symmetric hexagon of area t circumscribed about s. We telescope h in a vertical direction through the distance $b-a$, obtaining a hexagon h' of area t' (Fig. 7). The hexagon h' contains a site s' arising from a square of side-length a having a vertical axis. Therefore the width w' of h' in a horizontal direction is at least $2+a$. Thus

$$t \geqq t' + w'(b-a) \geqq t' + (2+a)(b-a).$$

Replacing t' by the area $\bar{t}$ of the least centro-symmetric hexagon $\bar{h}$ containing s', we obtain

$$t \geqq \bar{t} + (2+a)(b-a).$$

But since $\bar{h}$ has a pair of vertical sides, the sum $\bar{t} + (2+a)(b-a)$ equals the area of the hexagon arising from $\bar{h}$ by telescopic elongation in a vertical direction through the distance $b-a$ (Fig. 8). Since this hexagon contains s, it is proved to be the minimal hexagon of all centro-symmetric hexagons containing s.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES,
BUDAPEST

(Received April 20, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 43—48.

ON THE SUMMABILITY OF THE FOURIER SERIES OF L^2 INTEGRABLE FUNCTIONS, III.

by

E. MAKAI

§ 1

Let

$$f(x) = \frac{a_0}{2} + \sum_{\nu=1}^{n} (a_\nu \cos \nu x + b_\nu \sin \nu x)$$

be a trigonometrical polynomial with complex coefficients, $s_k(x; f)$ its k'th partial sum and

$$\|f\| = \left\{ \frac{a_0 \bar{a}_0}{2} + \sum_{\nu=1}^{n} (a_\nu \bar{a}_\nu + b_\nu \bar{b}_\nu) \right\}^{1/2}.$$

In a previous paper [2] I have verified that for $n \leqq 53$

$$\left| \sum_{r=1}^{n} s_{k_r}\left(\frac{2\pi}{n} r; f\right) \right| \leqq \sqrt{\frac{3}{2}}\, n \|f\| \tag{1}$$

($k_r = 0, 1, \ldots, n$; $r = 1, \ldots, n$) with equality if and only if

$$k_1 = k_2 = \ldots = k_n = n \quad \text{and} \quad f(x) = a(\tfrac{1}{2} + \cos nx). \tag{2}$$

If (1) would hold for an infinity of n's then the Fourier series of any L^2 integrable function would converge almost everywhere.

Let now be $F(z) = c_0 + c_1 z + \ldots + c_n z^n$, $|||F|||^2 = c_0\bar{c}_0 + c_1\bar{c}_1 + \ldots + c_n\bar{c}_n$ and $F_k(z) = c_0 + c_1 z + \ldots + c_k z^k$, further $\Phi(z) = \gamma_1 z + \gamma_2 z^2 + \ldots + \gamma_n z^n$, $|||\Phi|||^2 = \gamma_1\bar{\gamma}_1 + \gamma_2\bar{\gamma}_2 + \ldots + \gamma_n\bar{\gamma}_n$ and $\Phi_k(z) = \gamma_1 z + \gamma_2 z^2 + \ldots + \gamma_k z^k$.

In § 1 it will be shown that

whenever for a particular n (1) *and* (2) *are true, then for this n*

$$\left| \sum_{r=1}^{n} F_{k_r}(e^{2\pi i r/n}) \right| \leqq \sqrt{2}\, n |||F||| \tag{3}$$

and

$$\left| \sum_{r=1}^{n} \Phi_{k_r}(e^{2\pi i r/n}) \right| \leqq n |||\Phi||| \tag{3'}$$

hold with equality if and only if

$$k_1 = k_2 = \ldots = k_n = n \quad \textit{and} \quad F(z) = c(1 + z^n),\ \Phi(z) = \gamma z^n, \quad \text{respectively.} \tag{4}$$

In particular (3) *and* (3′) *are true for* $n \leqq 53$ [1], *with equality if and only if* (4) *hold.*

Indeed, using the notations $\vartheta_r = 2\pi r/n$ and $\varepsilon_{\nu,k} = 1$ if $\nu \leqq k$ and $\varepsilon_{\nu,k} = 0$ if $\nu > k$, we have

$$\left|\sum_{r=1}^{n} F_{k_r}(e^{i\vartheta_r})\right|^2 = \left|\sum_{r=1}^{n}\sum_{\nu=0}^{n} \varepsilon_{\nu,k_r} c_\nu e^{i\nu\vartheta_r}\right|^2 =$$

$$= \left|\sum_{\nu=0}^{n} c_\nu \sum_{r=1}^{n} \varepsilon_{\nu,k_r} e^{i\nu\vartheta_r}\right|^2 \leqq \sum_{\nu=0}^{n} c_\nu \bar{c}_\nu \sum_{\nu=0}^{n} \sum_{p,q=1}^{n} \varepsilon_{\nu,k_p}\varepsilon_{\nu,k_q} e^{i\nu(\vartheta_p - \vartheta_q)}$$

by Schwarz's inequality. Since this last quantity is positive we may write

$$\left|\sum_{r=1}^{n} F_{k_r}(e^{i\vartheta_r})\right| \leqq |||F||| \left\{\sum_{p,q=1}^{n} \sum_{\nu=0}^{\min(k_p,k_q)} \cos \nu(\vartheta_p - \vartheta_q)\right\}^{1/2} =$$

(5)

$$= |||F||| \left\{\sum_{p,q=1}^{n} \left[\frac{1}{2} + D_{\min(k_p,k_q)}(\vartheta_p - \vartheta_q)\right]\right\}^{1/2}$$

where $D_l(x) = 1/2 + \cos x + \cos 2x + \ldots + \cos lx$ is Dirichlet's kernel.

Equality stands in (5) only if

$$c_\nu = c' \sum_{r=1}^{n} \varepsilon_{\nu,k_r} e^{-i\nu\vartheta_r} \tag{6}$$

where c' is independent of ν.

Now I have shown [1] that for a fixed sequence $k_1, k_2, \ldots, k_n$

$$\left|\sum_{r=1}^{n} s_{k_r}(\vartheta_r; f)\right| \leqq \left\{\sum_{p,q=1}^{n} D_{\min(k_p,k_q)}(\vartheta_p - \vartheta_q)\right\}^{1/2}$$

where equality is attained for a particular f.

Hence

$$\max_{|||F|||=1} \left|\sum_{r=1}^{n} F_{k_r}(e^{i\vartheta_r})\right|^2 - \max_{\|f\|=1} \left|\sum_{r=1}^{n} s_{k_r}(\vartheta_r; f)\right|^2 = \frac{n^2}{2} \tag{7}$$

and whenever the modulus of the second term on the left hand side is less than $3n^2/2$, we have

$$\max_{|||F|||=1} \left|\sum_{r=1}^{n} F_{k_r}(e^{i\vartheta_r})\right|^2 < 2n^2.$$

This is true, as previously mentioned, if $\min(k_1, k_2, \ldots, k_n) < n$ and $n \leqq 53$ (possibly for the other n's, too). It rests to envisage the case $k_1 = k_2 = \ldots = k_n = n$.

[1] If either (3) or (3′) would hold for all n's or at least for an infinity of n's then in view of our main formula (7) or of the analogous equality

$$\max_{\|f\|=1} \left|\sum_{r=1}^{n} s_{k_r}(\vartheta_r; f)\right|^2 - \max_{|||\Phi|||=1} \left|\sum_{r=1}^{n} \Phi_{k_r}(e^{i\vartheta_r})\right|^2 = \frac{n^2}{2}$$

it would follow again that the Fourier series of every L^2 integrable function would converge almost everywhere.

Then we have, again by (7) or directly from (5)

$$\left| \sum_{r=1}^{n}{}' F(e^{i\theta_r}) \right|^2 \leqq 2n^2 |||F|||^2$$

and equality stands here by (6) if and only if

$$c_\nu = c' \sum_{r=1}^{n}{}' e^{-i\nu 2\pi r/n}$$

i. e., $c_0 = c_n = c$, $c_1 = c_2 = \ldots = c_{n-1} = 0$.

Thus we have proved the statement about the polynomials $F(z)$. The corresponding statements (3′) and (4) about $\Phi(z)$ are deduced in a similar way.

§ 2

Let π_n' be the class of the not identically vanishing n'th order real trigonometrical polynomials

$$f = f(x) = \sum_{\nu=1}^{n}{}' (a_\nu \cos \nu x + b_\nu \sin \nu x) \tag{1}$$

with the norm

$$\|f\| = \left\{ \sum_{\nu=1}^{n}{}' (a_\nu^2 + b_\nu^2) \right\}^{1/2}.$$

Let further $s_k(x; f)$ be the k'th partial sum of (1) and m a natural number. In Part II of this paper [2] I verified up to $m = 53$ an equivalent of the following

CONJECTURE. *If* $m|n$, *then*

$$\frac{1}{m} \sum_{r=1}^{m}{}' \max_{k=0,1,\ldots,n} s_k\left(\frac{2\pi}{m} r; f\right) \leqq \sqrt{\frac{n}{m}} \|f\| \qquad (f \in \pi_n') \tag{2}$$

with equality if and only if

$$f(x) = c(\cos mx + \cos 2mx + \cos 3mx + \ldots + \cos nx), \quad c > 0. \tag{3}$$

Consider the special case $m = n$ of this conjecture. It asserts that

$$\frac{1}{m} \sum_{r=1}^{m}{}' \max_{k=0,1,\ldots,m} s_k\left(\frac{2\pi}{m} r; f\right) \Big/ \|f\| \leqq 1, \quad \text{if} \quad f \in \pi_m' \tag{4}$$

with equality only if

$$f(x) = c \cos mx, \quad c > 0. \tag{5}$$

We are going to show that this last weaker form implies the whole conjecture. Stated more explicitly:

THEOREM. *If for some* m (4) *holds, then for any multiple* n *of* m (2) *is true; further if for the same* m *the sign of equality in* (4) *holds only for positive multiples of* $\cos mx$, *then in* (2) *equality holds only if* $f(x)$ *is given by* (3).

In the following let us introduce the notations $x_r = 2\pi r/m$,

$$\Gamma_n^{(m)} = \max_{f \in \pi_n'} \frac{1}{m} \sum_{r=1}^{m} \max_{k=0,1,\ldots,n} s_k(x_r; f)/\|f\|$$

and let us denote by $\hat{f} = \sum_{\nu=1}^{n} (\hat{a}_\nu \cos \nu x + \hat{b}_\nu \sin \nu x)$ an extremal function of this maximum problem, i. e.

$$\Gamma_n^{(m)} = \frac{1}{m} \sum_{r=1}^{m} \max_{k=0,1,\ldots,n} s_k(x_r; \hat{f})/\|\hat{f}\|.$$

We shall need the following

LEMMA. *If* $m|n$, *then* $\Gamma_n^{(m)} \leqq \sqrt{n/m}\, \Gamma_n^{(m)}$.

Indeed, introducing the functions

$$f_l(x) = \sum_{\nu=(l-1)m+1}^{lm} (\hat{a}_\nu \cos \nu x + \hat{b}_\nu \sin \nu x) =$$

$$= \sum_{\mu=1}^{m} (\hat{a}_{(l-1)m+\mu} \cos [(l-1)m+\mu]x + \hat{b}_{(l-1)m+\mu} \sin [(l-1)m+\mu]x)$$

and

$$\varphi_l(x) = \sum_{\mu=1}^{m} (\hat{a}_{(l-1)m+\mu} \cos \mu x + \hat{b}_{(l-1)m+\mu} \sin \mu x)$$

$(l = 1, 2, \ldots, n|m)$ we may write

$$\max_{0 \leqq k \leqq n} s_k(x_r; \hat{f}) = \max_{0 \leqq k \leqq n} s_k\left(x_r; \sum_{l=1}^{n/m} f_l\right) \leqq \sum_{l=1}^{n/m} \max_{(l-1)m \leqq k \leqq lm} s_k(x_r; f_l) = \sum_{l=1}^{n/m} \max_{0 \leqq k \leqq m} s_k(x_r; \varphi_l)$$

since we have $\exp i\mu x_r = \exp i[(l-1)m+\mu]x_r$ for $r = 1, 2, \ldots, m$, hence on the places x_r we have

$$s_{(l-1)m+\mu}(x_r; f_l) = s_\mu(x_r; \varphi_l) \qquad (r, \mu = 1, 2, \ldots, m).$$

Moreover

$$\Gamma_n^{(m)} \|\hat{f}\| = \frac{1}{m} \sum_{r=1}^{m} \max_{0 \leqq k \leqq n} s_k(x_r; \hat{f}) \leqq \tag{6}$$

$$\leqq \sum_{l=1}^{n/m} \frac{1}{m} \sum_{r=1}^{m} \max_{0 \leqq k \leqq m} s_k(x_r; \varphi_l) \leqq \sum_{l=1}^{n/m} \Gamma_m^{(m)} \|\varphi_l\| =$$

$$= \sum_{l=1}^{n/m} \Gamma_m^{(m)} \|f_l\| \leqq \Gamma_m^{(m)} \sqrt{\frac{n}{m}} \left\{ \sum_{l=1}^{n/m} \|f_l\|^2 \right\}^{1/2} = \Gamma_m^{(m)} \sqrt{\frac{n}{m}}\, \|\hat{f}\|$$

by using in turn the definition of $\Gamma_m^{(m)}$ and Cauchy's inequality.

From our lemma it follows that if for a particular m $\Gamma_m^{(m)} = 1$, and $m|n$, then $\Gamma_n^{(m)} \leqq \sqrt{n/m}$. On the other hand by the definition of $\Gamma_n^{(m)}$, supposing again $m|n$,

$$\Gamma_n^{(m)} \geqq \frac{1}{m} \sum_{r=1}^{m} \max_{k=0,1,\ldots,n} s_k(x_r; \cos mx + \cos 2mx + \ldots + \cos nx) \Big/ \frac{1}{\sqrt{n/m}}$$

and since for $l=1, 2, \ldots$ $\cos lmx_r = 1$, we may write by taking everywhere $k=n$

$$\Gamma_n^{(m)} \geqq \frac{1}{m} \cdot \frac{1}{\sqrt{\frac{n}{m}}} \sum_{r=1}^{m} s_n(x_r; \cos mx + \cos 2mx + \ldots + \cos nx) = \sqrt{\frac{n}{m}}.$$

Hence if $\Gamma_m^{(m)} = 1$, then

$$\Gamma_n^{(m)} = \sqrt{\frac{n}{m}} \qquad (n = m, 2m, 3m, \ldots). \tag{7}$$

It rests to show that if for some m (5) is the only extremal function of (4), then (3) is the unique extremal function of (2). Indeed, by our assumption it follows that

$$\Gamma_m^{(m)} = \frac{1}{m} \sum_{r=1}^{m} \max_{k=0, 1, \ldots, m} s_k(x_r; \cos mx) = 1. \tag{8}$$

Hence by (7) and (8) there must stand equality everywhere in (6).

On the place of the last inequality in (6) there stands the sign of equality only if

$$\|\varphi_1\| = \|\varphi_2\| = \ldots = \|\varphi_{n/m}\|$$

and the next but last inequality degenerates into equality by our assumption if and only if

$$\varphi_l = c_l \cos mx.$$

Combining the last two equalities we have

$$c_l = c\varepsilon_l \qquad (l=1, 2, \ldots, n/m)$$

where $\varepsilon_l = \pm 1$, hence every extremal function of the problem is of the form

$$c(\varepsilon_1 \cos mx + \varepsilon_2 \cos 2mx + \ldots + \varepsilon_{n/m} \cos nx).$$

Since the partial sums of these functions on the places x_r either vanish or are equal to one of the quantities $c(\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_g)$, where $0 < g \leqq n/m$, we can conclude that in case of an extremal function we have necessarily

$$\varepsilon_1 = \varepsilon_2 = \ldots = \varepsilon_{n/m} = 1, \quad c > 0. \tag{9}$$

Indeed, for $r=1, 2, \ldots, m$

$$s_k(x_r; \varepsilon_1 \cos mx + \ldots + \varepsilon_{n/m} \cos nx) = \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_{[k/m]} \leqq n/m$$

with equality if and only if $k=n$, and (9) holds.

This proves the unicity of the extremal function under our assumption.

Finally we want to point out that in Part I of this paper [1] we formulated the conjecture that if π_n is the class of the n'th order trigonometrical polynomials

$$f(x) = \frac{a_0}{2} + \sum_{\nu=1}^{n} (a_\nu \cos \nu x + b_\nu \sin \nu x)$$

and $m|n$, then

$$(9) \qquad \frac{1}{m}\left|\sum_{r=1}^{m} \max_{k=0,1,\ldots,m} s_k\left(\frac{2\pi}{m}r;f\right)\right| \leqq \sqrt{\frac{n}{m}}\left\{\frac{|a_0|^2}{2}+\sum_{\nu=1}^{n}\left(|a_\nu|^2+|b_\nu|^2\right)\right\}^{1/2}$$

with equality if and only if

$$(10) \qquad f=c(\tfrac{1}{2}+\cos mx+\cos 2mx+\ldots+\cos nx).$$

Now we can state that if for a particular m (4) and (5) hold, then for this m (9) and (10) are true, too.

This is an obvious consequence of our theorem and of Lemmas 1 and 2 of Part II of this paper.

REFERENCES

[1] MAKAI, E.: On the summability of the Fourier series of L^2 integrable functions, I, *Publ. Math., Debrecen* **11** (1964) 101—118.

[2] MAKAI, E.: On the summability of the Fourier series of L^2 integrable functions, II, *Publ. Math., Debrecen* **12** (1965) 89—106.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received April 22, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 49—51.

ÜBER DIE VERZERRUNGSEIGENSCHAFTEN DER KONFORMEN ABBILDUNG DES EINHEITSKREISES AUF „ϱ_0-KONVEXE" GEBIETE II.

von

K. SZILÁRD

In der ersten Mitteilung über den hier zu behandelnden Gegenstand (s. [1]) wurde folgender Satz bewiesen.

Es sei die Funktion $w=f(z)$ im Inneren des Einheitskreises der Ebene der komplexen Veränderlichen z definiert, dort analytisch, $f(0)=0$ und $|f'(0)|=1$, ferner sei die Menge $\mathfrak{G}$ der Bildpunkte w in der w-Ebene „ϱ_0-konvexartig" in bezug auf einen solchen Randpunkt C von $\mathfrak{G}$, der von dem Punkte $w=0$ einen minimalen Abstand R besitzt (d. h. jeder Punkt w mit $|w|<R$ sei ein Bildpunkt). Wir setzen auch voraus, daß die Halbgerade aus $w=0$ durch C keinen Bildpunkt w mit $|w|\geqq R$ enthält. Dann behaupten wir:

$$R \geqq \frac{1}{8}\left[\sqrt{(4\varrho_0-1)^2+32\varrho_0}-(4\varrho_0-1)\right]. \tag{1}$$

Die rechte Seite dieser Ungleichung, die wir durch $R_0(\varrho_0)$ bezeichnen wollen, ist eine monoton wachsende Funktion von ϱ_0. Es ist $R_0(0)=\frac{1}{4}$ und $\lim\limits_{\varrho_0\to\infty} R_0(\varrho_0)=\frac{1}{2}$. Der Termin „$\varrho_0$-konvexartig in bezug auf einen Randpunkt" wurde in der zitierten Arbeit [1] erklärt.

Wir wollen uns nun von der Voraussetzung „dass die Halbgerade aus $w=0$ durch C keinen Bildpunkt w mit $|w|\geqq R$ enthält" befreien, also beweisen, daß die Ungleichung (1) ohne diese zusätzliche Voraussetzung gilt. Der Beweis gelingt für den Fall, daß die Abbildung, welche die Funktion $w=f(z)$ von der Kreisfläche $|z|<1$ liefert, *schlicht* ist, was wir für die folgenden Ausführungen auch annehmen wollen. Somit lautet der Satz, den wir beweisen wollen, folgendermaßen.

Die für $|z|<1$ definierte analytische Funktion $w=f(z)$ verwirkliche eine schlichte Abbildung des Einheitskreises auf ein Gebiet $\mathfrak{G}$ der w-Ebene, es sei $f(0)=0$ und $|f'(0)|=1$, ferner sei das Gebiet $\mathfrak{G}$ „ϱ_0-konvexartig" in bezug auf einen seiner Randpunkte C, der von dem Punkte $w=0$ einen minimalen Abstand R besitzt. Dann gilt die Ungleichung (1).

Beweis. Wir betrachten das Gebiet $\mathfrak{G}^*$, welches wir durch Symmetrisierung nach Pólya aus dem Gebiet $\mathfrak{G}$ in bezug auf eine Halbgerade, die durch den Punkt $w=0$ geht, erhalten haben (s. [2] und [3]). Um diese Halbgerade zu fixieren, nehmen wir ohne die Allgemeinheit einzuschränken an, dass der Randpunkt C auf der negativen Hälfte der reellen Achse der w-Ebene liegt und somit dort die Abszisse $w=-R$ besitzt (s. Fig. 1). Da das Gebiet $\mathfrak{G}$ in bezug auf diesen Randpunkt „ϱ_0-konvexartig" ist, so enthält der Kreis mit dem Mittelpunkt $w=-\varrho_0-R$ und mit dem Radius ϱ_0 (der also den Punkt $w=-R$ auf seinem Rande erhält), in seinem Inneren

keinen Punkt des Gebietes $\mathfrak{G}$. Die Halbgerade, in bezug auf welche wir die Symmetrisierung nach Pólya vornehmen, soll aus dem Kreismittelpunkt $w=-\varrho_0-R$ ausgehen und durch den Punkt $w=0$ gehen (also insbesondere die ganze positive Hälfte der reellen Achse der w-Ebene enthalten).[1] Das Gebiet $\mathfrak{G}^*$, welches wir durch Pólya-Symmetrisierung in bezug auf diese Halbgerade erhalten haben, enthält, genau so wie $\mathfrak{G}$, die ganze Kreisfläche $|w|<R$ und enthält keinen Punkt w der abgeschlossenen Kreisfläche $|w+R+\varrho_0|\leqq\varrho_0$. Doch enthält es auch keinen Punkt der negativen Hälfte der reellen Achse mit einer Abszisse $w<-R$. Dies folgt daraus, dass wegen der Schlichtheit der Abbildung durch die Funktion $w=f(z)$ das Gebiet $\mathfrak{G}$ einfach zusammenhängend ist, folglich das Gebiet $\mathfrak{G}^*$ auch (s. [3], Seite 71). Würde $\mathfrak{G}^*$ einen Punkt der reellen Achse mit einer Abszisse $w<-R-2\varrho_0$ enthalten, so würde es eine ganze Kreislinie mit dem Mittelpunkt $w=-R-\varrho_0$ durch diesen Punkt und (wegen des einfachen Zusammenhanges von $\mathfrak{G}^*$), auch die ganze zugehörige Kreisscheibe enthalten, was, wie wir gesehen haben, nicht der Fall sein kann. In bezug auf die beiden Gebiete $\mathfrak{G}$ und $\mathfrak{G}^*$ und den Punkt $w=0$, der in beiden Gebieten enthalten ist, gilt folgender Satz von Pólya und Szegő (s. [3], Seite 81).

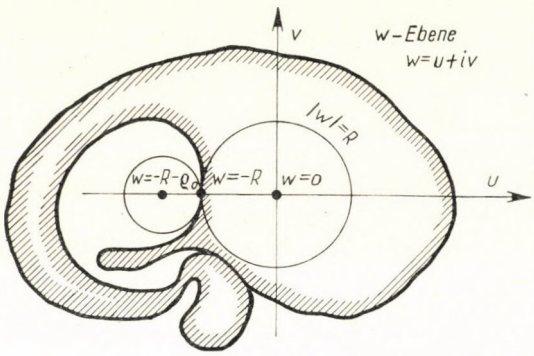


Fig. 1

„Nehmen wir an, daß a_0 ein Punkt des Gebietes $\mathfrak{G}$ in der w-Ebene ist und, daß das Gebiet $\mathfrak{G}^*$ aus $\mathfrak{G}$ durch Symmetrisierung in bezug auf eine Gerade, oder Halbgerade, welche durch a_0 geht, entstanden ist. Es seien r_0 und r_0^* die inneren Radii von $\mathfrak{G}$, bzw. $\mathfrak{G}^*$ in bezug auf a_0. Dann ist $r_0\leqq r_0^*$."

In unserem Falle ist $a_0=0$ und auf Grund dieses Satzes können wir behaupten, daß wenn die Funktion $w=f(z)=z+a_2z^2+\ldots$ eine schlichte konforme Abbildung des Einheitskreises der z-Ebene auf das Gebiet $\mathfrak{G}$ verwirklicht (so daß also der innere Radius r_0 von $\mathfrak{G}$ in bezug auf $w=0$ gleich 1 ist), so gilt für eine Funktion $w^*=f^*(z)$, welche eine schlichte Abbildung des Einheitskreises der z-Ebene auf das Gebiet $\mathfrak{G}^*$ mit $f^*(0)=0$ und $f^{*\prime}(0)=K\neq 0$ liefert:

$$|K|\geqq 1.$$

Es bedeutet hier keine Einschränkung der Allgemeinheit, wenn wir die ersten Ableitungen, $f'(0)$ und $f^{*\prime}(0)=K$ reell und positiv annehmen, also statt der vorigen Ungleichung $K\geqq 1$ schreiben.

Die Funktion $w^*=f^*(z)$ hat die Potenzreihenentwicklung

$$f^*(z)=Kz+a_2^*z^2+\ldots$$

[1] Diese Symmetrisierung von $\mathfrak{G}$ kann ausgeführt werden, wenn es nur im Randpunkte C „ϱ_0-konvexartig" ist. Im übrigen kann $\mathfrak{G}$ einen beliebig komplizierten Rand haben.

und die Funktion

$$\frac{1}{K}f^*(z) = z + \frac{a_2^*}{K}z^2 + \dots$$

verwirklicht eine konforme Abbildung der Kreisscheibe $|z|<1$ auf ein Gebiet der w-Ebene, welches aus $\mathfrak{G}^*$ durch eine Ähnlichkeitstransformation in bezug auf das Ähnlichkeitszentrum $w=0$ im Verhältnis K:1 entstanden ist. Nach dieser Transformation müssen wir ϱ_0 durch $\frac{\varrho_0}{K}$ und R durch $\frac{R}{K}$ ersetzen. Das entstandene Gebiet ist ein solches, für welches die Ungleichung (1) bereits (s. [1]) bewiesen wurde, sodaß wir behaupten können, daß die Ungleichung

$$\frac{R}{K} \geqq R_0\left(\frac{\varrho_0}{K}\right)$$

gilt. Es ist nun nicht schwer zu zeigen, daß

$$KR_0\left(\frac{\varrho_0}{K}\right) \geqq R_0(\varrho_0)$$

ist. Dies folgt daraus, daß $K \geqq 1$ und, daß die Kurve, welche die Abhängigkeit der Größe $R_0(\varrho_0)$ von ϱ_0 veranschaulicht, wie man sich z. B. durch zweimaliges Differenzieren von $R_0(\varrho_0)$ nach ϱ_0 überzeugen kann, konkav ist, (s. Fig. 2), woraus schließlich folgt:

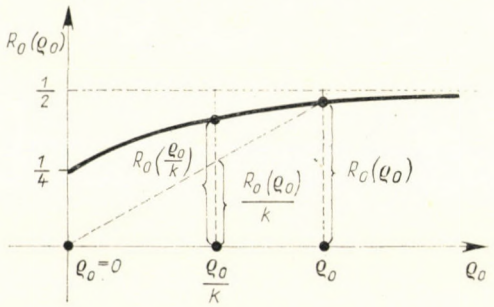


Fig. 2

$$R \geqq R_0(\varrho_0),$$

womit die Behauptung des Satzes bewiesen ist. Für den Grenzfall $\varrho_0 \to 0$ ergibt sich die Koebesche Konstante $\frac{1}{4}$; $(R \geqq \frac{1}{4})$.

LITERATURVERZEICHNIS

[1] SZILÁRD, K.: Über die Verzerrungseigenschaften der konformen Abbildung des Einheitskreises auf „ϱ_0-konvexe" Gebiete, I, *Studia Sci. Math. Hung.* **1** (1966) 133—136.

[2] PÓLYA, G. und SZEGŐ, G.: *Isoperimetric Inequalities in Mathematical Physics*, Princeton, 1951.

[3] HAYMAN, W. K.: *Multivalent Functions*, Cambridge Tracts in Mathematics and Mathematical Physics, No. 48, Cambridge, 1958.

MATHEMATISCHES INSTITUT DER UNGARISCHEN AKADEMIE DER WISSENSCHAFTEN, BUDAPEST

(Eingegangen: 26. April, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 53—61.

ANWENDUNG DER THEORIE DER DIFFERENTIALUNGLEICHUNGEN AUF ZWEI NEUE RANDWERTAUFGABEN FÜR PARABOLISCHE DIFFERENTIALGLEICHUNGEN

von

H. BECKER

Bei Wärmeleitungsproblemen wird gewöhnlich eine der drei Randwertaufgaben behandelt: Außer der Anfangsverteilung der Temperatur in einem Körper ist auf dem Rande für jeden späteren Zeitpunkt entweder die Temperatur oder der Wärmefluß oder eine Kombination beider vorgeschrieben. Man denkt sich etwa bei der ersten Randwertaufgabe (RWA) den Körper D in ein Wärmebad gebracht, das dem Rand ∂D die vorgeschriebene Temperatur verleiht und dessen Wärmekapazität im Vergleich zu der des Körpers so groß ist, daß der Wärmefluß durch ∂D die Temperatur $U(t)$ des Bades nicht beeinflußt. G. FREUD [1] und G. ADLER [2, 3] ließen diese Voraussetzung fallen. Damit ist — außer für die Zeit $t=0$ — die Temperatur $U(t)$ selbst unbekannt; statt dessen besteht eine Wärmebilanz, die $U(t)$ mit den äußeren Wärmequellen und dem Wärmefluß durch ∂D verbindet. Entsprechend läßt sich auch die dritte RWA modifizieren.

Die so entstehenden neuen Probleme nennen wir hier (im Anschluß an ADLER) vierte bzw. fünfte RWA der Wärmeleitungsgleichung. Für diese Aufgaben hat ADLER [3] Existenz- und Eindeutigkeitssätze angegeben.

In der vorliegenden Arbeit werden die Eindeutigkeitssätze unabhängig von der Existenztheorie mit Hilfe der Theorie der Differentialungleichungen hergeleitet. Die Stärke dieser vergleichsweise elementaren Methode besteht darin, daß ohne Vergrößerung des Aufwandes sofort Aussagen über große Klassen nichtlinearer Probleme und allgemeiner Bereiche gewonnen werden können. Den Ausgangspunkt bildet ein dem Problem angepaßtes Lemma von Nagumo-Westphal, wonach sich unter einfachen Voraussetzungen Ungleichungen auf dem Rande ins Innere fortpflanzen. In der dann folgenden Verschärfung dieses Lemmas, bei der in Voraussetzung und Behauptung die $<$-Zeichen durch $\leqq$ ersetzt werden, steckt bereits ein Eindeutigkeitssatz. Außerdem läßt sich aus dem Lemma ein Abschätzungssatz herleiten für Lösungen „benachbarter" Differentialgleichungen mit „benachbarten" Randbedingungen. Aus diesem Abschätzungssatz liest man unmittelbar einen Satz über die stetige Abhängigkeit einer Lösung von den Randwerten ab. Damit ist unser Vorgehen bei den in §1 formulierten RWAn skizziert. Die in § 2 wiedergegebenen Hilfssätze spielen eine zentrale Rolle in der Theorie der Differentialungleichungen; wir entnehmen sie der Monographie [4], an die wir uns auch sonst in den Bezeichnungen und der Darstellung sehr eng anschließen. Der § 3 bringt ein zu den betrachteten Problemen gehöriges Lemma von Nagumo-Westphal und den daraus folgenden Eindeutigkeitssatz. An einem Gegenbeispiel wird gezeigt, daß eine wünschenswerte Erweiterung dieses Lemmas nicht erwartet werden kann. Wie in [4] lassen sich aus den Hilfssätzen und dem Lemma eine Reihe von Ergebnissen her-

leiten, von denen wir in § 4 zwei einfache Abschätzungssätze wiedergeben, die die stetige Abhängigkeit von den Randbedingungen und der rechten Seite der Differentialgleichung enthalten. — Bemerkenswert ist, daß die lokalen Methoden der Differentialungleichungen auch hier zugkräftig sind, wo eine integrale Nebenbedingung besteht.

Ich möchte Herrn Prof. Dr. Wolfgang Walter, der diese Arbeit anregte und in Diskussionen förderte, meinen Dank aussprechen. Herrn Dipl.-Phys. M. Lenhard danke ich für Erörterungen im Zusammenhang mit dem Gegenbeispiel des § 3.

§ 1. Bezeichnungen. Problemstellung

Es sei J_0 das Intervall $0<t\leqq T$, J das Intervall $0\leqq t\leqq T$. Die Klasse aller (reellwertigen) Funktionen $\Phi(t)$, die auf J stetig und auf J_0 differenzierbar sind, nennen wir Z. Der m-dimensionale Euklidische Raum, d. h. die Menge aller m-tupel reeller Zahlen $x=(x_1, \ldots, x_m)$, versehen mit der Norm $|x|=\sqrt{x_1^2+\ldots+x_m^2}$, bezeichnen wir mit E^m. In diesem Raum sei ein (offenes) beschränktes Gebiet D vorgelegt, dessen Berandung ∂D einen endlichen $m-1$-dimensionalen Inhalt hat (es genügt vorauszusetzen, daß ∂D bezüglich eines $m-1$-dimensionalen Maßes, etwa des Haußdorffschen Maßes, meßbar ist und ein endliches Maß besitzt). Ferner sei $\overline{D}=D\cup\partial D$ die abgeschlossene Hülle von D und G das topologische Produkt von J_0 und D. Der Teil des Randes von G, der von den Mengen $\{0\}\times\overline{D}$ und $J_0\times\partial D$ gebildet wird, heiße $\partial' G$.

Für eine auf G einmal nach t und zweimal nach $x_1, \ldots, x_m$ differenzierbare Funktion $\varphi(t, x)$ bedeutet φ_t die Ableitung nach t, φ_x den Vektor der ersten Ableitungen nach $x_1, \ldots, x_m$ und φ_{xx} die $m\times m$-Matrix der zweiten Ableitungen. Die Funktion $f(t, x, z, p, r)$ sei definiert für alle $(t, x)\in G$, $(z, p)\in M$, $r\in M_r$, worin M eine Teilmenge des E^{m+1} und M_r eine Menge symmetrischer $m\times m$-Matrizen ist. Die Differentialgleichung

$$u_t=f(t, x, u, u_x, u_{xx})$$

heißt parabolisch, wenn

$$f(t, x, z, p, r)\leqq f(t, x, z, p, \bar{r}) \tag{1}$$

gilt, falls die Argumente im Definitionsbereich von f liegen und $\bar{r}-r$ positiv semidefinit ist.

Es bezeichnet Z^* die Klasse aller Funktionen $\varphi(t, x)$ mit den folgenden Eigenschaften: 1. φ ist auf $\overline{G}=J\times\overline{D}$ stetig und in G einmal nach t und zweimal stetig nach x differenzierbar. 2. In jedem Punkt $(t, x)\in J_0\times\partial D$ besitzt φ eine äußere Normalableitung φ_n im Sinne von WALTER [4, S. 222], d. h. für jedes $x\in\partial D$ ist eine gegen x konvergente Folge $x_{(k)}\in D$ ausgezeichnet, und der Grenzwert

$$\varphi_n(t, x)=-\lim_{k\to\infty}\frac{\varphi(t, x_{(k)})-\varphi(t, x)}{|x_{(k)}-x|} \tag{2}$$

existiert. 3. Die Funktion φ_n ist für jedes feste $t\in J_0$ über ∂D integrierbar.

Die im folgenden auftretenden Integrale erstrecken sich stets über ∂D.

$$P\varphi = \varphi_t - f(t, x, \varphi, \varphi_x, \varphi_{xx}) \qquad (\varphi \in Z^*) \tag{3}$$

heißt Defekt von φ.

Die unten angegebenen Eindeutigkeitsaussagen beruhen auf einer verallgemeinerten Lipschitzabschätzung mittels einer Funktion $\omega(t, z)$. Die Funktion $\omega(t, z)$ gehört zur Klasse E bzw. $\tilde{E}$, wenn sie für $t \in J_0$ und $z \geqq 0$ erklärt ist und die folgende Eigenschaft besitzt: Zu jedem $\varepsilon > 0$ gibt es ein $\delta > 0$ und ein $\varrho(t) \in Z$, so daß

$$E\colon\ \delta \leqq \varrho \leqq \varepsilon, \quad \varrho' > \omega(t, \varrho), \quad \varrho' > 0 \quad \text{in} \quad J_0 \tag{4a}$$

$$\tilde{E}\colon\ \delta \leqq \varrho \leqq \varepsilon, \quad \varrho' > \omega(t, \varrho) + \delta, \quad \varrho' > \delta \quad \text{in} \quad J_0 \tag{4b}$$

gilt.

Faßt man $\varrho(t) \in Z$ als in $\bar{G}$ definierte Funktion auf, so ist $\varrho \in Z^*$.

Wir beschäftigen uns hier mit Problemen der folgenden Art:

Gesucht werden eine Lösung $u(t, x) \in Z^*$ der parabolischen Differentialgleichung

$$u_t = f(t, x, u, u_x, u_{xx}) \qquad ((t, x) \in G) \tag{5a}$$

und eine Funktion $U(t) \in Z$, die den Randbedingungen für $t = 0$

$$u(0, x) = u_0(x) \qquad (x \in \bar{D}) \tag{5b}$$

$$U(0) = U_0, \tag{5c}$$

der Beziehung

$$\int u_n(t, x)\, do + U'(t) = Q(t) \qquad (t \in J_0) \tag{5d}$$

sowie der Randbedingung auf $J_0 \times \partial D$

$$R[u, U] = 0 \tag{5e}$$

genügen.

Der Randausdruck $R = 0$ hat die Gestalt

$$u(t, x) - U(t) = 0 \qquad \textit{(vierte RWA)} \tag{5e$_1$}$$

oder

$$u_n(t, x) + \vartheta(t, x, u(t, x), U(t)) = 0 \qquad \textit{(fünfte RWA)}. \tag{5e$_2$}$$

Dabei sind u_0 und Q gegebene stetige Funktionen auf $\bar{D}$ bzw. J und U_0 eine gegebene Konstante. Die Funktion $\vartheta(t, x, z_1, z_2)$ sei für $t \in J_0$, $x \in \partial D$, $z_{1,2} \in E^1$ definiert, in z_1 monoton wachsend und in z_2 stark monoton fallend. Dies ist z. B. der Fall, wenn bei (5e$_2$) eine Randbedingung dritter Art

$$u_n + \alpha(u - U) = 0 \qquad (\alpha > 0) \tag{6}$$

vorliegt.

Die Beziehung (5d) stellt die eingangs erwähnte Bilanz dar. Bedeutet u die Temperatur im Körper D, U die des Wärmebades B, so ist in geeigneten Einheiten

und bei passender Normierung des Oberflächenmaßes $\int u_n do$ die nach D hineinfließende Wärmemenge, U' die zur Erwärmung von B benötigte und Q die von außen zugeführte Wärmemenge, jeweils bezogen auf die Zeiteinheit.

Ohne Änderungen in den Beweisen bleibt das folgende übrigens gültig, wenn man das Integral über ∂D ersetzt durch eine Schar monotoner Funktionale $F(t; .)$, die erklärt sind für Funktionen auf ∂D. Lediglich die dritte Eigenschaft der Funktionen aus Z^* muß geändert und dem Definitionsbereich dieser Funktionale angepaßt werden.

§ 2. Hilfssätze

Die beiden folgenden Hilfssätze sind für unser Vorgehen grundlegend. Der erste ist im wesentlichen der Satz 8. II in [4], der zweite ist ein auf die vorliegenden Verhältnisse zugeschnittener Spezialfall von Satz 24. I in [4]. Auf die Beweise darf deshalb hier verzichtet werden.

HILFSSATZ 1: Es seien $\Phi(t)$, $\Psi(t)$ zwei in $I_0: 0<t\leqq\bar{t}$ differenzierbare, in $0\leqq t\leqq\bar{t}$ stetige Funktionen mit der Eigenschaft:

(7) Ist $\Phi(t_0)=\Psi(t_0)$ für ein $t_0\in I_0$, so gilt $\Phi'(t_0)<\Psi'(t_0)$.

Dann liegt genau einer der beiden Fälle vor:

(8a) $\Phi<\Psi$ in I_0,

(8b) Es gibt ein $t^*>0$ mit $\Phi\geqq\Psi$ in $0\leqq t\leqq t^*$.

HILFSSATZ 2: Es seien $\varphi(t, x)$, $\psi(t, x)$ zwei Funktionen aus Z^* mit der Eigenschaft:

(9) Ist $\varphi=\psi$, $\varphi_x=\psi_x$ und $\psi_{xx}-\varphi_{xx}$ positiv semidefinit an einer Stelle in G, so gilt an dieser Stelle $\varphi_t<\psi_t$. Dann liegt genau einer der beiden Fälle vor:

(10a) $\varphi<\psi$ in G.

(10b) Es existiert ein $(\bar{t},\bar{x})\in\partial' G$ so daß

(11) $\varphi(t, x)<\psi(t, x)$ für $0<t\leqq\bar{t}$, $x\in D$

und

(12) $\varphi(\bar{t}, \bar{x})\geqq\psi(\bar{t}, \bar{x})$.

Man beachte, daß für $\bar{t}>0$ der Punkt $\bar{x}$ auf ∂D liegt und daß die Ungleichung (11) noch bei $t=\bar{t}$ erfüllt ist; wegen der Stetigkeit von φ und ψ in $\bar{G}$ gilt dann in (12) die Gleichheit:

(13) $\varphi(\bar{t}, \bar{x})=\psi(\bar{t}, \bar{x})$.

§ 3. Ein Nagumo-Westphal-Lemma

Die Funktionen Φ, Ψ seien aus Z, die Funktionen φ, ψ aus Z^*, und es gelte

(14a) $P\varphi < P\psi \quad in \quad G = J_0 \times D$

(14b) $\varphi(0, x) < \psi(0, x) \quad für \quad x \in \overline{D}$

(14c) $\Phi(0) < \Psi(0)$

(14d) $\int \varphi_n do + \Phi'(t) < \int \psi_n do + \Psi'(t) \quad für \quad t \in J_0$

(14e) $R[\varphi, \Phi] = 0, \; R[\psi, \Psi] = 0 \quad auf \quad J_0 \times \partial D.$

Dann ist

$$\varphi < \psi \quad in \quad G, \; \Phi < \Psi \quad in \quad J_0.$$

Der Beweis beruht auf einer gleichzeitigen Anwendung der beiden Hilfssätze. Wir weisen zunächst (9) nach. Ist $\varphi = \psi$, $\varphi_x = \psi_x$ und $\psi_{xx} - \varphi_{xx}$ positiv semidefinit an einer Stelle in G, so gilt mit (1), (3) und (14a) dort

$$0 < P\psi - P\varphi = \psi_t - \varphi_t - f(t, x, \psi, \psi_x, \psi_{xx}) + f(t, x, \psi, \psi_x, \varphi_{xx}) \leqq \psi_t - \varphi_t.$$

Der Hilfssatz 2 ist also anwendbar. Nehmen wir an, es liege der Fall (10b) vor. Wegen (12) und (14b) ist $\bar{t} > 0$. Nun soll der Hilfssatz 1 auf das Intervall $I_0 : 0 < t \leqq \bar{t}$ und die Funktionen Φ, Ψ angewandt werden. Der Nachweis von (7) und die weiteren Schlüsse müssen für die vierte und fünfte RWA getrennt geführt werden. 1°. Hat (14e) die Form

(15) $$\varphi(t, x) = \Phi(t), \quad \psi(t, x) = \Psi(t) \quad (t \in J_0, \; x \in \partial D)$$

und ist $\Phi = \Psi$ in $t_0 \in I_0$, dann ist $\varphi(t_0, x) = \psi(t_0, x)$ auf ∂D, also wegen (11) und (2) $\varphi_n(t_0, x) \leqq \psi_n(t_0, x)$ auf ∂D. Hieraus und aus (14d) folgt $\Phi'(t_0) < \Psi'(t_0)$; damit ist (7) erfüllt. Da der Fall (8b) wegen (14c) nicht eintreten kann, gilt nach dem ersten Hilfssatz $\Phi < \Psi$ auf I_0, insbesondere an der Stelle $\bar{t}$. 2°. Aus $\Phi(\bar{t}) < \Psi(\bar{t})$ und (15) folgt $\varphi(\bar{t}, x) < \psi(\bar{t}, x)$ auf ∂D. Dies steht für $x = \bar{x}$ im Widerspruch zu (13). Also liegt nicht (10b), sondern (10a) vor.

3°. Nun zeigen wir, daß auch für die fünfte RWA der Fall (10a) vorliegt. Die Randbedingung (14e) lautet jetzt

(16) $$\varphi_n + \vartheta(t, x, \varphi, \Phi) = 0, \quad \psi_n + \vartheta(t, x, \psi, \Psi) = 0 \qquad (t \in J_0, \; x \in \partial D).$$

Mit (14d) ergibt sich aus $\Phi = \Psi$ in $t_0 \in I_0$ und aus der Annahme (10b)

$$\Phi'(t_0) - \Psi'(t_0) < \int \vartheta(t_0, x, \varphi, \Phi(t_0))\, do - \int \vartheta(t_0, x, \psi, \Phi(t_0))\, do \leqq 0,$$

letzteres wegen der Monotonie von ϑ. Damit ist auch hier (7) erfüllt. Weil der Fall (8b) wegen (14c) nicht eintreten kann, gilt wieder $\Phi < \Psi$ auf I_0. 4°. Aus $\Phi(\bar{t}) < \Psi(\bar{t})$, (13), (16) und der starken Monotonie von ϑ im letzten Argument ergibt sich

$$\varphi_n(\bar{t}, \bar{x}) - \psi_n(\bar{t}, \bar{x}) = \vartheta(\bar{t}, \bar{x}, \varphi, \Psi(\bar{t})) - \vartheta(\bar{t}, \bar{x}, \varphi, \Phi(\bar{t})) < 0.$$

Aus (11), (13) folgt aber $\varphi_n \leqq \psi_n$ in $(\bar{t}, \bar{x})$. Dieser Widerspruch zeigt, daß auch hier (10a) gilt.

5°. Gehen wir die Punkte 1°. bzw. 3°. erneut durch, wobei wir $\bar{t}$ durch T, d. h. I_0 durch J_0 ersetzen und an Stelle von (10b) die bereits bewiesenen Ungleichungen $\varphi < \psi$ in G, $\varphi \leqq \psi$ in $\bar{G}$ benutzen, so ergibt sich $\Phi < \Psi$ in J_0. Das Lemma ist vollständig bewiesen.

Die folgende Abwandlung des Lemmas läßt sich auch für die fünfte RWA leicht durchführen, wenn ϑ die spezielle Form

$$\vartheta(t, x, u, U) = \vartheta^*(t, x, u - U) \tag{17}$$

besitzt; $\vartheta^*(t, x, z)$ soll dabei bzgl. z stark monoton wachsend sein. Die Randbedingung (6) ist z. B. von dieser Gestalt.

COROLLAR: *Es seien* $V, W \in Z$ *und* $v, w \in Z^*$. *Die Funktion* f *genüge für ein* $\omega \in E$ *der einseitigen Abschätzung*

$$f(t, x, w+z, w_x, w_{xx}) - f(t, x, w, w_x, w_{xx}) \leqq \omega(t, z) \quad \textit{für} \quad z > 0, \tag{18}$$

falls die linke Seite definiert ist. Ferner gelte

$$Pv \leqq Pw \quad \textit{in} \quad G \tag{19a}$$

$$v(0, x) \leqq w(0, x) \quad \textit{in} \quad \bar{D} \tag{19b}$$

$$V(0) \leqq W(0) \tag{19c}$$

$$\int v_n do + V'(t) \leqq \int w_n do + W'(t) \quad \textit{in} \quad J_0 \tag{19d}$$

$$R[v, V] = 0, \; R[w, W] = 0 \quad \textit{in} \quad J_0 \times \partial D. \tag{19e}$$

(Liegt die fünfte RWA vor, so möge (17) *gelten.) Dann ist*

$$v \leqq w \quad \textit{in} \quad \bar{G}, \quad V \leqq W \quad \textit{in} \quad J.$$

BEWEIS: Zu $\varepsilon > 0$ wählen wir ein $\delta > 0$ und ein $\varrho \in Z$ derart, daß (4a) gilt, und setzen $\varphi = v$, $\Phi = V$, $\psi = w + \varrho$, $\Psi = W + \varrho$. Dann sind, wie wir sehen werden, die Voraussetzungen des Lemmas erfüllt. Die Ungleichungen (14b, c) sind evident. Da $\varrho_n = 0$ und $\varrho' > 0$ ist, gilt auch (14d). (14e) folgt ebenfalls sofort, im Falle der fünften RWA wegen (17). Der Nachweis von (14a) bedarf einer kleinen Rechnung:

$$P\psi - P\varphi = \varrho' + w_t - v_t - f(t, x, w+\varrho, w_x, w_{xx}) + f(t, x, v, v_x, v_{xx}) \geqq$$

$$\geqq \varrho' + Pw - Pv - \omega(t, \varrho) > 0.$$

Also ist $v < w + \varepsilon$ und $V < W + \varepsilon$ in G bzw. J_0. Weil ε beliebig war, folgt die Behauptung $v \leqq w$ in $\bar{G}$ und $V \leqq W$ in J.

EINDEUTIGKEITSSATZ: *Genügt* f *einer Abschätzung* (18), *dann hat die RWA* (5a—e), (17) *höchstens eine Lösung* (u, U).

Für zwei Lösungen (v, V), (w, W) liest man aus dem Corollar $v \leqq w$ und $w \leqq v$, also $v = w$ in $\bar{G}$ ab. Analog folgt $V = W$ in J. Auf die Einschränkung (17) kann im Corollar und im Eindeutigkeitssatz verzichtet werden, wenn ϑ einseitigen Abschätzungen in u und U und f einer verschärften Abschätzung (vgl. [4, 25. VIII δ]) genügen. Mit geringen Änderungen verläuft dann der Beweis des Corollars wie der des Lemmas.

BEMERKUNG: Dem Lemma scheint ein Mangel anzuhaften: In (14e) und (19e) muß das Gleichheitszeichen gelten. Das engt für Rechnungen den Kreis der Ober- und Unterfunktionen stark ein und macht Einschränkungen im folgenden Abschätzungssatz nötig. Es wäre darum wünschenswert, z. B. (19e) bzgl. der vierten RWA durch „$v \leqq V$, $w \geqq W$ auf $J_0 \times \partial D$" (oder die umgekehrten Ungleichungen) zu ersetzen. Die folgende Betrachtung zeigt jedoch, daß dann Lemma und Corollar falsch werden. Es sei $w(t, r)$ die Lösung der ersten RWA der Wärmeleitungsgleichung $u_t = \Delta u$ für die Kugel D: $|x| = r < R$ im E^3 mit $w(0, r) \equiv 0$ und $w(t, R) = h(t) = t(t_0 - t)$ für $t \geqq 0$; $t_0 > 0$. Dazu konstruieren wir durch die Beziehung

$$\int_D w(t, x)\, dx + W(t) = 0 \qquad (t \geqq 0) \tag{20}$$

eine Funktion $W \in Z$, die wegen $w_t = \Delta w$ und der Greenschen Formel der Gleichung

$$\int w_n do + W'(t) = 0 \qquad (t > 0)$$

genügt. Da $w(t, r) > 0$ in $(0 < t \leqq t_0) \times D$ gilt, ist $W(t)$ für $0 < t \leqq t_0$ negativ. Dann existiert rechts von t_0 eine Stelle t_1 derart, daß $w(t, R) = h(t) \geqq W(t)$ für $0 \leqq t \leqq t_1$ und $w(t_1, R) < 0$ ist. Auf Grund der Stetigkeit von w in $J \times \bar{D}$ gilt $w(t_1, r) < 0$ für passende $r < R$.

Setzt man noch $v \equiv 0$, $V \equiv 0$ und wählt man im Corollar für J das Intervall $0 \leqq t \leqq t_1$, so sind alle Voraussetzungen (19a—e) bzgl. der vierten RWA erfüllt mit Ausnahme der Abschwächung: $w(t, x) \geqq W(t)$ auf $J_0 \times \partial D$ von (e), und es gilt $v > w$ für $t = t_1$ und gewisse Punkte aus D sowie $V > W$ in J_0.

Ähnlich zeigt man, daß auch eine entsprechende Abschwächung des Lemmas nicht gelten kann; man wähle etwa $\varphi = \Phi = -\varepsilon(1 + t)$ mit $\varepsilon > 0$ und $\psi = w$, $\Psi = W$ mit den oben konstruierten Funktionen w, W.

Die umgekehrte Abänderung von (19e) in $w \leqq W$ ist auf Grund der Verhältnisse bei der ersten RWA (vgl. [4, 25. II]) sicher nicht möglich. Anhand des obigen Gegenbeispiels sieht man dies ein, wenn man h durch $-h$ und folglich (w, W) durch $(-w, -W)$ ersetzt: Für $0 < t \leqq t_0$ trifft dann die Behauptung des Corollars nicht mehr zu.

Auch für die fünfte RWA kann man auf demselben Prinzip beruhende Beispiele dafür finden, daß in der Randbedingung (19e) $w_n + \alpha(w - W) = 0$ $(\alpha > 0)$ nicht das $\geqq$-Zeichen zulässig ist. Betrachtet man unter sonst gleichen Gegebenheiten wie oben die dritte RWA $w_n + \alpha(w - h) = 0$ für die Kugel D, so gilt im Intervall $0 \leqq t \leqq t_1$ die Ungleichung $w_n + \alpha(w - W) \geqq 0$, und durch Wahl von α läßt sich auch hier $w(t_1, r) < 0$ in der Nähe des Randes erreichen, da nach (20) $W(t_1) < 0$ gelten muß.

Daß in diesem Falle die Randbedingung nicht verletzt werden darf, ist nicht so überraschend wie im Falle der vierten RWA, weil schon bei gewissen Sätzen über die dritte RWA (vgl. [4, 31. X, XII]) die Randbedingung exakt erfüllt werden muß.

§ 4. Abschätzungssatz

Mit den von WALTER [4] behandelten Methoden lassen sich auch in unserem Falle Abschätzungssätze beweisen, wobei man sich hier auf das Lemma aus § 3 beruft. Ein Beispiel dafür ist der folgende

ABSCHÄTZUNGSSATZ: *Es seien* $V, W \in Z$ *und* $v, w \in Z^*$. *Die Funktion* f *genüge den Abschätzungen*

$$f(t, x, w, w_x, w_{xx}) - f(t, x, w - \varrho, w_x, w_{xx}) \leqq \omega(t, \varrho)$$

(21a) $$f(t, x, w + \overline{\varrho}, w_x, w_{xx}) - f(t, x, w, w_x, w_{xx}) \leqq \overline{\omega}(t, \overline{\varrho}),$$

wobei Funktionen $\varrho, \overline{\varrho} \in Z$ *und Funktionen* $\delta, \overline{\delta}$ *auf* J *existieren mögen, die den Ungleichungen*

(21b) $$\varrho' > \omega(t, \varrho) + \delta(t), \quad \overline{\varrho}' > \overline{\omega}(t, \overline{\varrho}) + \overline{\delta}(t) \quad \text{in} \quad J_0$$

genügen. Ferner gelte

(22a) $$Pv = 0, \quad -\overline{\delta}(t) \leqq Pw \leqq \delta(t) \quad \text{in} \quad G$$

(22b) $$-\overline{\varrho}(0) < w - v < \varrho(0) \quad \text{für} \quad t = 0, \; x \in \overline{D}$$

(22c) $$-\overline{\varrho}(0) < W(0) - V(0) < \varrho(0)$$

(22d) $$-\overline{\varrho}' < \int w_n do + W' - \int v_n do - V' < \varrho' \quad \text{in} \quad J_0$$

(22e) $$R[v, V] = 0, \quad R[w, W] = 0 \quad \text{auf} \quad J_0 \times \partial D.$$

(Wenn die fünfte RWA vorliegt, soll (17) *gelten.)*

Dann ist

$$-\overline{\varrho}(t) < w - v < \varrho(t) \quad \text{in} \quad G, \quad -\varrho(t) < W - V < \varrho(t) \quad \text{in} \quad J_0.$$

Zum Beweis setzen wir $\varphi = w - \varrho, \psi = v, \Phi = W - \varrho, \Psi = V$ und prüfen die Voraussetzungen des Nagumo—Westphal-Lemmas nach. Nur der Nachweis von (14a) sei erwähnt, die übrigen sind noch einfacher:

$$P\varphi - P\psi = w_t - \varrho' - f(t, x, w - \varrho, w_x, w_{xx}) \leqq Pw - \varrho' + \omega(t, \varrho) \leqq$$
$$\leqq \delta(t) - \varrho' + \omega(t, \varrho) < 0.$$

Also gilt $\varphi < \psi$ in G, $\Phi < \Psi$ in J_0, und das ist die eine Hälfte der Behauptung; die andere ergibt sich entsprechend. Es sei darauf hingewiesen, daß die im Satz vorkommenden Schranken $\varrho, \overline{\varrho}, \delta, \overline{\delta}$ keineswegs positiv sein müssen. Nimmt man dies jedoch an, so kann man dem Satz die folgende einfachere Form geben.

COROLLAR: *Es seien* $U, W \in Z$ *und* $u, w \in Z^*$. *Die Funktionen* $f, \varrho \equiv \overline{\varrho}$ *und* δ *mögen* (21a, b) *erfüllen. Ferner gelte* (17) *und*

(23a) $$Pu = 0, \quad |Pw| \leqq \delta(t) \quad \text{in} \quad G$$

(23b) $$|w - u| < \varrho(0) \quad \text{für} \quad t = 0, \; x \in \overline{D}$$

(23c) $$|W(0) - U(0)| < \varrho(0)$$

(23d) $$\left| \int w_n do + W' - \int u_n do - U' \right| < \varrho' \quad \text{in} \quad J_0$$

(23e) $$R[u, U] = 0 \quad R[w, W] = 0 \quad \text{auf} \quad J_0 \times \partial D.$$

Dann ist $|w - u| < \varrho(t)$ *in* $\overline{G}$, $|W - U| < \varrho(t)$ *in* J.

Die Bedingungen (21a, b) können ohne Kenntnis einer Lösung (u, U) des Problems (5a—e) geprüft werden, die (u, U) betreffenden Angaben in (23b—e) lassen sich der Problemstellung entnehmen. Der Abschätzungssatz gestattet also, falls nur die Existenz einer Lösung gesichert ist, die Güte einer Näherung (w, W) zu beurteilen.

Unter Beachtung der Definition (4b) der Klasse $\tilde{E}$ folgt aus dem Corollar der

SATZ: *Die Funktion f genüge für ein $\omega \in \tilde{E}$ den Abschätzungen* (21a). *Die Lösung (u, U) der RWA* (5a—e), (17) *hängt (wenn sie existiert) stetig von der rechten Seite der Differentialgleichung und den Vorgaben u_0, U_0 und Q ab, falls die Randbedingung auf $J_0 \times \partial D$ exakt erfüllt bleibt: Zu jedem $\varepsilon > 0$ existiert ein $\delta > 0$, so daß aus $W \in Z$, $w \in Z^*$, $|Pw| < \delta$ in G, $|w(0, x) - u_0(x)| < \delta$ in $\bar{D}$, $|W(0) - U_0| < \delta$,*

$$\left| \int w_n do + W' - Q \right| < \delta \quad \text{in} \quad J_0$$

und $R[w, W] = 0$ folgt:

$$|w - u| < \varepsilon \quad \text{in} \quad G, \qquad |W - U| < \varepsilon \quad \text{in} \quad J_0.$$

LITERATURVERZEICHNIS

[1] FREUD, G.: Über Wärmeleitungs- und Diffusionsprobleme mit zusammengesetzten Randbedingungen, I., *Magyar Tud. Akad. Alkalm. Mat. Int. Közl.* **3** (1955) 369—394.

[2] ADLER, G.: Über Wärmeleitungs- und Diffusionsprobleme mit zusammengesetzten Randbedingungen, II., *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **1** (1956) 167—183.

[3] ADLER, G.: Un type nouveau des problèmes aux limites de la conduction de la chaleur, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **4** (1959) 109—127.

[4] WALTER, W.: *Differential- und Integral-Ungleichungen und ihre Anwendungen bei Abschätzungs- und Eindeutigkeitsproblemen*, Springer Tracts in Natural Philosophy, Vol. 2, Springer-Verlag, 1964.

MATHEMATISCHES INSTITUT III DER TECHNISCHEN HOCHSCHULE, KARLSRUHE

(Eingegangen: 27. April, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 63—66.

ON APPROXIMATION BY POSITIVE LINEAR METHODS, I

by

G. FREUD

We refer to the following theorem of P. P. KOROVKIN [4]: Let $C_{2\pi}$ be the space of 2π-periodic continuous functions, $\{A_n\}$ a sequence of linear transformations of $C_{2\pi}$ into $C_{2\pi}$. We denote the A_n-transform of $f(x)$ by $A_n\{f(t);\, x\}$. Let us suppose that all A_n are positive in the following sence: If $f(x) \geqq 0$ for all real x, then $A_n(f;x) \geqq 0$ for all real x. Now, KOROVKIN's theorem states that if for such a sequence $\{A_n\}$ we have for a numerical sequence $\lambda_n \to 0$

$$A_n(1;\, x) \equiv 1,\quad A_n(\cos t;\, x) = \cos x + O(\lambda_n^2), \tag{1}$$

$$A_n(\sin t;\, x) = \sin x + O(\lambda_n^2)$$

then we have for an arbitrary $f \in C_{2\pi}$

$$|A_n(f;\, x) - f(x)| \leqq K_1 \omega(f;\, \lambda_n) \tag{2}$$

where

$$\omega(f;\, \delta) = \max_{|h| \leqq \delta} |f(x+h) - f(x)| \tag{3}$$

is the modulus of continuity of f.

As an immediate consequence of this relation we deduce as a necessary and sufficient condition for

$$\|A_n(f) - f\| \to 0 \tag{3}$$

that (1) should be satisfied with some sequence $\lambda_n \to 0$. This means that a necessary and sufficient condition for (3) to hold for all $f \in C_{2\pi}$ is, that it should hold for just the three functions $f_0 \equiv 1, f_1(x) = \cos x$ and $f_2(x) = \sin x$. A phenomenon of this kind was first observed by H. BOHMAN [1]. As it was proved by P. P. KOROVKIN himself, by a proper choice of the A_n we can achieve, a., that A_n transforms $C_{2\pi}$ into the subspace T_n of trigonometric polynomials of order n at most; b., that $\lambda_n = 0(n^{-2})$. We then conclude

$$f(x) - A_n(f;\, x) = O(1)\omega(f;\, n^{-1})$$

i. e. our sequence is jacksonian.

Though we obtain in this way a rather simple proof of Jackson's theorem, our condition is no more of a genuine test condition type, as we were assuming in (1) a better order of approximation for f_0, f_1 and f_2, as it would follow from (2). The reason of this slight defect of a very beautiful result was pointed out in a former

paper of the author [2]. To obtain a more precise result we must replace $\omega(f;\delta)$ (see [3]), by the smoothness modulus

$$\omega_2(f;\delta)=\max|f(x+h)+f(x-h)-2f(x)|. \tag{4}$$

We then obtained — nevertheless only under additional conditions concerning the A_n — that from (1) follows

$$|A_n(f;x)-f(x)|\leqq K_2\omega_2(f;\lambda_n).$$

In this way, if A_n sends $C_{2\pi}$ into T_n, then (1) to hold with $\lambda_n=0(1/n)$ is necessary as well as sufficient for $\{A_n\}$ being zygmundian. (And — as a consequence — sufficient for being jacksonian). The additional conditions concerning the A_n where that they should commute with the operators of translation T_n and the symmetry operator S, defined by

$$T_h\{f(t);x\}=f(x+h) \quad \text{and} \quad S\{f(t);x\}=f(-x).$$

Recently I found that this additional assumptions can be dropped entirely. Indeed, we are going to prove an even more general statement, where $\omega(f;\delta)$ is replaced by the higher order continuity modulus

$$\omega_m(f;\delta)=\max_{|h|\leqq\delta}\sum_{\nu=0}^{m}\binom{m}{\nu}(-1)^\nu f(x+\nu h) \tag{5}$$

with arbitrary $m>1$ (see "Theorem" to follow).

LEMMA: *Let B be a linear operator $C_{2\pi}\to C_{2\pi}$, $B^{(m)}$ the restriction of B to the space $C_{2\pi}^{(m)}$ of 2π-periodic function having a continuous 2π periodic m-th derivative, then for an arbitrary $f\in C_{2\pi}$ and an arbitrary integer ν we have for all x*

$$\|B(f)\|_{C_{2\pi}}=\operatorname{Max}|B(f;x)|\leqq K_m(\|B\|+\nu^m\|B^{(m)}\|)\omega_m(f;\nu^{-1}) \tag{6}$$

where the norms of the linear operators B, resp. $B^{(m)}$ are defined as

$$\|B\|=\sup_{\|f\|_{C_{2\pi}}\leqq 1}\|B(f)\|_{C_{2\pi}}, \quad \|B^{(m)}\|=\sup_{\substack{f\in C_{2\pi}^{(m)}\\ \|f^{(m)}\|_{C_{2\pi}}\leqq 1}}\|B^m(f)\|_{C_{2\pi}}.$$

This lemma was proved for the case $m=2$ by the author [3], and it was conjectured by the author to hold even for arbitrary integer m. The proof of the general case was given by G. I. SUNOUCHY [5].

We arrived at the point to formulate and prove our main theorem. In all what follows we assume $\{A_n\}$ to be a sequence of linear operators $C_{2\pi}\to C_{2\pi}$, which are positive in the sense explained. Let further $m>1$ be an integer and $\{\lambda_n\}$ a positive null-sequence.

THEOREM: *It is a necessary and sufficient condition that*

$$|A_n(f;x)-f(x)|\leqq K\omega_m(f;\lambda_n) \tag{7}$$

should hold for some fixed K and all $f\in C_{2\pi}$, that

$$A_n(f_0;x)\equiv 1, \quad A_n(f_\nu;x)-f_\nu(x)=O(\lambda_n^m); \qquad \nu=1,2 \tag{8}$$

where

$$f_0(x) \equiv 1, \quad f_1(x) = \cos x, \quad f_2(x) = \sin x.$$

As (8) means exactly that (7) is satisfied for the special cases $f = f_\nu$, $\nu = 0, 1, 2$, we obtained a genuine test condition.

PROOF OF THE THEOREM: we prove that for an arbitrary m-times continuously differentiable 2π periodic function $g(x)$ the relation

$$\|A_n(g) - g\|_{2\pi} \leqq K_1 \lambda_n^{-m} \|g^{(m)}\|_{C_{2\pi}} \tag{9}$$

is satisfied. To this end we observe that for every $f \in C_{2\pi}^{(m)}$ the second derivative of the function

$$g(x) = f(x) - f(t) - f'(t) \sin(t - x)$$

satisfies the inequality

$$|g''(\xi)| = |f''(\xi) - f'(t) \sin(\xi - t)| \leqq \|f'\| + \|f''\| \leqq 2(2\pi)^{m-1} \|f^{(m)}\| \text{ [1]}.$$

We obtain in this way, using $(x-t)^2 = \pi^2 \sin^2 \frac{x-t}{2}$

$$2^{-m} \pi^{m+1} \|f^{(m)}\| \sin^2 \frac{x-t}{2} \leqq \tag{10}$$

$$\leqq f(x) - f(t) - f'(t) \sin(x - t) \leqq 2^m \pi^{m+1} \|f^{(m)}\| \sin^2 \frac{x-t}{2}.$$

Now we observe that — as a consequence of (8) — we have

$$A_n\{\sin(x - t); x\} = O(\lambda_n^m) \tag{11}$$

and

$$A_n\left\{\sin^2 \frac{x-t}{2}; x\right\} = O(\lambda_n^m). \tag{12}$$

In applying the operator $A_n\{g; x\}$ to the inequality (10) [2] and using (11) and (12) we obtain

$$f(x) - A_n(f; x) = O(\lambda_n^m).$$

That means, that for $B(f; x) = A_n(f; x) - f(x)$ we have

$$\|B\| \leqq 2 \quad \text{and} \quad \|B^{(m)}\| = O(\lambda_n^m).$$

[1] For let $f \in C_{2\pi}^{(k)}$, $k \geqq 2$, then $\int_{-\pi}^{\pi} f^{(k-1)}(t)\, dt = f^{(k-2)}(\pi) - f^{(k-2)}(-\pi) = 0$, so that $f^{(k-1)}(\xi) = 0$ for some $\xi \in [-\pi, \pi]$ and then we have $|f^{(k-1)}(x)| = \left| \int_{\xi}^{x} f^{(k)}(t)\, dt \right| < 2\pi \|f^{(k)}\|$, and finally $\|f^{k-1}\| \leqq 2\pi \|f^{(k)}\|$. This establishes the second part of our inequality.

[2] This is legitimate, because A_n is positive.

We insert this inequalities in our lemma and take $v = \left[\frac{1}{\lambda_n}\right]$. It follows

$$\|A_n(f) - f\| = O(1)\omega_m(f; v^{-1}) < O(1)\omega_m(f; \lambda_n).$$

Q. e. d.

REFERENCES

[1] BOHMAN, H.: On approximation of continuous and analytic functions, *Archiv für Mathematik* **2** (1952) 43—52.
[2] FREUD, G.: Über positive Zygmundsche Approximationsverfahren, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6A** (1961) 71—75.
[3] FREUD, G.: Sui procedimenti lineari d'approssimazione, *Atti. Accad. Naz. Lincei. Rend. Cl. Sci. Fis. Mat. Nat.* **26** (1959) 641—643.
[4] КОРОВКИН, П. П.: *Линейные операторы и теория приближения*, Москва, 1959
[5] SUNOUCHI, G.: Remark on page 183 in I. S. N. M. Vol. 5, *On Approximation Theory*, edited by P. L. Butzer and J. Korevaar, Birkhäuser Verl. Basel—Stuttgart, 1964.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received May 10, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 67—72.

ON A THEOREM OF L. ALPÁR CONCERNING FOURIER SERIES OF POWERS OF CERTAIN FUNCTIONS

by

G. HALÁSZ

L. ALPÁR [1], investigating a problem on convergence of Fourier series, has proved the following theorem:

If $f(x)$ is real valued and analytic on the real line, periodical with period 2π then the partial sums

$$\sum_{n=m_1}^{m_2} a_{nv} e^{inx}$$

of the Fourier series

$$e^{ivf(x)} = \sum_{n=-\infty}^{\infty} a_{nv} e^{inx}$$

are bounded by a constant $K = K(f)$ independent of v, m_1, m_2 and x.

His proof is very complicated, using saddle point method for suitably chosen paths. In this paper we give a simpler proof based on VAN DER CORPUT's ideas. There will be no need to deform the path of integration and this, at the same time, enables us to relax the very strict condition of analyticity of $f(x)$. However, for the sake of brevity we shall not consider uniformity in x but confine our attention to a fixed point. Accordingly, the conditions too, imposed on $f(x)$ will be local as given by the following

THEOREM. *Let $f(x)$ be measurable in $[-\pi, \pi]$, twice continuously differentiable in a neighbourhood of 0 satisfying*

$$f''(\pm x) = \text{const } x^p + O(x^{p+\varepsilon}) \qquad (x \to +0)$$

with a $p > -1$, $\varepsilon > 0$, the const $\neq 0$ (*this* const $\neq 0$ *may be different in the two cases* $\pm$)[1]. *If*

$$e^{ivf(x)} \sim \sum_{n=-\infty}^{\infty} a_{nv} e^{inx}$$

then for the partial sums at 0

$$\left| \sum_{n=m_1}^{m_2} a_{nv} \right| < K$$

independently of v, m_1 and m_2.

[1] If $p < 0$ or $p = 0$ and the two constants are not equal then $f''(x)$ does not tend to finite limit as $x \to 0$ and in that case we allow that $f''(0)$ should not exist though $f'(0)$ must exist in any case.

ALPÁR conjectured that the statement holds if only $f''(x)$ is continuous. This was disproved by Y. KATZNELSON, the example can be found in [1]. That function vanished in infinite order at $x=0$, a possibility which is ruled out in our theorem by the basic condition on the asymptotic behaviour of $f''(x)$.

COROLLARY. *If $f(x)$ satisfies the hypotheses of the theorem and $F(u)$ has an absolutely convergent Fourier series then that of $F[f(x)]$ converges (if not absolutely) at $x=0$.*

For the (simple) connection between theorem and corollary the reader is referred to ALPÁR's paper ([2], formuli (1.8), (2.3)). Here we only prove our theorem, but first a lemma of Van der Corput type.

LEMMA. *If $\varphi(x)$ is continuously differentiable in $[a, b]$ and $[-b, -a]$ $(0 \leqq a \leqq b)$ and is of constant sign and monotone separately in each of the two intervals, furthermore*

$$0 < \frac{1}{M_1} \leqq \left| \frac{\varphi(x)}{\varphi(-x)} \right| \leqq M_1 < +\infty \qquad (x \in (a, b)), \tag{1}$$

$$\left| \frac{\varphi(x)}{x\varphi'(x)} \right| \leqq M_2 < +\infty \qquad (x \in (a, b) \cup (-b, -a)), \tag{2}$$

$$\operatorname*{Var}_{(-b, -a) \cup (a, b)} \frac{\varphi(x)}{x\varphi'(x)} \leqq M_3 < +\infty \tag{3}$$

then

$$\left| \int_a^b \frac{e^{i\varphi(x)} - e^{i\varphi(-x)}}{x}\, dx \right| < C(M_1, M_2, M_3)$$

and this bound depends only on M_1, M_2, M_3 and neither on other properties of $\varphi(x)$ nor on a, b.

We split the range of integration $a \leqq x \leqq b$ into two parts according as $|\varphi(x)| \leqq 1$ or $|\varphi(x)| \geqq 1$. Since $\varphi(x)$ is of constant sign and monotone these sets, say I_1 and I_2 are intervals. On I_1 we write the integrand in the form

$$\frac{e^{i\varphi(x)} - 1}{x} + \frac{1 - e^{i\varphi(-x)}}{x}$$

and integrate these separately. Here $|\varphi(-x)| \leqq M_1 |\varphi(x)| \leqq M_1$ and as the constant 1 in the definition of $I_{1,2}$ had no special significance it is enough to consider the first term:

$$\left| \int_{I_1} \frac{e^{i\varphi(x)} - 1}{x}\, dx \right| \leqq \int_{I_1} \frac{|\varphi(x)|}{x}\, dx = \int_{I_1} \left| \frac{\varphi(x)}{x\varphi'(x)} \right| |\varphi'(x)|\, dx \leqq$$

$$\leqq M_2 \int_{I_1} |\varphi'(x)|\, dx = M_2 \left| \int_{I_1} \varphi'(x)\, dx \right| \leqq M_2.$$

For the second term we would get $M_1 M_2$.

In the case of I_2 we leave the numerator in the original form and integrate the two terms separately. Again we may confine ourselves to the first one since in I_2 $|\varphi(-x)| \geqq \frac{1}{M_1} |\varphi(x)| \geqq \frac{1}{M_1}$ a positive constant which is the only essential fact. We have

$$\int_{I_2} \frac{e^{i\varphi(x)}}{x}\, dx = \int_{I_2} \varphi'(x) \frac{e^{i\varphi(x)}}{\varphi(x)} \cdot \frac{\varphi(x)}{x\varphi'(x)}\, dx$$

or by partial integration if $S(t) = \int_t^\infty \frac{e^{iu}}{u}\, du$

$$\left| \int_{I_2} \right| \leqq 2 \max_{x \in I_2} \left| \frac{\varphi(x)}{x\varphi'(x)} \right| \cdot \max_{x \in I_2} |S(|\varphi(x)|)| + \max_{x \in I_2} |S(|\varphi(x)|)| \cdot \operatorname*{Var}_{I_2} \frac{\varphi(x)}{x\varphi'(x)}.$$

But in I_2 $|\varphi(x)| \geqq 1$, consequently $S(|\varphi(x)|)$ is bounded and so all quantities occurring on the right hand side are bounded and the proof is finished.

Before passing to the proof of our theorem, let us make a remark which points out the background of that theorem. The partial sums of the Fourier series can be expressed by an integral similar to that of our lemma and we shall have to apply the lemma to $\varphi(x) = vf(x)$ and similar functions. But v cancels out in each of the conditions, showing that the situation is in fact independent of v. Also, in Alpár's proof to estimate the Fourier coefficients, not their sums it was essential in which order $f''(x)$ vanishes. Here, however, it is not. Let us put e. g. $\varphi(x) = x^q$ then

$$\frac{\varphi(x)}{x\varphi'(x)} = \frac{x^q}{xqx^{q-1}} \equiv \frac{1}{q}$$

a constant which shows that the conditions of the lemma are fulfilled independently of q. Although our functions will not be so simple as x^q but they will turn out to imitate it, possibly with different q's in different intervals.

Now we turn to the proof. The partial sums of the Fourier series in question, beginning with the term $a_{0,v}$ are given by

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ivf(x)} \frac{1 - e^{-imx}}{1 - e^{-ix}}\, dx$$

where $m \geqq 0$ is an integer.

In $[-\pi, \pi]$ $\frac{1}{1 - e^{-ix}} = -\frac{i}{x} + O(1)$ while the numerator cannot exceed 2 in absolute value, making it possible to replace the denominator by x. Also we can disregard integration outside $(-\tau, \tau)$ if τ is fixed and the integral to be estimated

takes the form

$$\int_{-\tau}^{\tau} e^{ivf(x)} \frac{1-e^{-imx}}{x}\,dx = \int_{-\tau}^{\tau} \frac{e^{ivf(x)} - e^{ivf(x)-imx}}{x}\,dx =$$

$$= \int_{0}^{\tau} \frac{e^{ivf(x)} - e^{ivf(-x)}}{x}\,dx - \int_{0}^{\tau} \frac{e^{ivf(x)-imx} - e^{ivf(-x)+imx}}{x}\,dx.$$

It is sufficient to deal with the second integral since it reduces to the first if $m=0$. If we prove, which we shall indeed, the boundedness of the second integral for all real m not necessarily integer then we may even suppose $f'(0)=0$ (substracting from $f(x)$ $f'(0)x$ which can be included into mx) and independently of anything that $f(0)=0$ since it only means a multiplication by the factor $e^{-ivf(0)}$. In this case we have, by integrating the formula satisfied by $f''(x)$, together with itself

$$\begin{aligned} f''(\pm x) &= \text{const}\, x^{p} + O(x^{p+\varepsilon}) \\ f'(\pm x) &= \text{const}\, x^{p+1} + O(x^{p+1+\varepsilon}) \qquad (x \to +0) \\ f(\pm x) &= \text{const}\, x^{p+2} + O(x^{p+2+\varepsilon}). \end{aligned} \tag{4}$$

Now fix τ so the main terms should dominate in these asymptotic expressions. Here and in that follows we mean by the domination of a quantity over an other the exact meaning that the dominating one should be at least twice as much in modulus as the other.

We apply the lemma with

$$\varphi(x) = vf(x) - mx$$

$$\varphi'(x) = vf'(x) - m.$$

Its conditions would not be satisfied on the whole interval $(-\tau, \tau)$ and we have to exclude values where the two terms of $\varphi(x)$ and $\varphi'(x)$ are approximately equal. This happens when

$$|v|\,|x|^{p+2} \sim |mx|$$

$$|v|\,|x|^{p+1} \sim |m|$$

$$|x| \sim \left|\frac{m}{v}\right|^{\frac{1}{p+1}}$$ [2, 3]

and we can find two constants $0<a<A$ such that for $0<x<\tau_1 \overset{\text{def}}{=} \min\left[a\left|\frac{m}{v}\right|^{\frac{1}{p+1}}, \tau\right]$ and $-\tau_1<x<0$ the second, for $\tau_2 \overset{\text{def}}{=} \min\left[A\left|\frac{m}{v}\right|^{\frac{1}{p+1}}, \tau\right] < x < \tau$ and $-\tau < x < -\tau_2$

[2] $\sim$ means that the quotient of the two sides lies between two positive constants.

[3] Assume $m \neq 0$, $v \neq 0$. v too may be any real number, not necessarily integer so that taking limits afterwards, we can get the result for these excluded cases.

the first terms dominate in the expressions of $\varphi(x)$ and $\varphi'(x)$. On the range (τ_1, τ_2), instead of using the lemma, we use the trivial estimation

$$\left| \int_{\tau_1}^{\tau_2} \frac{e^{i\varphi(x)} - e^{i\varphi(-x)}}{x} dx \right| \leqq 2 \int_{\tau_1}^{\tau_2} \frac{1}{x} dx = 2 \log \frac{\tau_2}{\tau_1} \leqq 2 \log \frac{A}{a} \qquad \text{(a constant).}$$

On the remaining parts $(0, \tau_1)$, (τ_2, τ), however, we can verify the requirements of our lemma, and that is what we shall do in the sequel.

In $(0, \tau_1)$, (τ_2, τ) and the symmetric intervals either the first or the second terms of $\varphi(x)$ and $\varphi'(x)$ dominate so that both $\varphi(x)$ and $\varphi'(x)$ are of constant sign and the qualitative properties required are satisfied.

As to (1), in $(-\tau_1, \tau_1)$ the second is the dominating term hence

$$\frac{1}{3} = \frac{\frac{1}{2}|mx|}{\frac{3}{2}|mx|} \leqq \left| \frac{\varphi(x)}{\varphi(-x)} \right| = \left| \frac{vf(x) - mx}{vf(-x) + mx} \right| \leqq \frac{\frac{3}{2}|mx|}{\frac{1}{2}|mx|} = 3$$

while in $(-\tau, -\tau_2) \cup (\tau_2, \tau)$

$$1 = \frac{|x|^{p+2}}{|x|^{p+2}} \sim \frac{\frac{1}{2}|vf(x)|}{\frac{3}{2}|vf(-x)|} \leqq \left| \frac{\varphi(x)}{\varphi(-x)} \right| \leqq \frac{\frac{3}{2}|vf(x)|}{\frac{1}{2}|vf(-x)|} \sim \frac{|x|^{p+2}}{|x|^{p+2}} = 1.$$

Similarly in the case of (2), in $(-\tau_1, \tau_1)$

$$\left| \frac{\varphi(x)}{x\varphi'(x)} \right| = \left| \frac{vf(x) - mx}{x(vf'(x) - m)} \right| \leqq \frac{\frac{3}{2}|mx|}{|x|\frac{1}{2}|m|} = 3,$$

in $[-\tau, -\tau_2], [\tau_2, \tau]$

$$\left| \frac{\varphi(x)}{x\varphi'(x)} \right| \leqq \frac{\frac{3}{2}|vf(x)|}{|x|\frac{1}{2}|vf'(x)|} \sim \frac{|x|^{p+2}}{|x| \cdot |x|^{p+1}} = 1.$$

To verify (3) it will be more tiresome. Since the situation is entirely the same in the symmetric intervals, let us confine ourselves to estimating the variation over $(0, \tau_1) \cup (\tau_2, \tau)$. We have to compute the derivative and integrate it in absolute value.

$$\frac{d}{dx} \frac{\varphi(x)}{x\varphi'(x)} = \frac{x\varphi'^2(x) - \varphi(x)\varphi'(x) - x\varphi(x)\varphi''(x)}{x^2\varphi'^2(x)}. \tag{5}$$

Let us introduce for the numerator the abbreviation $V[\varphi(x)]$. For example $V(x^q) \equiv 0$ since, as we observed already, in this case $\dfrac{\varphi(x)}{x\varphi'(x)}$ is constant, its derivative zero. In the numerator we substitute $\varphi(x) = vf(x) - mx$ and carry out the derivations and multiplications. Those terms to which only mx contributes cancel out because they give $V(mx) \equiv 0$. Next we collect terms originating solely from $vf(x)$. In the numerator it is $V[vf(x)]$ and the whole expression is

$$\frac{V[vf(x)]}{x^2\varphi'^2(x)} = \frac{v^2[xf'^2(x) - f(x)f'(x) - xf(x)f''(x)]}{x^2(vf'(x) - m)^2}.$$

Each term in the square bracket is $O(x^{2p+3})$. Nothing better is needed for the integration over $(0, \tau_1)$ namely in the denominator it is m that dominates and

$$\int_0^{\tau_1} \frac{v^2 x^{2p+3}}{x^2 \; m^2} dx = \frac{v^2}{m^2} \frac{\tau_1^{2p+2}}{2p+2} \leqq \frac{v^2}{m^2} \frac{a^{2p+2} \frac{m^2}{v^2}}{2p+2} = O(1).$$

In the case of the second interval, however, the estimation $O(x^{2p+3})$ would not be strong enough. But we can obtain a better one if we put the asymptotic formulas (4) in the place of $f(x), f'(x), f''(x)$. The main terms of exponent $2p+3$ cancel out, their contibution being $V(x^{p+2}) \equiv 0$, while the O-terms have exponents greater than $2p+3$ by at least ε so that the quantity $V[(f(x)]$ in the square bracket is in fact $O(x^{2p+3+\varepsilon})$ and the whole fraction in (τ_2, τ), owing to the fact that here $vf'(x)$ dominates over m

$$O\left(\frac{v^2 x^{2p+3+\varepsilon}}{x^2 v^2 f'^2(x)}\right) = O\left(\frac{x^{2p+3+\varepsilon}}{x^2 x^{2p+2}}\right) = O(x^{\varepsilon-1})$$

which can be integrated even from 0 to τ.

Now only the mixed terms in (5) remain to be estimated:

$$\frac{vm[-2xf'(x)+f(x)+xf'(x)+x^2 f''(x)]}{x^2(vf'(x)-m)^2}.$$

In the square bracket each term is $O(x^{p+2})$. This bound is applicable to $(0, \tau_1)$ where

$$\int_0^{\tau_1} \frac{|vm| x^{p+2}}{x^2 m^2} dx = O(1) \left|\frac{v}{m}\right| \tau_1^{p+1} = O(1)$$

as well as to (τ_2, τ) where (if $\tau_2 < \tau$)

$$\int_{\tau_2}^{\tau} \frac{|vm| x^{p+2}}{x^2 v^2 f'^2(x)} dx = O(1) \left|\frac{m}{v}\right| \int_{\tau_2}^{\tau} \frac{x^{p+2}}{x^2 x^{2p+2}} dx = O(1) \left|\frac{m}{v}\right| \int_{\tau_2}^{\infty} \frac{1}{x^{p+2}} =$$

$$= O(1) \left|\frac{m}{v}\right| \frac{1}{\tau_2^{p+1}} = O(1)$$

and having verified (3) of our lemma, the proof of the theorem is completed.

REFERENCES

[1] ALPÁR, L.: Sur une classe particulière de séries de Fourier ayant de sommes partielles bornées, *Studia Sci. Math. Hung.* **1** (1966) 189—204.

[2] ALPÁR, L.: Sur certaines transformées des séries de puissance absolument convergentes sur la frontière de leur cercle de convergence, *A Magyar Tudományos Akadémia Matematikai Kutatóintézetének Közleményei* **7** (1962) 287—316.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received May 31, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 73—80.

GENERALIZATION OF TWO THEOREMS OF G. FREUD CONCERNING THE RATIONAL APPROXIMATION

by

J. SZABADOS

1. In his paper [1] G. FREUD has proved among others the following two theorems:

THEOREM 1. *Let $f(x)$ be a continuous function in* [0, 1] *and let us assume that there exist polynomials $p_n^{\{k\}}(x)$ of degree n for which*

$$|f(x)-p_n^{\{k\}}(x)|<\varepsilon_n \quad (x\in I_k=[\xi_k, \xi_{k+1}],\ n=0, 1, 2, \ldots)$$

holds for $k=1, \ldots, s$ where $0=\xi_1<\xi_2<\ldots<\xi_{s+1}=1$ and $\lim_{n\to\infty}\varepsilon_n=0$. Then there exist rational functions $r_N(x)$ of degree N for which

$$(1.1)\quad |f(x)-r_N(x)| = O(1)\left(\varepsilon_{2\left[\frac{N}{4s}\right]}+s\|f\|e^{-\frac{1}{2}\sqrt{2\left[\frac{N}{4s}\right]}}\right) \qquad (x\in[0,1],\ N>8s).$$

Denote by $V^{(r)}(a, b)$ $(r=1, 2, \ldots)$ the class of the functions $f(x)$ which are continuously differentiable $r-1$-times, and $f^{(r-1)}(x)$ is an integralfunction of a function $f^{[r]}(x)$ of bounded variation (in a finite interval $[a, b]$).

THEOREM 2. *If $f(x)\in V^{(r)}(0, 1)$ then there exist rational functions $r_N(x)$ of degree N for which*

$$|f(x)-r_N(x)| = V(f^{[r]})O\left(\frac{\log^2 N}{N^{r+1}}\right) \qquad (x\in[0,1],\ N\geqq 2r+1)$$

where $V(f^{[r]})$ is the total variation of $f^{[r]}(x)$.

In this paper we are going to prove two more general theorems instead of these ones. We shall use the same notations.

THEOREM 1A. *Let $f(x)$ be a continuous function in* [0, 1] *and let us assume that there exist rational functions $r_n^{\{k\}}(x)$ of degree n for which*

$$|f(x)-r_n^{\{k\}}(x)|<\varepsilon_n \qquad (x\in I_k,\ n=0, 1, 2, \ldots)$$

for $k=1, \ldots, s$ where $\lim_{n\to\infty}\varepsilon_n=0$. Further let

$$(1.2)\qquad a_n(\delta_n) = \max_{1\leqq k\leqq s}\ \max_{x\in(J_k\cup J_{k+1})\cap[0,1]} |r_n^{\{k\}'}(x)|$$

where

$$J_k=[\xi_k-\delta_n,\ \xi_k+\delta_n] \qquad (k=1,\ \ldots,\ s+1)$$

($\delta_n>0$ *is such small that* $a_n(\delta_n)$ *is finite, further* $\delta_n<\frac{1}{2}\min\limits_{1\leqq k\leqq s}(\xi_{k+1}-\xi_k)$), *and*

$$M=\sup_n \max_{1\leqq k\leqq s}\max_{x\in I_k}|r_n^{\{k\}}(x)|.$$

Then there exist rational functions $r_N(x)$ *of degree* N *for which*

$$|f(x)-r_N(x)|\leqq 2\varepsilon_n+e^{-\sqrt{n}} \qquad (x\in[0,\ 1],\ n\geqq n_0) \tag{1. 3}$$

where n *is the greatest integer satisfying the inequality*

$$5n+\frac{9}{4}\left\{\log\left[3a_n(\delta_n)+\frac{3M}{2}\left(\frac{9s}{\delta_n}\right)^{2/3}\right]+\sqrt{n}\right\}^2\leqq\frac{N}{s}. \tag{1. 4}$$

THEOREM 2A. *Let* $f(x)$ *be continuous in* [0, 1] *and*

$$f(x)\in V^{(r)}(\xi_k,\ \xi_{k+1}) \qquad (k=1,\ \ldots,\ s;\ 0=\xi_1<\xi_2<\ldots<\xi_{s+1}=1).$$

Then there exist rational functions $r_N(x)$ *of degree* N *for which*

$$|f(x)-r_N(x)|=s^{r+1}V(f^{[r]})O\left(\frac{\log^2 N}{N^{r+1}}\right)+e^{-\sqrt{\frac{N}{14s}}}$$

for sufficiently large N*'s, where* $V(f^{[r]})$ *is the maximum of the total variations of* $f^{[r]}(x)$ *in the intervals* $(\xi_k,\ \xi_{k+1})$ $(k=1,\ \ldots,\ s)$.

2. First we prove that Theorem 1A implies Theorem 1 (this is obvious in connection with Theorem 2 and 2A). For this purpose we have to estimate (1. 2) if $r_n^{\{k\}}(x)$ are especially polynomials. Let

$$X^{\{k\}}=\frac{2}{\xi_{k+1}-\xi_k}\left(x-\frac{\xi_k+\xi_{k+1}}{2}\right)$$

and $T_n(X^{\{k\}})$ the transformated Chebyshev-polynomial. Using the classical Markov's and Bernstein's inequalities, we get for $\delta_n=1/n^2$

$$a_n(\delta_n)\leqq\max_{1\leqq k\leqq s}\left(\frac{2n^2}{\xi_{k+1}-\xi_k+2\delta_n}\max_{x\in[\xi_k-\delta_n,\ \xi_{k+1}+\delta_n]}|r_n^{\{k\}}(x)|\right)\leqq$$

$$\leqq\max_{1\leqq k\leqq s}\left(\frac{2n^2M}{\xi_{k+1}-\xi_k}\max_{x\in[\xi_k-\delta_n,\ \xi_{k+1}+\delta_n]}|T_n(X^{\{k\}})|\right)\leqq$$

$$\leqq\max_{1\leqq k\leqq s}\frac{2n^2M}{\xi_{k+1}-\xi_k}\left(|X^{\{k\}}|+\sqrt{X^{\{k\}2}-1}\right)^n\leqq$$

$$\leqq\max_{1\leqq k\leqq s}\frac{2n^2M}{\xi_{k+1}-\xi_k}\left(1+\frac{2}{\xi_{k+1}-\xi_k}\delta_n+\sqrt{\frac{4}{\xi_{k+1}-\xi_k}\delta_n+\frac{4}{(\xi_{k+1}-\xi_k)^2}\delta_n^2}\right)^n\leqq$$

$$\leqq\max_{1\leqq k\leqq s}\frac{2n^2M}{\xi_{k+1}-\xi_k}\left(1+2\sqrt{\frac{5}{\xi_{k+1}-\xi_k}\delta_n}\right)^n\leqq\max_{1\leqq k\leqq s}\frac{2n^2M}{\xi_{k+1}-\xi_k}e^{2\sqrt{\frac{5}{\xi_{k+1}-\xi_k}}}=O(n^2).$$

Thus the condition (1. 4) means that for sufficiently large n's

$$5n+\frac{9}{4}\left(O(\log n)+\sqrt{n}\right)^2 \leqq 8n \leqq \frac{N}{s} \qquad (n \geqq n_1)$$

i. e. we may choose $n=\left[\frac{N}{8s}\right]$ which shows that the right hands of (1. 1) and (1. 3) are of the same magnitude.

3. PROOF OF THEOREM 1A. We construct a new approximating rational function as follows. Let

$$\bar{R}_n^{\{k\}}(x) = \frac{r_n^{\{k\}}(x)}{1+\eta_n^2 r_n^{\{k\}}(x)^2}$$

where

(3. 1) $$\eta_n = \frac{1}{2M}\left(\frac{9s}{\delta_n e^{\sqrt{m}}}\right)^{1/3},$$

(3. 2) $$m = \left[\frac{9}{4}\left\{\log\left(3a_n(\delta_n)+\frac{3M}{2}\left(\frac{9s}{\delta_n}\right)^{2/3}\right)+\sqrt{n}\right\}^2\right].$$

Clearly

(3. 3) $$\lim_{n\to\infty}\eta_n = 0.$$

$\bar{R}_n^{\{k\}}(x)$ is a rational function of degree $2n$ and has the following properties:

(3. 4) $$|f(x)-\bar{R}_n^{\{k\}}(x)| \leqq |f(x)-r_n^{\{k\}}(x)|+|r_n^{\{k\}}(x)-\bar{R}_n^{\{k\}}(x)| \leqq$$

$$\leqq \varepsilon_n+\frac{|r_n^{\{k\}}(x)|^3}{1+\eta_n^2 r_n^{\{k\}}(x)^2}\eta_n^2 \leqq \varepsilon_n+M^3\eta_n^2 \qquad (x\in I_k,\ k=1,\ldots,s),$$

(3. 5) $$|\bar{R}_n^{\{k\}}(x)| \leqq \frac{1}{2\eta_n} \qquad (x\in[0,1]),$$

(3. 6) $$|\bar{R}_n^{\{k\}'}(x)| = \frac{|1-\eta_n^2 r_n^{\{k\}}(x)^2|}{[1+\eta_n^2 r_n^{\{k\}}(x)]^2}|r_n^{\{k\}'}(x)| \leqq a_n(\delta_n)$$

$$(x\in(J_k\cup J_{k+1})\cap[0,1];\ k=1,\ldots,s).$$

Now we have to change the $\bar{R}_n^{\{k\}}(x)$'s so that they coincide with $f(x)$ at the endpoints of I_k. For this purpose let

(3. 7) $$R_n^{\{k\}}(x) = \bar{R}_n^{\{k\}}(x)+$$

$$+\frac{f(\xi_k)-\bar{R}_n^{\{k\}}(\xi_k)+\bar{R}_n^{\{k\}}(\xi_{k+1})-f(\xi_{k+1})}{\xi_k-\xi_{k+1}}(x-\xi_{k+1})+f(\xi_{k+1})-\bar{R}_n^{\{k\}}(\xi_{k+1}),$$

which is a rational function of degree $2n+1$. Hence

(3. 8) $$R_n^{\{k\}}(\xi_k) = f(\xi_k), \quad R_n^{\{k\}}(\xi_{k+1}) = f(\xi_{k+1})$$

and taking into account (3. 4), (3. 7), (3. 5) and (3. 3) we obtain

$$(3.9)\qquad |f(x)-R_n^{\{k\}}(x)| \leqq |f(x)-\bar{R}_n^{\{k\}}(x)|+|\bar{R}_n^{\{k\}}(x)-R_n^{\{k\}}(x)| \leqq$$

$$\leqq \varepsilon_n+M^3\eta_n^2+\varepsilon_n+M^3\eta_n^2 = 2\varepsilon_n+2M^3\eta_n^2 \qquad (x\in I_k;\ k=1,\ldots,s),$$

$$(3.10)\quad |R_n^{\{k\}}(x)| \leqq \frac{1}{2\eta_n}+(\varepsilon_n+M^3\eta_n^2)\left[2\max_{1\leqq k\leqq s}\frac{|x-\xi_k|}{\xi_{k+1}-\xi_k}+1\right] \leqq \frac{1}{\eta_n} \quad (x\in[0,1],\ n\geqq n_2),$$

$$(3.11)\qquad |R_n^{\{k\}'}(x)| \leqq a_n(\delta_n)+\frac{2(\varepsilon_n+M^3\eta_n^2)}{\min\limits_{1\leqq k\leqq s}(\xi_{k+1}-\xi_k)} \leqq a_n(\delta_n)+\frac{s}{2\delta_n\eta_n}$$

$$(x\in(J_k\cup J_{k+1})\cap[0,1],\ n\geqq n_3).$$

Now let

$$(3.12)\qquad f_n(x)=\frac{R_n^{\{1\}}(x)+R_n^{\{s\}}(x)}{2}+\frac{1}{2}\sum_{k=2}^{s}|x-\xi_k|\,\frac{R_n^{\{k\}}(x)-R_n^{\{k-1\}}(x)}{x-\xi_k}.$$

Evidently

$$f_n(x)=R_n^{\{k\}}(x) \qquad (x\in I_k,\ k=1,2,\ldots,s),$$

thus we get by (3. 8)

$$(3.13)\qquad |f(x)-f_n(x)|\leqq 2\varepsilon_n+2M^3\eta_n^2 \qquad (x\in[0,1]).$$

Further let $F_m(x-\xi_k)$ be the NEWMAN's rational function of degree $\leqq m$ then (see [2])

$$(3.14)\qquad ||x-\xi_k|-F_m(x-\xi_k)|\leqq 3e^{-\sqrt{m}} \qquad (x\in[\xi_k-1,\ \xi_k+1]).$$

Consider the rational function

$$(3.15)\qquad r_N(x)=\frac{R_n^{\{1\}}(x)+R_n^{\{s\}}(x)}{2}+\frac{1}{2}\sum_{k=2}^{s}F_m(x-\xi_k)\,\frac{R_n^{\{k\}}(x)-R_n^{\{k-1\}}(x)}{x-\xi_k}$$

which is of degree not greater than (see (3. 2) and (1. 4))

$$2(2n+1)+(s-1)(m+4n+1) \leqq s(m+5n) \leqq$$

$$\leqq s\left(5n+\frac{9}{4}\left\{\log\left(3a_n(\delta_n)+\frac{3M}{2}\left(\frac{9s}{\delta_n}\right)^{2/3}\right)+\sqrt{n}\right\}^2\right) \leqq N.$$

From (3. 13), (3. 12), (3. 15) and (3. 14) we get

$$|f(x)-r_N(x)| \leqq |f(x)-f_n(x)|+|f_n(x)-r_N(x)| \leqq$$

$$\leqq 2\varepsilon_n+2M^3\eta_n^2+\frac{3e^{-\sqrt{m}}}{2}\sum_{k=2}^{s}\frac{|R_n^{\{k\}}(x)-R_n^{\{k-1\}}(x)|}{|x-\xi_k|}.$$

First of all let $x\notin J_k$ $(k=1,\ldots,s)$. Then $|x-\xi_k|>\delta_n$, thus from (3. 10)

$$\sum_{k=2}^{s}\frac{|R_n^{\{k\}}(x)-R_n^{\{k-1\}}(x)|}{|x-\xi_k|} \leqq 2\,\frac{s-1}{\delta_n\eta_n} \qquad (x\notin J_k;\ k=1,\ldots,s).$$

Should it happen that $x \in J_l$ then by (3. 10), (1. 2) and (3. 11)

$$\sum_{k=2}^{s} \frac{|R_n^{\{k\}}(x) - R_n^{\{k-1\}}(x)|}{|x-\xi_k|} \leqq 2\,\frac{s-2}{\delta_n \eta_n} + \left| \frac{R_n^{\{l\}}(x) - R_n^{\{l\}}(\xi_l)}{x-\xi_l} \right| + \left| \frac{R_n^{\{l-1\}}(\xi_l) - R_n^{\{l-1\}}(x)}{\xi_l - x} \right| \leqq$$

$$\leqq 2\,\frac{s-2}{\delta_n \eta_n} + 2\left[a_n(\delta_n) + \frac{s}{2\delta_n \eta_n} \right] < \frac{3s}{\delta_n \eta_n} + 2a_n(\delta_n).$$

Hence we have for all $x \in [0, 1]$ by (3. 1) and (3. 2)

$$|f(x) - r_N(x)| \leqq 2\varepsilon_n + 2M^3 \eta_n^2 + \frac{3e^{-\sqrt{m}}}{2}\left(\frac{3s}{\delta_n \eta_n} + 2a_n(\delta_n) \right) =$$

$$= 2\varepsilon_n + \frac{M}{2}\left(\frac{9s}{\delta_n e^{\sqrt{m}}} \right)^{2/3} + M\left(\frac{9s}{\delta_n e^{\sqrt{m}}} \right)^{2/3} + \frac{3a_n(\delta_n)}{e^{\sqrt{m}}} \leqq$$

$$\leqq 2\varepsilon_n + \left[\frac{3M}{2}\left(\frac{9s}{\delta_n} \right)^{2/3} + 3a_n(\delta_n) \right] e^{-\frac{2}{3}\sqrt{m}} = 2\varepsilon_n + e^{-\sqrt{n}} \qquad (n \geqq \max(n_1, n_2, n_3) = n_0)$$

which proves Theorem 1A.

4. PROOF OF THEOREM 2A. First we prove a lemma concerning the NEWMAN's rational function.

LEMMA. *Let*

$$G_n(x) = \frac{F_n(x)}{x} = \frac{p(x) - p(-x)}{p(x) + p(-x)}$$

where

$$p(x) = \prod_{k=0}^{n-1} (x + \xi^k), \quad \xi = e^{-\frac{1}{\sqrt{n}}}.$$

Then

$$|G_n(x)| \leqq 5 \qquad \left(|x| \leqq 1 + \frac{1}{n} \right)$$

and

$$|G_n'(x)| = O\left(n^{5/2} e^{6\sqrt{n}}\right) \qquad \left(|x| \leqq 1 + \frac{1}{n} \right).$$

It is sufficient to prove the statements for $0 \leqq x \leqq 1 + 1/n$. We will make use of the inequalities

$$1 + x \leqq e^x, \quad 1 + 2x \geqq e^x \qquad (0 \leqq x \leqq 1),$$

$$\sqrt{n} \leqq \frac{1}{1-\xi} \leqq 2\sqrt{n},$$

and the Markov's inequality to estimate $|p'(x)|$ and $|p'(-x)|$.

Case 1: $0 \leqq x \leqq \xi^n = e^{-\sqrt{n}}$. Then

$$0 \leqq p(-x) \leqq p(x) = \xi^{\frac{n(n-1)}{2}} \prod_{k=0}^{n-1} \left(1+\frac{x}{\xi^k}\right) \leqq e^{-\frac{1}{2}\sqrt{n}(n-1)} e^{\frac{x}{\xi^{n-1}} \frac{1-\xi^n}{1-\xi}} \leqq e^{-\frac{1}{2}\sqrt{n}(n-1)+2\sqrt{n}},$$

$$p(x) \geqq e^{-\frac{1}{2}\sqrt{n}(n-1)}, \quad |p'(x)| \leqq 2n^2 e^{-\frac{1}{2}\sqrt{n}(n-1)+3\sqrt{n}}, \quad |p'(-x)| \leqq 2n^2 e^{-\frac{1}{2}\sqrt{n}(n-1)+2\sqrt{n}}.$$

Thus

$$0 \leqq G_n(x) \leqq 1,$$

$$|G_n'(x)| \leqq 2\frac{|p'(x)p(-x)|+|p'(-x)p(x)|}{[p(x)+p(-x)]^2} \leqq 2\frac{|p'(x)|+|p'(-x)|}{p(x)} \leqq$$

$$\leqq \frac{8n^2 e^{-\frac{1}{2}\sqrt{n}(n-1)+3\sqrt{n}}}{e^{-\frac{1}{2}\sqrt{n}(n-1)}} = 8n^2 e^{3\sqrt{n}}.$$

Case 2: $\xi^{k+1} \leqq x \leqq \xi^k$ $(k=0, 1, \ldots, n-1)$. Then

$$0 < p(x) = x^{n-k-1} \prod_{j=k+1}^{n-1} \left(1+\frac{\xi^j}{x}\right) \xi^{\frac{k(k+1)}{2}} \prod_{j=0}^{k} \left(1+\frac{x}{\xi^j}\right) \leqq$$

$$\leqq x^{n-k-1} \xi^{\frac{k(k+1)}{2}} e^{u(x)} \leqq \xi^{\frac{k(2n-k-1)}{2}} e^{4\sqrt{n}}$$

where

$$u(x) = \frac{\xi^{k+1}}{x} \cdot \frac{1-\xi^{n-k-1}}{1-\xi} + \frac{x}{\xi^k} \frac{1-\xi^{k+1}}{1-\xi}, \quad \frac{1}{3} < u(x) \leqq 4\sqrt{n}.$$

Analogously

$$p(x) \geqq x^{n-k-1} \xi^{\frac{k(k+1)}{2}} e^{\frac{1}{2}u(x)} > x^{n-k-1} \xi^{\frac{k(k+1)}{2}} e^{\frac{1}{6}},$$

$$|p(-x)| \leqq x^{n-k-1} \xi^{\frac{k(k+1)}{2}} e^{-u(x)} \leqq \xi^{\frac{k(2n-k-1)}{2}} e^{-\frac{1}{3}} \leqq \xi^{\frac{k(2n-k-1)}{2}},$$

$$|p'(x)| \leqq \frac{2n^2 \xi^{\frac{k(2n-k-1)}{2}} e^{4\sqrt{n}}}{\xi^k - \xi^{k+1}} \leqq 4n^{\frac{5}{2}} \xi^{\frac{k(2n-k-3)}{2}} e^{4\sqrt{n}},$$

$$|p'(-x)| \leqq 4n^{\frac{5}{2}} \xi^{\frac{k(2n-k-3)}{2}}.$$

Thus

$$|G_n(x)| \leqq 1 + \frac{2x^{n-k-1} \xi^{\frac{k(k+1)}{2}} e^{-\frac{1}{3}}}{x^{n-k-1} \xi^{\frac{k(k+1)}{2}} \left(e^{\frac{1}{6}} - e^{-\frac{1}{3}}\right)} = 1 + \frac{2}{\sqrt{e}-1} < 5,$$

$$|G_n'(x)| \leqq 2\frac{2 \cdot 4n^{\frac{5}{2}} \xi^{k(2n-k-2)} e^{4\sqrt{n}}}{x^{2n-2k-2} \xi^{k(k+1)} \left(e^{\frac{1}{6}} - e^{-\frac{1}{3}}\right)^2} = O\left(\xi^{k+2-2n} n^{\frac{5}{2}} e^{4\sqrt{n}}\right) \leqq O\left(n^{\frac{5}{2}} e^{6\sqrt{n}}\right).$$

Case 3: $1 \leqq x \leqq 1+\frac{1}{n}$. Then

$$0 < p(x) = x^n \prod_{j=0}^{n-1}\left(1+\frac{\xi^j}{x}\right) \leqq x^n e^{\frac{1}{x}\frac{1-\xi^n}{1-\xi}} \leqq e \cdot e^{\frac{2\sqrt{n}}{x}} \leqq e^{2\sqrt{n}+1}, \quad p(x) \geqq e^{\frac{1}{2}\sqrt{n}},$$

$$|p(-x)| \leqq e e^{-\frac{\sqrt{n}}{2x}} \leqq e^{-\frac{1}{4}\sqrt{n}+1}, \quad |p'(x)| \leqq 2n^3 e^{2\sqrt{n}+1}, \quad |p'(-x)| \leqq 2n^3 e^{-\frac{1}{4}\sqrt{n}+1}.$$

Thus

$$|G_n(x)| \leqq 1 + \frac{2e^{-\frac{1}{4}\sqrt{n}+1}}{e^{\frac{1}{2}\sqrt{n}} - e^{-\frac{1}{4}\sqrt{n}+1}} = 1 + \frac{2e}{e^{\frac{3}{4}\sqrt{n}} - e} < 5,$$

$$|G_n'(x)| \leqq 2\,\frac{2\cdot 2n^3\, e^{\frac{7}{4}\sqrt{n}+2}}{e\sqrt{n}\left(1-e^{-\frac{3}{4}\sqrt{n}+1}\right)^2} = O\left(n^3 e^{\frac{3}{4}\sqrt{n}}\right)$$

and the lemma is proved.

Now we turn to the proof of Theorem 2A. Let us apply Theorem 2 in the intervals $I_k=[\xi_k, \xi_{k+1}]$ $(k=1, \ldots, s)$ instead of $[0, 1]$. There exist rational functions

$$r_n^{\{k\}}(x) = \sum_{j=0}^{r-1} \frac{f^{(j)}(\xi_k)}{j!}(x-\xi_k)^j + \frac{f^{[r]}(\xi_k)}{r!}(x-\xi_k)^r + \tag{4. 1}$$

$$+ \frac{q_{v,k}^{\{1\}}(x)+q_{v,k}^{\{t\}}(x)}{2} + \frac{1}{2}\sum_{l=2}^{t} G_v(x-\zeta_{l,k})\left[q_{v,k}^{\{l\}}(x)-q_{v,k}^{\{l-1\}}(x)\right]$$

of degree n, for which

$$\left|f(x)-r_n^{\{k\}}(x)\right| = V(f^{[r]})\cdot O\left(\frac{\log^2 n}{n^{r+1}}\right) \qquad (x\in I_k,\ k=1, \ldots, s). \tag{4. 2}$$

Here the $q_{v,k}^{\{l\}}(x)$'s are suitable rational functions of degree v having the properties (cf. [1])

$$|q_{v,k}^{\{l\}}(x)| = O(n^{r+5})\,V(f^{[r]}) \qquad (x\in[-2,2]), \tag{4. 3}$$

$$|q_{v,k}^{\{l\}'}(x)| = O(n^{3r+16})\,V(f^{[r]}) \quad (x\in[-2,2]), \tag{4. 4}$$

furthermore

$$v = 10^4 r^2[\log^2 n], \quad t = \left[10^{-6} r^{-2}\frac{n}{\log^2 n}\right] \tag{4. 5}$$

and

$$\xi_k = \zeta_{0,k} < \zeta_{1,k} < \ldots < \zeta_{t,k} = \xi_{k+1}.$$

Now we have by (4. 1), (4. 3), (4. 4), (4. 5) and the lemma

$$|r_n^{\{k\}'}(x)| = O\left[n^{r+5} + \frac{n}{\log^2 n}(\log^5 n \cdot n^{600r}\cdot n^{r+5} + n^{3r+16})\right] =$$

$$= O(\log^3 n \cdot n^{601r+6}) \qquad \left(|x| \leqq 1+\frac{1}{n}\right).$$

Thus we can apply Theorem 1A with (see (4. 2))

$$\varepsilon_n = O\left(\frac{\log^2 n}{n^{r+1}}\right)V(f^{[r]}), \quad a_n(\delta_n) = O(\log^2 n \cdot n^{601r+6}), \quad \delta_n = \frac{1}{n}.$$

The condition (1. 4) gives

$$5n + \frac{9}{4}(\sqrt{n}+\sqrt{n})^2 \leqq \frac{N}{s}$$

i. e.

$$n \leqq \frac{N}{14s}.$$

Correspondingly, there exists a rational function $r_N(x)$ of degree N for which

$$|f(x)-r_N(x)| = O\left(\frac{\log^2 \frac{N}{14s}}{\left(\frac{N}{14s}\right)^{r+1}}\right)V(f^{[r]}) + e^{-\sqrt{\frac{N}{14s}}} = s^{r+1}O\left(\frac{\log^2 N}{N^{r+1}}\right)V(f^{[r]}) + e^{-\sqrt{\frac{N}{14s}}}.$$

REMARK. It would have been possible to prove Theorem 2A without using Theorem 1A, but this way seems to be simpler.

REFERENCES

[1] FREUD, G.: Über die Approximation reeller Funktionen durch rationale gebrochene Funktionen, *Acta Math. Acad. Sci. Hung.* **17** (1966) 313—324.

[2] NEWMAN, D. J.: Rational approximation to $|x|$, *Michigan Math. Journal* **11** (1964) 11—14.

EÖTVÖS LORÁND UNIVERSITY, BUDAPEST

(Received May 31, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 81—89.

A NOTE ON THE SOLUTION OF INTEGRAL EQUATIONS OF CONVOLUTION TYPE OF THE THIRD KIND BY APPLICATION OF THE OPERATIONAL CALCULUS OF MIKUSIŃSKI

by

T. FÉNYES

Introduction

In paper [1] we have solved the integral equation

$$(t+a)f(t)+\int_0^t f(\tau)g(t-\tau)\,d\tau = h(t) \tag{1}$$

by application of the operational calculus of Mikusiński. The functions occurring in (1) are defined and locally integrable over the interval $\langle 0, \infty)$, a is an arbitrary real number.

The equation (1) has the following operational form

$$Df-af-fg = -h \tag{2}$$

in which D denotes the symbol of the so-called algebraic derivative. We have proved that (2) is solvable in the operator field, if the quantity

$$g(+0) = \lim_{t\to+0} g(t)$$

exists and if the function

$$\frac{g(t)-g(+0)}{t}$$

is also locally integrable.

In this case every solution of the algebraic differential equation (2) is a finite-order derivative of a continuous function in the generalized operational sense and can be identified with a Schwartz distribution having a support bounded on the left.

In (1) we have given an explicit general operational solution of (2)

We have also considered the problem of existence of the locally integrable solutions of the algebraic differential equation (2). We have proved that a homogenous equation ($h\equiv 0$) has non-trivial locally integrable solutions, if and only if

$$a\leqq 0 \quad \text{and} \quad g(+0)<0. \tag{3}$$

We have also discussed the problem of existence of the locally integrable solutions in the inhomogenous case $h\neq 0$ assuming a to be non-negative.

The above discussions are based on the theorem of the algebraic integrability of operators. This theorem has been proved by E. GESZTELYI [2] and states the following fact.

Let

$$x = s^k e^{-as} f$$

where f denotes a locally integrable function defined on $\langle 0, \infty)$, a a real number, k a non-negative integer, s the differential-operator, then x is algebraic integrable and

$$\int x\,ds = s^{k+2}e^{-as}u(t),$$

$$u(t) = \begin{cases} -(t+a)^{k+1}\displaystyle\int_0^t \frac{F(\tau)}{(\tau+a)^{k+2}}\,d\tau, & \text{if} \quad a \neq 0 \\ -t^{k+1}\displaystyle\int_\Omega^t \frac{F(\tau)\,d\tau}{\tau^{k+2}}, & \text{if} \quad a = 0,\ \Omega > 0 \text{ arbitrary} \end{cases} \tag{4}$$

$$F(t) = \int_0^t f(\tau)\,d\tau,$$

The determination by applying (4) of the locally integrable solutions of the inhomogeneous integral equation (1) is rather complicated and the formulae expressing the results are very involved [see 1. 4. §.].

In this paper we shall apply a simple generalization of Gesztelyi's theorem, the application of which will be very convenient for the determination of the locally integrable solutions of (1). The formulae will become much more simple.

This is the only purpose of this Note.

1. The Generalization of Gesztelyi's Theorem

THEOREM 1. *Let*

$$x = s^k e^{-as} f,$$

where f is a locally integrable function defined on the interval $\langle 0, \infty)$, $a \geqq 0$, k an arbitrary real number, then

$$\int x\,ds = s^{k+2}e^{-as}u(t)$$

where

$$u(t) = \begin{cases} -(t+a)^{k+1}\displaystyle\int_0^t \frac{F(\tau)\,d\tau}{(\tau+a)^{k+2}}, & \text{if} \quad a > 0 \\ -t^{k+1}\displaystyle\int_0^t \frac{F(\tau)\,d\tau}{\tau^{k+2}}, & \text{if} \quad a = 0 \text{ and } k < -1 \\ -t^{k+1}\displaystyle\int_\varepsilon^t \frac{F(\tau)\,d\tau}{\tau^{k+2}}, & \text{if} \quad a = 0 \text{ and } k \geqq -1 \end{cases}$$

in which the number $\varepsilon > 0$ *is arbitrary, and*

$$F(t) = \int_0^t f(\tau)\, d\tau,$$

$u(t)$ *is continuous on* $\langle 0, \infty)$ *except the case* $a = 0$, $k = -1$, *where, in general, it is only locally integrable.*

The proof of this theorem is wholly analogous to that which has been published in the paper [2] of E. GESZTELYI and so we omit it. Nevertheless, we notice, that Theorem 1 refers only to the cases where $a \geqq 0$ and loses its meaning if $a < 0$.

2. The Locally Integrable Solutions of the Inhomogeneous Integral Equation

We have shown [1] that, if $g(+0)$ exists and if $\dfrac{g(t)-g(+0)}{t}$ is locally integrable, a particular solution of (2) can be written as

$$f = \left[-\int h e^{-as} s^{-g(+0)} (1 + \{G_0\})\, ds\right] e^{as} s^{g(+0)} (1 + \{G\}) \tag{2. 1}$$

where

$$G(t) = \sum_{i=1}^{\infty} \left\{\frac{g(t)-g(+0)}{-t}\right\}^i \frac{1}{i!},$$
$$G_0(t) = \sum_{i=1}^{\infty} \left\{\frac{g(t)-g(+0)}{t}\right\}^i \frac{1}{i!}. \tag{2. 2}$$

The powers are to be understood in the operational sense. In the sequel we shall deal with the problem of the local integrability of (2. 1) by assuming a to be non-negative.

First let $a > 0$. Applying Theorem 1 we get for the algebraic integral occurring in (2. 1)

$$-\int h e^{-as} s^{-g(+0)} (1 + \{G_0\})\, ds = \tag{2. 3}$$

$$= s^{-g(+0)+2} e^{-as} \left\{(t+a)^{-g(+0)+1} \int_0^t \frac{lh + lh * G_0}{(\tau + a)^{-g(+0)+2}}\, d\tau\right\}.$$

Substituting this in (2. 1) we get

$$f = (1 + \{G\}) s^2 \left\{(t+a)^{-g(+0)+1} \int_0^t \frac{lh + lh * G_0}{(\tau + a)^{-g(+0)+2}}\, d\tau\right\}.$$

It can be easily seen that the multiplication of the operator s^2 and of the function occurring in the second bracket of the above formula corresponds to two times

differentiation of this function. So we obtain the function:

$$\frac{h+h*G_0}{t+a}-g(+0)\frac{lh+lh*G_0}{(t+a)^2}-$$

$$-g(+0)(1-g(+0))(t+a)^{-g(+0)-1}\int_0^t\frac{lh+lh*G_0}{(\tau+a)^{2-g(+0)}}\,d\tau.$$

Integration by parts gives $(g(+0)\neq 0, 1)$

$$\int_0^t\frac{lh+lh*G_0}{(\tau+a)^{2-g(+0)}}\,d\tau=\frac{lh+lh*G_0}{(g(0)-1)(t+a)^{1-g(+0)}}+\frac{1}{1-g(+0)}\int_0^t\frac{h+h*G_0}{(\tau+a)^{1-g(0)}}\,d\tau.$$

By carrying out every substitutions, we get the wanted locally integrable solution which can be written in the form

$$(2.4)\quad f=(1+\{G\})\left\{\frac{h+h*G_0}{t+a}-g(+0)(t+a)^{-g(+0)-1}\int_0^t\frac{h+h*G_0}{(\tau+a)^{1-g(+0)}}\,d\tau\right\}.$$

However, it can be easily seen that (2. 4) holds for every values of $g(0)$. In [1] we have shown that in the case $a>0$ the homogeneous integral equation (1) has no locally integrable solutions except the trivial solution. Referring to this we have the following

THEOREM 2. *Let* $g(+0)$ *exist and let the functions* $\dfrac{g(t)-g(+0)}{t}$ *and* $h(t)$ *be locally integrable. Then the integral equation*

$$(t+a)f(t)+\int_0^t f(\tau)g(t-\tau)\,d\tau=h(t),\qquad (a>0)$$

has only one locally integrable solution which can be written in the following form

$$f(t)=U(t)+U(t)*G(t)$$

where

$$G(t)=\sum_{i=1}^{\infty}\left\{\frac{g(t)-g(+0)}{-t}\right\}^i\frac{1}{i!}$$

$$U(t)=\frac{h+h*G_0}{t+a}-g(+0)(t+a)^{-g(+0)-1}\int_0^t\frac{h+h*G_0}{(\tau+a)^{-g(+0)+1}}\,d\tau$$

$$G_0(t)=\sum_{i=1}^{\infty}\left\{\frac{g(t)-g(+0)}{t}\right\}^i\frac{1}{i!}.$$

Now we will discuss the local integrable solutions in the case $a=0$. The local integrability of $h(t)$, $\frac{g(t)-g(+0)}{t}$ does not guarantee the existence of locally integrable solutions of the equation

$$tf(t)+f*g=h.$$

As an example, let us consider the integral equation

$$tf(t)+\int_0^t f(\tau)\,d\tau = \{1\}$$

the operational solutions of which can be written in the form

$$f=Cs+1 \qquad (C \text{ is an arbitrary number}).$$

Consequently there are no locally integrable solutions. We shall give later a simple sufficient condition concerning $h(t)$ which admits the existence of at least one locally integrable solution. We make use of the following three statements which have been proved in paper [1].

1. Let the functions $a(t)$, $\frac{a(t)}{t}$, $b(t)$ be locally integrable on the interval $\langle 0, \infty)$, then the function

$$\frac{a(t)*b(t)}{t}$$

is also locally integrable on this interval.

2. Let $\frac{h(t)}{t}$ be locally integrable on $\langle 0, \infty)$, then the function $\frac{lh(t)}{t}$ is absolutely continuous and tends to zero as t tends to zero. Moreover, the function $\frac{lh(t)}{t^2}$ is also locally integrable on $\langle 0, \infty)$. Here l denotes the integral operator.

3. Let $F(t)$ be a locally integrable function on $\langle 0, \infty)$, $k\neq 0$ a real number, $\Omega \geqq 0$. The functions

$$t^{k-1}\int_0^t \frac{F(\tau)\,d\tau}{\tau^k}, \qquad k<0$$

$$t^{k-1}\int_\Omega^t \frac{F(\tau)\,d\tau}{\tau^k}, \qquad k>0$$

are locally integrable on $\langle 0, \infty)$. (Of course, the choice $\Omega=0$ is only permissible if the second integral remains convergent).

Let us rewrite again (2. 1) substituting $a=0$

$$f=-\Big[\int hs^{-g(+0)}\big(1+\{G_0\}\big)\,ds\Big]s^{g(+0)}\big(1+\{G\}\big). \tag{2. 5}$$

By Theorem 1 we get

$$-\int hs^{-g(+0)}(1+\{G_0\})\,ds =$$

$$= \begin{cases} s^{-g(+0)+2}\left\{t^{1-g(+0)}\displaystyle\int_0^t \frac{lh+lh*G_0}{\tau^{2-g(+0)}}\,d\tau\right\}, & \text{if} \quad g(+0) > 1 \\ s^{-g(+0)+2}\left\{t^{1-g(+0)}\displaystyle\int_\varepsilon^t \frac{lh+lh*G_0}{\tau^{2-g(+0)}}\,d\tau\right\}, & \text{if} \quad g(+0) \leqq 1 \end{cases}$$

and

$$f = (1+\{G\})s^2\left\{t^{-g(+0)+1}\int_0^t \frac{lh+lh*G_0}{\tau^{2-g(+0)}}\,d\tau\right\}, \quad \text{if} \quad g(+0) > 1$$

(2. 6)

$$f = (1+\{G\})s^2\left\{t^{-g(+0)+1}\int_\varepsilon^t \frac{lh+lh*G_0}{\tau^{2-g(+0)}}\,d\tau\right\}, \quad \text{if} \quad g(+0) \leqq 1.$$

The operator (2. 6) is not a function in general. Nevertheless, it is evident from the elements of the operational calculus that (2. 6) is a function, if and only if the function occurring in the second bracket of the corresponding formula in (2. 6) has a locally integrable second derivative and if this function itself with its first derivative tends to zero as t tends to zero. If these conditions are satisfied the multiplication of the operator s^2 and this function is equivalent to two times differentiation of it.

We show that if $\dfrac{h(t)}{t}$ is locally integrable then the operator (2. 6) is a function.

In the sequel we require the local integrability of $\dfrac{h(t)}{t}$.

Then it may be chosen the value $\varepsilon = 0$ in the second equation of (2. 6) provided that $g(+0) \geqq 0$, because the integral remains convergent. This fact can be easily seen by application of the above statements 1 and 2. So we can write (2. 6) in the more convenient form

$$f = (1+\{G\})s^2\left\{t^{-g(+0)+1}\int_0^t \frac{lh+lh*G_0}{\tau^{-g(+0)+2}}\,d\tau\right\}, \quad \text{if} \quad g(+0) \geqq 0,$$

(2. 7)

$$f = (1+\{G\})s^2\left\{t^{-g(+0)+1}\int_\varepsilon^t \frac{lh+lh*G_0}{\tau^{-g(+0)+2}}\,d\tau\right\}, \quad \text{if} \quad g(+0) < 0.$$

When $g(+0) \geqq 0$ the function occurring in the bracket after s^2 in the first formula in (2. 7) is absolutely continuous and tends to zero as $t \to 0$. Its derivative

$$\frac{lh+lh*G_0}{t} + (1-g(+0))t^{-g(+0)}\int_0^t \frac{lh+lh*G_0}{\tau^{2-g(+0)}}\,d\tau \tag{2. 8}$$

also tends to zero because of the required condition on the function $\frac{h(t)}{t}$. This can be seen by application of the statements 1 and 2 and of the Bernouilli—L'Hospital rule. Moreover (2. 8) is also absolutely continuous, its derivative is

$$\frac{h+h*G_0}{t}-g(+0)\frac{lh+lh*G_0}{t^2}-g(+0)(1-g(+0))t^{-g(+0)-1}\int_0^t\frac{lh+lh*G_0}{\tau^{-g(+0)+2}}\,d\tau.$$

Wholly analogously to the case $a>0$, this expression can be reduced to the following form

$$\frac{h+h*G_0}{t}-g(+0)t^{-g(+0)-1}\int_0^t\frac{h+h*G_0}{\tau^{-g(+0)+1}}\,d\tau. \tag{2. 9}$$

This is a locally integrable function (see the statements 1 and 3). By (2. 9) and the first formula of (2. 7) we get the required particular solution as follows:

$$f=(1+\{G\})\left\{\frac{h+h*G_0}{t}-g(+0)\,t^{-g(+0)-1}\int_0^t\frac{h+h*G_0}{\tau^{1-g(+0)}}\,d\tau\right\}. \tag{2. 10}$$

When $g(+0)<0$ we start from the second formula of (2. 7). Analogously to (2. 9) we get the function

$$\frac{h+h*G_0}{t}+\gamma g(+0)(1-g(+0))t^{-g(+0)-1}-g(+0)t^{-g(+0)-1}\int_\varepsilon^t\frac{h+h*G_0}{\tau^{-g(+0)+1}}\,d\tau \tag{2. 11}$$

where the constant γ is determined by the constant $\varepsilon>0$. The required particular solution can be written as follows:

$$f=(1+\{G\})\left\{\frac{h+h*G_0}{t}-\gamma g(+0)(1-g(+0))t^{-g(0)-1}-\right.$$
$$\left.-g(+0)\,t^{-g(+0)-1}\int_\varepsilon^t\frac{h+h*G_0}{\tau^{1-g(+0)}}\,d\tau\right\}. \tag{2. 12}$$

Comparing (2. 10) with (2. 12) we see that these functions differ from each other in the function

$$C(1+\{G\})\{t^{-g(+0)-1}\}\qquad(C=\text{constant}). \tag{2. 13}$$

However, we have proved in paper [1] that (2. 13) is the general solution of the homogeneous integral equation

$$tf+f*g=0.$$

So we get the following

THEOREM 2. *Let the value* $g(+0)$ *exist and let the functions* $\frac{g(t)-g(+0)}{t}$ *and* $\frac{h(t)}{t}$ *be locally integrable on* $\langle 0, \infty)$. *When* $g(+0) \geqq 0$ *the integral equation*

$$tf(t)+\int_0^t f(\tau)g(t-\tau)\,d\tau = h(t) \tag{2.14}$$

has exactly one locally integrable solution on $\langle 0, \infty)$ *which can be written as follows:*

$$f(t)=U(t)+U(t)*G(t)$$

where

$$U(t) = \frac{h(t)+h(t)*G_0(t)}{t} - g(+0)t^{-g(+0)-1}\int_0^t \frac{h+h*G_0}{\tau^{1-g(0)}}\,d\tau$$

If $g(+0)<0$ *the integral equation* (2.14) *has an infinite number of locally integrable solutions that has the following form:*

$$f(t)=U(t)+U(t)*G(t)$$

where

$$U(t) = Ct^{-g(+0)-1} + \frac{h(t)+h(t)*G_0(t)}{t} - g(+0)t^{-g(+0)-1}\int_\varepsilon^t \frac{h+h*G_0}{\tau^{-g(+0)+1}}\,d\tau$$

($\varepsilon>0$, C arbitrary).

Any two solutions differ from each other in one solution of the homogeneous integral equation.

REMARK. The local integrability of $\frac{h(t)}{t}$ is only a sufficient condition and is not necessary. We can easily give examples of the discussed integral equation where the function $\frac{h(t)}{t}$ is not locally integrable, although the integral equation

$$tf+f*g=h$$

has a locally integrable solution.

As an example let us assume that

$$f(t) = \begin{cases} \dfrac{1}{t\log^2 t}, & \text{if} \quad 0 \leqq t \leqq \delta < 1 \\ 0 \qquad\quad , & \text{if} \quad t > \delta \end{cases}$$

$$g(t) = 1.$$

Here $f(t)$, $g(t)$ are locally integrable but a simple calculation shows that $\frac{h(t)}{t}$ is not locally integrable on $\langle 0, \infty)$.

REFERENCES

[1] FÉNYES, T.: Anwendung der Mikusińskischen Operatorenrechnung zur Lösung von Integralgleichungen dritter Art von Faltungstypus, *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei* **9** (1965) 365—399.

[2] GESZTELYI, E.: Anwendung der Operatorenrechnung auf lineare Differentialgleichungen mit Polynom-Koeffizienten, *Publicationes Mathematicae* **10** (1963) 215—243.

[3] BUTZER, P. L.: Die Anwendung des Operatorenkalküls von Jan Mikusinski auf lineare Integralgleichungen vom Faltungstypus, *Archive for Rational Mechanics and Analysis* **2** (1958) 114—128.

[4] FÉNYES, T.—KOSIK, P.: Über das algebraische Integral der Mikusińskischen Operatoren, *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei* **9** (1964) 21—34.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received June 7, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 91—114.

FIXPUNKTSÄTZE FÜR ITERATIONEN MIT VERÄNDERLICHEN OPERATOREN

von

T. FREY

1. §. Einleitende Bemerkungen

J. SCHRÖDER hat vor einigen Jahren erkannt, daß man die Güte der benützten Näherungsverfahren bzw. numerischen Methoden viel tiefer charakterisieren kann, wenn die Abstände der betrachteten Funktionen nicht mit Hilfe von Zahlen, d. h. mit gewöhnlicher Metrik, sondern mit Hilfe von Elementen eines halbgeordneten linearen Raumes, d. h. mit einer sog. Pseudometrik angegeben sind (s. z. B. [1], [2]). Er hat nämlich einen Fixpunktsatz für Iterationsverfahren angegeben, mit Hilfe welcher man in vielen Fällen nicht nur Existenz, aber auch Konvergenz bzw. Konvergenzschnelle von Iterationsfolgen gegen die gesuchte Lösung beweisen kann (s. z. B. [3], [4]). Dieser Satz ist jedoch nur dann anwendbar, wenn man „konvexe" Operatoren betrachtet, die daneben beim Verlauf des Prozesses konstant bleiben. In den interessantesten praktischen Fällen sind jedoch die Operatoren oft „konkav" und bleiben daneben oft nicht konstant.

Wir geben nachstehend einige Verallgemeinerungen dieses SCHRÖDERschen Fixpunktsatzes in den benannten Richtungen an, und zeigen mit Hilfe einiger Beispiele die Anwendbarkeit dieser Sätze. Zuerst einige Bezeichnungen und Definitionen:

Wir nennen die Menge $\mathfrak{M}$ einen halbgeordneten, linearen Raum über dem vollgeordneten Ring K, falls

a) $\mathfrak{M}$ halbgeordnet ist, d. h. eine Ordnungsrelation (bezeichnet durch $\leqq$) zwischen gewissen Elementen von $\mathfrak{M}$ besteht, mit den Eigenschaften:

1°, $h \leqq h$ für alle $h \in \mathfrak{M}$;

2°, $\leqq$ ist transitiv, d. h. aus $f \leqq g$; $g \leqq h$ folgt ja $f \leqq h$;

3°, aus $f \leqq g$ und $g \leqq f$ folgt ja $f = g$;

b) Linearität und Halbordnung von $\mathfrak{M}$ sind einander kompatibel, d. h.

4°, aus $f \geqq \Theta$; $f \in \mathfrak{M}$; $c \geqq 0$, $c \in K$ folgt ja $cf \geqq \Theta$ wobei Θ das Nullelement der Abelschen Gruppe $\mathfrak{M}$, 0 aber dasjenige des Ringes K bezeichnet;

5°, aus $f_j \leqq g_j$ $(j = 1, 2;\ f_j, g_j \in \mathfrak{M})$ folgt ja $f_1 + f_2 \leqq g_1 + g_2$;

c) Die Topologie von $\mathfrak{M}$ ist kompatibel mit der Halbordnung und mit der Linearität von $\mathfrak{M}$, d. h. es gibt einen abstrakten Limesbegriff über $\mathfrak{M}$ bzw. K (bezeichnet durch $\lim_{n \to \infty} e_n = e$) mit den Eigenschaften:

6°, eine jede konstante Folge ist konvergent, d. h. ist $e_n \equiv e$ $(n = 1, 2, \ldots)$, so ist $\lim_{n \to \infty} e_n = e$ $(e_n, e \in \mathfrak{M}$ bzw $\in K)$.

7°, jede Teilfolge einer konvergenten Folge ist konvergent und strebt nach demselben Limeselement;

8°, strebt die Folge $f_n \in \mathfrak{M}$ bzw. $g_n \in \mathfrak{M}$ nach f bzw. g, so ist auch $f_n + g_n$ konvergent, und strebt gegen $f+g$;

9°, strebt die Folge $f_n \in \mathfrak{M}$ nach f; $c_n \in K$ aber nach c, so ist auch $\lim_{n\to\infty} c_n f_n = cf$ gültig;

10°, strebt die Folge $f_n \in \mathfrak{M}$ nach f, und ist für alle n $f_n \geqq \Theta$ gültig, so gilt auch $f \geqq \Theta$; und endlich

11°, strebt die Folge $f_n \geqq \Theta$, $f_n \in \mathfrak{M}$ $(n = 1, 2, \ldots)$ nach Θ, und gilt für jedes n $\Theta \leqq g_n \leqq f_n$, so steht auch $\lim_{n\to\infty} g_n = \Theta$ fest.

Wir nennen nun die Menge R einen pseudometrischen Raum, wenn zu jedem Elementenpaar s, t von R ein Element $\varrho(s, t)$ eines linearen, halbgeordneten Raumes $\mathfrak{M}$ zugeordnet ist, mit den Eigenschaften:

12°, $\varrho(s, t) = \Theta_{\mathfrak{M}}$ dann und nur dann, wenn $s = t$ ist;

13°, $\varrho(s, t) \leqq \varrho(s, u) + \varrho(t, u)$ steht für jedes $u \in R$ fest.

Es ist nun leicht einzusehen, daß dieser Pseudoabstand ϱ symmetrisch und ein Element der sog. positiven Halbgruppe von $\mathfrak{M}$ ist, d. h. auch

$$\varrho(s, t) = \varrho(t, s)$$

auch

$$\varrho(s, t) \geqq \Theta_{\mathfrak{M}}$$

gültig ist.

Die Topologie des Raumes R ist somit durch diejenige von $\mathfrak{M}$ definiert: $v_n \in R \to v \in R$ ist nämlich gleichbedeutend mit $\lim_{n\to\infty} \varrho(v_n, v) = \Theta_{\mathfrak{M}}$. Es leuchtet ein, daß dieser Limesbegriff eindeutig definiert ist. Man kann somit auch die Begriffe Vollständigkeit, Kompaktheit, Stetigkeit usf. in R einführen.

Es sei nun R bzw. S ein pseudometrischer Raum, mit Pseudoabständen ϱ bzw. σ des halbgeordneten, linearen Raumes $\mathfrak{M}$ bzw. $\mathfrak{N}$. Wir nennen nun den Operator T, definiert über $r \subseteqq R$, mit eindeutiger Bildmenge in S, einen beschränkten Operator in r, falls man einen stetigen, beschränkten, positiven Operator P, definiert über $\mathfrak{M}$, mit einer Bildmenge in $\mathfrak{N}$ so angeben kann, daß für jedes $u, v \in r$

$$\sigma(Tu, Tv) \leqq P\varrho(u, v)$$

feststeht.

Ist daneben auch $P\Theta_{\mathfrak{M}} = \Theta_{\mathfrak{N}}$ gültig, so ist T auch stetig in r.

Man kann auch die Topologie des Raumes $\mathfrak{N}$ auf die Menge der Operatoren (den Begriff der schwachen Konvergenz), definiert über r, mit Bildmenge in S, übertragen. Wir sagen nämlich, daß die Folge T_n solcher Operatoren $\mathfrak{N}$-schwach nach T strebt, falls für jedes $u \in r$

$$\lim_{n\to\infty} \sigma(T_n u, Tu) = \Theta_{\mathfrak{N}}$$

feststeht.

Es ist zweckmäßig, auch den Begriff der ($\mathfrak{N}$-schwachen) gleichmäßigen Konvergenz von Operatorenfolgen einzuführen. Wir sagen nämlich, daß die Operatorenfolge T_n $\mathfrak{N}$-schwach gleichmäßig nach T über $r \subseteqq R$ strebt, falls man ein Element $w_0 \in R$, ferner eine Folge stetiger, positiver Operatoren P_n, definiert über $\mathfrak{M}$, mit Bildmengen in $\mathfrak{N}$ so angeben kann, daß

I°. die Folge P_n monoton in n ist, d. h. für jedes $\varrho \in \mathfrak{M}$ und jedes Indexpaar $m \geqq l$ — bzw. $m \leqq l$ — eine Relation

$$P_m \varrho \geqq P_l \varrho$$

gilt;

II°. die Folge P_n nach einem stetigen, positiven Operator $P_\infty = P$ strebt, d. h. für jedes $\varrho \in \mathfrak{M}$

$$\lim_{n \to \infty} (P_n \varrho - P_\infty \varrho) = \Theta_{\mathfrak{N}}$$

feststeht;

III°. die Relation

$$\sigma(T_m u,\ T_l u) \leqq P_m \varrho(u,\ w_0) - P_l \varrho(u,\ w_0)$$

für jedes $u \in r$ und für jedes Indexpaar $l \leqq m \leqq \infty$ — bzw. $m \leqq l \leqq \infty$ — gilt.

Man kann einsehen, dass jede gleichmäßig konvergente Operatorenfolge auch konvergent ist.

Den Begriff des Pseudoabstandes (den Begriff der starken Konvergenz) kann man im allgemeinen nicht direkt in den Raum der Operatoren übertragen, da im Raume $\mathfrak{N}$ im allgemeinen keine „obere Grenze" für eine Menge von Elementen existiert. Wie wir aber gleich sehen werden, braucht man diesen Begriff nicht unbedingt.

2. §. Die Fixpunktsätze für pseudometrische Räume

Wir betrachten nachstehend folgende Probleme: der Operator $T = T_\infty$ bildet den Teilraum $r \subseteq R$ des vollständigen, linearen, pseudometrischen Raumes R in R ab; es gibt ferner eine Folge von „numerischen Operatoren" T_n, die r in R abbilden, ferner ($\mathfrak{M}$-schwach) gleichmäßig nach T streben. Gibt man nun eine „approximierende" Folge $u_{n+1} = T_n u_n$ an, so kann man folgende Fragen betrachten:

1°. Strebt die Folge u_n nach einem Fixpunkt $u \in r$ des Operators T_∞?

2°. Ist dieser Fixpunkt eindeutig definiert?

3°. Wie ist die Konvergenzschnelle dieser Folge?

SATZ 1a. *Setzen wir voraus, daß die Operatorenfolge P_n, mit Hilfe welcher wir — und mit $w_0 \in R$ — die gleichmäßige Konvergenz von $\{T_n\}$ gegen $T = T_\infty$ darstellen, monoton nichtfallend nach P_∞ strebt, ferner die folgenden weiteren Bedingungen erfüllt sind:*

α) für jedes Paar von Elementen $u, v \in r$, und für jedes Indexpaar $m \leqq l \leqq \infty$ gilt die „Abschätzung"

$$\varrho(T_l u,\ T_m v) \leqq Q\varrho(u, v) + P_l[\varrho(u, v) + \varrho(v, w_0)] - P_m \varrho(v, w_0)$$

mit $Q\Theta_{\mathfrak{M}} = \Theta_{\mathfrak{M}}$;

β) die Paare der Folge P_n, ferner der Operator Q sind monoton „nichtfallende" und „nichtkonkave" Funktionen ihrer Argumente, d. h. für jede $\Theta \leqq \varrho \leqq \varrho^*$; $\Theta \leqq \sigma \leqq \sigma^*$ und $m \leqq l \leqq \infty$ ist

$$P_l(\varrho + \sigma) - P_m \varrho \leqq P_l(\varrho^* + \sigma^*) - P_m \varrho^*;$$

$$Q(\varrho + \sigma) - Q\varrho \leqq Q(\varrho^* + \sigma^*) - Q\varrho^*;$$

γ) $\tau \in \mathfrak{M}$ *ist ein geeignetes Element, mit dessen Hilfe man die Operatorenfolge* S_n:

$$S_n\sigma = Q\sigma + P_n\sigma + \tau$$

bzw. die Folge $S_n\sigma_n = \sigma_{n+1} \in \mathfrak{M}$ *definiert; ferner ist* $u_0 \in r$ *bzw.* $\sigma_0 \in \mathfrak{M}$ *das Anfangselement der Folge* $T_n u_n$ *bzw.* $S_n\sigma_n$, *zusammen mit* τ *so gewählt, daß*

$$\sigma_0 \geqq \varrho(u_0, w_0)$$

und

$$\sigma_1 = S_0\sigma_0 \geqq \sigma_0 + \varrho(u_0, T_0 u_0)$$

gilt.

δ) *Die Folge* S_n *ist „gleichgradig" vollstetig, d. h. ist die Menge* $M\{m_\alpha; \alpha \in \Gamma\} \subset \mathfrak{M}$ *beschränkt, so ist* $\hat{M}\{S_{n(\alpha)}m_\alpha; \alpha \in \Gamma\}$ *bei beliebiger Wahl der benützten Indexmenge* $n(\alpha)$ *kompakt. Die Folge* σ_n *sei beschränkt.*

ε) *Endlich ist* r *eine vollständige Menge, welche daneben die ganze Folge* $u_{n+1} = T_n u_n$ *enthält.*

Neben den oben angegebenen Bedingungen gibt es für die Gleichung $u = Tu$ *mindestens eine Lösung (d. h. der Operator* T *besitzt einen Fixpunkt)* in r; die Folge u_n *strebt nach einer solchen Lösung.*

BEWEIS. 1°. Wir zeigen zuerst, dass die Folge S_n bzw. σ_n konvergent ist, nach S_∞ bzw. $\sigma_\infty \in \mathfrak{M}$ strebt, und letztere genügt der Gleichung

$$\sigma_\infty = S_\infty\sigma_\infty .$$

Nun $S_n\sigma = Q\sigma + P_n\sigma + \tau$, und da hier P_n nach P_∞ strebt, so strebt auch S_n nach S_∞ mit $S_\infty\sigma = Q\sigma + P_\infty\sigma + \tau$.

Die Folge σ_n ist monoton wachsend: $\sigma_1 \geqq \sigma_0$ folgt nämlich aus γ), und gilt $\sigma_{n+1} \geqq \sigma_n$ für $n = 0, 1, 2, \ldots, N-1$, so ist nach β) bzw. γ)

$$\sigma_{N+1} - \sigma_N = S_N\sigma_N - S_{N+1}\sigma_{N-1} = Q\sigma_N + P_N\sigma_N + \tau -$$

$$-[Q\sigma_{N-1} + P_{N-1}\sigma_{N-1} + \tau] = (Q\sigma_N - Q\sigma_{N-1}) + (P_N\sigma_N - P_{N-1}\sigma_{N-1}) \geqq \Theta_{\mathfrak{M}},$$

d. h.

$$\sigma_{n+1} \geqq \sigma_n \tag{1}$$

auch für $n = N$, also für alle n gültig.

Nach δ) ist daneben die Folge σ_n beschränkt, und so nach δ) auch kompakt, da $\sigma_n = S_{n-1}\sigma_{n-1}$ gültig ist. σ_n hat also eine konvergente Teilfolge, und die Folge selbst strebt nach dem Grenzwert σ_∞ dieser Teilfolge, da sie monoton ist.

Nach β) bzw. γ) ist auch die Folge S_n (mit dem Index) monoton wachsend; es gilt also für jedes $k \geqq m$

$$S_k\sigma_m \geqq S_m\sigma_m = \sigma_{m+1}$$

folglich auch

$$S_\infty\sigma_m \geqq \sigma_{m+1},$$

also auch

$$S_\infty\sigma_\infty \geqq \sigma_\infty, \tag{2}$$

da S_∞ stetig ist.

Jedoch ist nach β) bzw. γ) S_n auch eine monoton wachsende Funktion des Argumentes. Somit für jedes $k \geqq m$

$$\sigma_{m+1} = S_m \sigma_m \leqq S_m \sigma_k \leqq S_m \sigma_\infty$$

also

$$\sigma_\infty \leqq S_\infty \sigma_\infty . \tag{3}$$

(2) und (3) zeigen nun die Gültigkeit unserer Behauptungen.

2°. Wir schätzen nun die Pseudoabstände der Elemente u_k und u_m, bzw. u_k und w_0, d. h.

$$\varrho_{k,m} = \varrho(u_k, u_m) \quad \text{bzw.} \quad \varrho_k = \varrho(u_k, w_0).$$

Nach γ) ist die Abschätzung $\sigma_0 \geqq \varrho_0$ und $\sigma_1 - \sigma_0 \geqq \varrho_{1,0}$ gültig. Wir zeigen nun durch vollständige Induktion, daß auch allgemein die Abschätzungen

$$\sigma_n \geqq \varrho_n \quad \text{und für } k > m \;\; \sigma_k - \sigma_m \geqq \varrho_{k,m} \tag{4}$$

gültig sind. Es sei vorausgesetzt, dass (4) für $n = 0, 1, 2, \ldots, N$ bzw. für $k = 1, 2, \ldots, N$ und für $m = 0, 1, \ldots, k-1$ schon gezeigt wurde. Ist nun $k = N+1$ und $0 < m \leqq N$, so

$$\begin{aligned}
\varrho_{N+1,m} = \varrho(u_{N+1}, u_m) = \varrho(T_N u_N, T_{m-1} u_{m-1}) \leqq \\
\leqq Q \varrho_{N,m-1} + P_N [\varrho_{N,m-1} + \varrho(u_{m-1}, w_0)] - P_{m-1} \varrho(u_{m-1}, w_0) \leqq \\
\leqq Q(\sigma_N - \sigma_{m-1}) + P_N [\sigma_N - \sigma_{m-1} + \sigma_{m-1}] - P_{m-1} \sigma_{m-1} \leqq \\
\leqq Q\sigma_N + P_N \sigma_N + \tau - \{Q\sigma_{m-1} + P_{m-1} \sigma_{m-1} + \tau\} = \\
= S_N \sigma_N - S_{m-1} \sigma_{m-1} = \sigma_{N+1} - \sigma_m ,
\end{aligned}$$

d. h. der zweite Teil von (4) ist für $k = N+1$ und $m = 1, 2, \ldots, N$ gültig. Ferner

$$\begin{aligned}
\varrho_{N+1,0} = \varrho(u_{N+1}, u_0) \leqq \varrho(u_{N+1}, u_1) + \varrho(u_1, u_0) \leqq \\
\leqq \sigma_{N+1} - \sigma_1 + \sigma_1 - \sigma_0 = \sigma_{N+1} - \sigma_0 ,
\end{aligned}$$

d. h. dieser zweite Teil ist auch für $m = 0$ gültig.

Endlich

$$\begin{aligned}
\varrho_{N+1} = \varrho(u_{N+1}, w_0) \leqq \varrho(u_{N+1}, u_0) + \varrho(u_0, w_0) \leqq \\
\leqq \sigma_{N+1} - \sigma_0 + \sigma_0 = \sigma_{N+1} ,
\end{aligned}$$

d. h. auch der erste Teil von (4) ist für N + 1 — und so für alle n bzw. k und $0 \leqq m \leqq$ $\leqq k-1$ — gültig.

3°. (4) hat zur Folge, daß $u_{n+1} = T_n u_n$ eine in sich konvergente Folge ist; nach ε) besitzt sie somit einen Grenzwert $u_\infty \in r$. Wir zeigen nun, daß $u_\infty = T_\infty u_\infty$ gültig ist.

$$\begin{aligned}
\varrho(u_\infty, T_\infty u_\infty) \leqq \varrho(u_{n+1}, T_\infty u_\infty) + \varrho(u_{n+1}, u_\infty) = \\
= \varrho(T_n u_n, T_\infty u_\infty) + \varrho(u_{n+1}, u_\infty) \leqq Q \varrho(u_n, u_\infty) + P_\infty [\varrho(u_\infty, u_n) + \\
+ \varrho(u_n, w_0)] - P_n \varrho(u_n, w_0) + \varrho(u_{n+1}, u_\infty) \leqq \\
\leqq Q[\sigma_\infty - \sigma_n] + P_\infty [\sigma_\infty - \sigma_n + \sigma_n] - P_n \sigma_n + \sigma_\infty - \sigma_{n+1} \leqq \\
\leqq Q\sigma_\infty + P_\infty \sigma_\infty + \tau - \{Q\sigma_n + P_n \sigma_n + \tau\} + \sigma_\infty - \sigma_{n+1} = \\
= S_\infty \sigma_\infty - S_n \sigma_n + \sigma_\infty - \sigma_{n+1} = 2(\sigma_\infty - \sigma_{n+1}) \to \Theta_{\mathfrak{M}}
\end{aligned}$$

falls $n \to \infty$. Also

$$\varrho(u_\infty, T_\infty u_\infty) \leqq \Theta_{\mathfrak{M}},$$

und daraus folgt unsere Behauptung.

Es sei noch bemerkt, daß man in α) Q mit einer monoton wachsenden Folge Q_n ersetzen kann.

SATZ 1b. *Betrachten wir wieder das Problem des Satzes* 1a, *jedoch setzen wir jetzt voraus, daß die Folge* P_n *monoton fallend nach* P_∞ *strebt, ferner die Paare* P_n, P_m *„konkav", und* $Q = 0$ *sind. Genauer:*

αb) *für* $k \leqq m \leqq \infty$

$$\varrho(T_k u, T_m v) \leqq P_k[\varrho(u, v) + \varrho(v, w_0)] - P_m \varrho(v, w_0);$$

βb) *für* $k \leqq m \leqq \infty$, $\Theta \leqq \varrho \leqq \varrho^*$ *und* $\Theta \leqq \sigma \leqq \sigma^*$

$$P_k(\varrho^* + \sigma) - P_m \varrho^* \leqq P_k(\varrho + \sigma^*) - P_m \varrho;$$

γb) $S_n \sigma = P_n \sigma + \tau$, *mit entsprechendem* $\tau \in \mathfrak{M}$; *ferner*

$$\sigma_0 \leqq \varrho(u_0, w_0) \quad \text{und} \quad \sigma_0 \geqq \sigma_1 + \varrho(u_0, T_0 u_0);$$

δb) *und* εb) *enthalten dieselben Voraussetzungen wie* δ) *und* ε) in Satz 1a), *wir setzen jedoch auch noch* $\sigma_n \geqq \Theta_{\mathfrak{M}}$ $(n = 0, 1, 2, \ldots)$ *voraus.*

Unter diesen Bedingungen ist die Behauptung des Satzes 1a) *wieder gültig.*

BEWEIS: Wir werden dem Gedankengang des Beweises des vorigen Satzes folgen und nur die Änderungen bemerken. In 1° kann man jetzt aus der Voraussetzung $\sigma_1 \leqq \sigma_0$, und aus $P_n \searrow P_\infty$ gleich zeigen, dass die Folge σ_n monoton fallend ist:

$$\sigma_N - \sigma_{N+1} = P_{N-1}\sigma_{N-1} - P_N \sigma_N \geqq \Theta, \quad \text{falls} \quad \sigma_{N-1} - \sigma_N \geqq \Theta.$$

Daraus und aus $\sigma_n \geqq \Theta$ folgt nun wieder, dass $\sigma_n \to \sigma_\infty$ feststeht. Da nun weiter auch σ_n, auch S_n bzw. P_n monoton fallend sind, so gilt einerseits für $k > m$

$$S_k \sigma_m \leqq S_m \sigma_m = \sigma_{m+1}$$

d. h.

$$S_\infty \sigma_m \leqq \sigma_{m+1},$$

resp. wegen der Stetigkeit von S_∞

$$S_\infty \sigma_\infty \leqq \sigma_\infty,$$

anderseits aber

$$\sigma_{m+1} = S_m \sigma_m \geqq S_m \sigma_\infty$$

d. h.

$$\sigma_\infty \geqq S_\infty \sigma_\infty,$$

woraus wieder $\sigma_\infty = S_\infty \sigma_\infty$ folgt.

2°. Hier beweisen wir jetzt mit Hilfe vollständiger Induktion, daß

$$\sigma_n \leqq \varrho_n \quad \text{und} \quad \sigma_k - \sigma_m \geqq \varrho_{m,k} \quad (k < m) \tag{5}$$

gilt. Für $n=0$ bzw. $k=0$, $m=1$ sind die Ungleichungen (5) vorausgesetzt; sind sie schon für $n=0, 1, \ldots, N$, bzw. $m=1, 2, \ldots, N$ und $0<k\leqq m-1$ bewiesen, so gilt (für $k>0$)

$$\varrho_{N+1,k}=\varrho(T_N u_N, T_{k-1}u_{k-1})\leqq P_{k-1}[\varrho(u_N, u_{k-1})+\varrho(u_N, w_0)]-$$

$$-P_N\varrho(u_N, w_0)\leqq P_{k-1}[\sigma_{k-1}-\sigma_N+\sigma_N]-P_N\sigma_N=\sigma_k-\sigma_{N+1},$$

da $\varrho_N\geqq\sigma_N$; für $k=0$ aber

$$\varrho_{N+1,0}\leqq\varrho_{N+1,1}+\varrho_{1,0}\leqq\sigma_1-\sigma_{N+1}+\sigma_0-\sigma_1=\sigma_0-\sigma_{N+1};$$

endlich

$$\varrho_{N+1}=\varrho(u_{N+1}, w_0)\geqq\varrho(w_0, u_0)-\varrho(u_0, u_{N+1})\geqq\sigma_0-[\sigma_0-\sigma_{N+1}]=\sigma_{N+1},$$

und so ist (5) für $n=N+1$, bzw. $m=N+1$, $0\leqq k\leqq N$, d. h. für alle n, m, k bewiesen.

Die Behauptungen in 3° haben Wort für Wort auch jetzt Gültigkeit.

SATZ 1c. *Betrachten wir wieder das Problem des Satzes* 1—1b), *und zwar unter folgenden Bedingungen:*

αc) *für* $k\leqq m\leqq\infty$ *sei*

$$\varrho(T_k u, T_m v)\leqq Q\varrho(u, v)+P_k[\varrho(u, v)+\varrho(v, w_0)]-P_m\varrho(v, w_0)$$

ferner

βc) *die Paare* P_k-P_m $(k\leqq m\leqq\infty)$ *erfüllen diejenigen Bedingungen, wie in* βb), *ferner ist* Q *auch eine stetige und monoton wachsende Funktion des Argumentes, endlich ist*

$$Q\Theta_{\mathfrak{M}}=P_\infty\Theta_{\mathfrak{M}}=\Theta_{\mathfrak{M}}.$$

γc) S_n *ist durch* $S_n\sigma=Q\sigma+P_n\sigma-P_{n+1}\Theta$ *definiert, ferner hat die Gleichung*

$$\sigma=S_\infty\sigma=Q\sigma+P_\infty\sigma$$

unter der Bedingung $\sigma=\varrho(u, v)$; $u, v\in r$, *nur die „triviale" Lösung* $\sigma=\Theta_{\mathfrak{M}}$. *Ferner ist*

$$\sigma_0\geqq\varrho(u_0, u_1)\quad\text{und}\quad\sigma_1<\sigma_0;$$

δc) *Außer den Bedingungen in* δ) *bzw.* δb) *soll auch vorausgesetzt werden, daß die Operatorenfolge* T_n *gleichgradig* $\mathfrak{M}$*-vollstetig ist (d. h. sie bildet eine* $\mathfrak{M}$*-beschränkte Menge in eine* $\mathfrak{M}$*-kompakte ab)*;

εc) *Außer den Bedingungen in* ε) *bzw.* εb) *soll auch vorausgesetzt werden, daß die Menge* r $\mathfrak{M}$*-beschränkt ist, d. h. für jedes* $u, v\in r$ *die Abschätzung* $\varrho(u, v)\leqq\varrho_r$ *feststeht.*

Unter diesen Bedingungen besitzt die Folge $u_{n+1}=T_n u_n$ *einen Grenzwert in* r, *welcher ein Fixpunkt von* T_∞ *ist.*

BEWEIS: 1°. Ebenso wie früher, ist mit vollständiger Induktion gleich einzusehen, daß die Folge σ_n monoton abnehmend und durch $\Theta_{\mathfrak{M}}$ begrenzt ist, also einen Grenzwert σ_∞ besitzt, welcher der Gleichung $\sigma_\infty=S_\infty\sigma_\infty$ genügt. Nach γc) ist also $\sigma_\infty=\Theta_{\mathfrak{M}}$ gültig.

2°. Wir beweisen jetzt durch vollständige Induktion die Gültigkeit von

$$\sigma_n \leqq \varrho(u_n, u_{n+1}). \tag{6}$$

(6) ist nämlich für $n=0$ nach γc) gültig. Ist es schon für $n=N$ bewiesen, so ist

$$\varrho(u_{N+1}, u_{N+2}) = \varrho(T_N u_N, T_{N+1} u_{N+1}) \leqq Q\varrho(u_N, u_{N+1}) +$$

$$+ P_N[\varrho(u_N, u_{N+1}) + \varrho(u_{N+1}, w_0)] - P_{N+1}\varrho(u_{N+1}, w_0) \leqq$$

$$\leqq Q\sigma_N + P_N[\sigma_N + \Theta] - P_{N+1}\Theta = S_N\sigma_N = \sigma_{N+1},$$

da Q monoton wachsend und die Paare P_N, P_{N+1} konkav sind. Somit ist (4) für alle n gültig.

3°. Da die Folge T_n gleichgradig $\mathfrak{M}$-vollstetig, ferner r $\mathfrak{M}$-beschränkt, endlich $u_n \in r$ für alle n ist, besitzt die Folge $\{u_n\}$ eine konvergente Teilfolge $\{u_{n_k}\}$, welche nach $u_\infty \in r$ strebt. Dann strebt aber (4) gemäß auch die Folge $\{u_{n_k+m}\}$ $(m=1, 2, \ldots)$ nach u_∞, d. h. die Folge $\{u_n\}$ selbst auch. Endlich

$$\varrho(u_\infty, T_\infty u_\infty) \leqq \varrho(u_\infty, u_n) + \varrho(T_{n-1}\, u_{n-1}, T_\infty u_\infty) \leqq$$

$$\leqq \varrho(u_n, u_\infty) + Q\varrho(u_{n-1}, u_\infty) + P_{n-1}[\varrho(u_{n-1}, u_\infty) +$$

$$+ \varrho(u_\infty, w_0)] - P_\infty \varrho(u_\infty, w_0) \leqq \varrho(u_n, u_\infty) + Q\varrho(u_{n-1}, u_\infty) +$$

$$+ P_{n-1}[\varrho(u_{n-1}, u_\infty)] \leqq \varrho(u_n, u_\infty) + Q\varrho(u_{n-1}, u_\infty) + P_m[\varrho(u_{n-1}, u_\infty)],$$

falls $m \leqq n-1$ ist. Da nun P_m eine monoton abnehmende Funktion des Indexes und des Argumentes, ferner $P_\infty\Theta = \Theta$ gültig ist, so strebt die rechte Seite unserer Ungleichung nach $P_m\Theta$, falls $n \to \infty$, bzw. nach Θ, falls dann auch $m \to \infty$ gilt. Folglich ist

$$\varrho(u_\infty, T_\infty u_\infty) \leqq \Theta_{\mathfrak{M}},$$

womit alles bewiesen ist.

SATZ 1d. *Betrachten wir das Problem bzw. die Bedingungen des Satzes* 1c), *setzen jetzt jedoch auch* $w_0 \in r$ *voraus, lassen dagegen aber zu, daß die Paare* P_k, P_m *nur asymptotisch „schwach konkav" sind, d. h.*

βd) *für* $k < m \leqq \infty$, $\Theta \leqq \varrho \leqq \varrho^*$; $\Theta \leqq \sigma \leqq \sigma^*$

$$P_k(\varrho^* + \sigma) - P_m\varrho^* \leqq P_k(\varrho + \sigma^*) - P_m\varrho + \alpha_{k,m} R(\varrho^* - \varrho)$$

gilt, wo R *ein positiver, monoton wachsender Operator ist, ferner* $\alpha_{k,m}$ *eine monoton fallende Zahlenfolge beider Indexe, für welche daneben* $\alpha_{k,\infty} \searrow 0$ *für* $k \to \infty$ *auch feststeht*;

γd) S_n *soll durch* $S_n\sigma = Q\sigma + P_n\sigma - P_{n+1}\Theta + \alpha_{n,n+1} R\varrho_r$ *definiert sein.*

Die weiteren Voraussetzungen des Satzes 1c) *beibehaltend, sind auch die Behauptungen dieses Satzes gültig.*

BEWEIS: 1°. Da $\alpha_{n,n+1}$ monoton fallend ist, kann man die Monotonität von σ_n ebenso wie in Satz 1c) beweisen; daraus folgt wiederum die Relation $\sigma_\infty = \Theta_{\mathfrak{M}}$.

2°. Auch die Ungleichung $\sigma_n \geqq \varrho(u_n, u_{n+1})$ ist durch vollständige Induktion leicht zu beweisen: nämlich

$$\varrho(u_{N+1}, u_{N+2}) = \varrho(T_N u_N, T_{N+1} u_{N+1}) \leqq Q\varrho(u_N, u_{N+1}) +$$
$$+ P_N[\varrho(u_N, u_{N+1}) + \varrho(u_{N+1}, w_0)] - P_{N+1}\varrho(u_{N+1}, w_0) \leqq$$
$$\leqq Q\sigma_N + P_N[\sigma_N + \Theta] - P_{N+1}\Theta + \alpha_{N,N+1} R\varrho(u_{N+1}, w_0) \leqq$$
$$\leqq Q\sigma_N + P_N\sigma_N - P_{N+1}\Theta + \alpha_{N,N+1} R\varrho_r = S_N\sigma_N = \sigma_{N+1},$$

falls $\varrho_{N,N+1} \leqq \sigma_N$ gültig war, und damit ist alles bewiesen.

3°. ist Wort für Wort aus Satz 1c) benützbar.

Es sei hier bemerkt, daß man statt $\alpha_{k,m}R$ auch den Operator $R_{k,m}$ benützen kann, falls $R_{k,m}$ monoton abnehmend mit seinen beiden Indexen ist, ferner $R_{k,\infty} \searrow 0$, für $k \to \infty$ auch gültig ist.

Es sei noch bemerkt, daß in Satz 1c) bzw. 1d) die Unizitätsbedingung der Lösung der Gleichung $\sigma = S_\infty \sigma$ die strengste Voraussetzung ist. Mit Hilfe von sehr schweren Hilfsmitteln kann man diese Voraussetzung abschwächen; wir werden darauf in einer späteren Arbeit zurückkommen.

SATZ 2. *In Satz* 1. *ist Voraussetzung* ε) *durch* ε_2): *r enthält die abgeschlossene Pseudokugel*

$$K\{v: \varrho(v, u_1) \leqq \sigma_\infty - \sigma_1\}$$

ersetzbar.

Diese Kugel enthält dann das Limeselement u_∞ *von* $\{u_n\}$, *und in K ist dieses der einzige Fixpunkt von* $T_\infty = T$.

Es sei ferner $n_0 \geqq 1$ *und* $\tau_{n_0} \in \mathfrak{M}$ *erfülle die Unglechung* $\tau_{n_0} \geqq \sigma_{n_0}$, *endlich die Folge*

$$\tau_{n+1} = S_n \tau_n \qquad (n = n_0, n_0 + 1; \ldots)$$

strebe auch nach σ_∞. *Dann enthält die Pseudokugel* $K^*\{v; \varrho(v, u_{n_0}) \leqq \tau_{n_0} - \sigma_{n_0}\}$ *höchstens einen Fixpunkt von* T_∞.

BEWEIS: 1°. Es sei zuerst bemerkt, daß in Satz 1. Punkt 1° die Voraussetzung ε) gar nicht benützt wurde, σ_∞ und somit K existiert also auch im Falle ε_2) statt ε). Es ist nun leicht zu sehen, daß $u_n \in K$ für $n \geqq 1$ gültig ist, da $T_n K \subseteqq K$ feststeht. Es sei denn $v \in K$ beliebig, und $n \geqq 1$; nun

$$\varrho(T_n v, u_1) = \varrho(T_n v, T_0 u_0) \leqq Q[\varrho(u_0, v)] + P_n[\varrho(u_0, v) +$$
$$+ \varrho(u_0, w_0)] - P_0\varrho(u_0, w_0) \leqq Q[\varrho(u_0, u_1) + \varrho(u_1, v) +$$
$$+ \varrho(u_0, w_0)] - Q[\varrho(u_0, w_0)] + P_\infty[\varrho(u_0, u_1) + \varrho(u_1, v) +$$
$$+ \varrho(u_0, w_0)] - P_0[\varrho(u_0, w_0)] \leqq Q(\sigma_1 - \sigma_0 + \sigma_\infty - \sigma_1 + \sigma_0) -$$
$$- Q(\sigma_0) + P_\infty(\sigma_1 - \sigma_0 + \sigma_\infty - \sigma_1 + \sigma_0) - P_0(\sigma_0) =$$
$$= S_\infty \sigma_\infty - S_0 \sigma_0 = \sigma_\infty - \sigma_1,$$

d. h. $T_n v \in K$, w. z. b. w. Da nun $T_n K \subseteqq K$ gültig ist, so ist ε) eine Folge von ε_2).

2°. Es sei vorausgesetzt, dass $w \in K$ auch ein Fixpunkt von T_∞ ist. Da $w \in K$, so ist $\varrho(w, u_1) \leqq \sigma_\infty - \sigma_1$. Dann kann man aber mit vollständiger Induktion leicht zeigen, daß

$$\varrho(w, u_n) \leqq \sigma_\infty - \sigma_n \tag{7}$$

für jedes $n \geqq 1$ gültig ist. Steht nämlich (7) für $n = N$ fest, so gilt

$$\varrho(w, u_{N+1}) = \varrho(T_\infty w, T_N u_N) \leqq Q[\varrho(w, u_N)] + P_\infty[\varrho(w, u_N) +$$
$$+ \varrho(u_N, w_0)] - P_N[\varrho(u_N, w_0)] \leqq Q[\varrho(w, u_N) + \varrho(u_N, w_0)] -$$
$$- Q\varrho(u_N, w_0) + P_\infty[\varrho(w, u_N) + \varrho_N] - P_N\varrho_N \leqq Q[\sigma_\infty - \sigma_N +$$
$$+ \sigma_N] - Q\sigma_N + P_\infty[\sigma_\infty - \sigma_N + \sigma_N] - P_N\sigma_N = S_\infty\sigma_\infty - S_N\sigma_N = \sigma_\infty - \sigma_{N+1},$$

d. h. (7) ist auch für $n = N+1$, d. h. für alle $n \geqq 1$ gültig. Daraus folgt aber gleich, daß $u_n \to w$, d. h. $w = u_\infty$ ist.

Die Behauptung über K^* ist Wort für Wort ähnlich wie (7) zu beweisen.

SATZ 2b. *Die Voraussetzung* εb) *in Satz* 1b) *ist durch* ε_2b): *r enthält die abgeschlossene Pseudokugel*

$$K\{v;\ \varrho(v, u_1) \leqq \sigma_1 - \sigma_\infty\}$$

ersetzbar.

Diese Pseudokugel enthält dann auch u_∞, *und dieser ist der einzige Fixpunkt von* T_∞ *in* K.

Ist ferner $n_0 \geqq 1$ *und* $\tau_{n_0} \leqq \sigma_{n_0}$, *strebt daneben* $\tau_{n+1} = S_n\tau_n$ $(n \geqq n_0)$ *auch nach* σ_∞, *so besitzt die Pseudokugel* $K^*\{v;\ \varrho(v, u_{n_0}) \leqq \sigma_{n_0} - \tau_{n_0}\}$ *höchstens einen Fixpunkt von* T_∞.

BEWEIS: 1°. Die Existenz von σ_∞ bzw. K ist jetzt wieder unabhängig von εb). Auch $T_nK \subseteqq K$ ist wieder leicht einzusehen: ist nämlich $v \in K$, so

$$\varrho(Tv, u_1) = \varrho(Tv, T_0u_0) \leqq P_0[\varrho(u_0, v) + \varrho(u_0, w_0)] -$$
$$- P\varrho(u_0, w_0) \leqq P_0[\varrho(u_0, u_1) + \varrho(u_1, v) + \varrho(u_0, w_0)] - P_\infty\varrho_0 \leqq$$
$$\leqq P_0[\sigma_0 - \sigma_1 + \sigma_1 - \sigma_\infty + \sigma_\infty] - P_\infty\sigma_\infty = S_0\sigma_0 - S_\infty\sigma_\infty = \sigma_1 - \sigma_\infty,$$

da P_n monoton abnehmend mit seinem Index ist, und $\varrho(u_0, w_0) \geqq \sigma_0$, um so mehr $\varrho(u_0, w_0) \geqq \sigma_\infty$. Damit haben wir also schon bewiesen, daß ε_2b) die Voraussetzungen von εb) mitzieht.

2°. Es sei nun wieder $w \in K$ ein Fixpunkt von T_∞. Wir zeigen durch vollständige Induktion, daß

$$\varrho(w, u_n) \leqq \sigma_n - \sigma_\infty \tag{8}$$

feststeht. Für $n = 1$ ist ja (8) gültig, da $w \in K$. Ist (8) schon für $n = N$ bewiesen, so gilt

$$\varrho(w, u_{N+1}) = \varrho(T_\infty w, T_N u_N) \leqq P_N[\varrho(u_N, w) + \varrho(u_N, w_0)] -$$
$$- P_\infty[\varrho(u_N, w_0)] \leqq P_N[\sigma_N - \sigma_\infty + \sigma_\infty] - P_\infty\sigma_\infty = S_N\sigma_N - S_\infty\sigma_\infty = \sigma_{N+1} - \sigma_\infty,$$

da $\varrho(u_N, w_0) \geqq \sigma_N$ nach (5), um so mehr also $\varrho(u_N, w_0) \geqq \sigma_\infty$. Damit haben wir also (8) für jedes $n \geqq 1$ bewiesen.

Im weiteren können wir den Gedankengang des Satzes 2 Wort für Wort folgen.

SATZ 2c. *Unter den Voraussetzungen des Satzes* 1c *besitzt* T_∞ *keinen Fixpunkt außer* u_∞ *im Gebiet* $K_\vartheta\{v: P_\infty[\varrho(u_\infty, v)] + Q[\varrho(u_\infty, v)] \leqq \vartheta\varrho(u_\infty, v)\} \cap r$, *falls* $0 \leqq \vartheta < 1$ *ist.*

BEWEIS: Es sei vorausgesetzt, daß $w \in K_\vartheta \cap r$ ein Fixpunkt ist. Dann aber ist

$$\varrho(u_\infty, w) = \varrho(T_\infty u_\infty, T_\infty w) \leqq Q\varrho(u_\infty, w) + P_\infty \varrho(u_\infty, w) \leqq \vartheta\varrho(u_\infty, w), \tag{10}$$

da $w \in K_\vartheta \cap r$. Ist nun $0 \leqq \vartheta < 1$, so folgt $\varrho(u_\infty, w) = \Theta$ aus (10), w. z. b. w.

SATZ 2d. *Unter den Voraussetzungen des Satzes* 1d *besitzt* T_∞ *keinen Fixpunkt außer* u_∞ *im Gebiet*

$$K_\vartheta\{v: P_\infty \varrho(u_\infty, v) + Q\varrho(u_\infty, v) \leqq \vartheta\varrho(u_\infty, v)\} \cap r, \quad \textit{falls} \quad 0 \leqq \vartheta < 1 \quad \textit{ist.}$$

Der BEWEIS entspricht dem des vorigen Satzes.

SATZ 3. *Unter den Voraussetzungen des Satzes* 1. *kann man die Konvergenzschnelle der Folge* u_n *nach* u_∞ *durch*

$$\varrho(u_\infty, u_n) \leqq \sigma_\infty - \sigma_n$$

abschätzen.

BEWEIS: Für beliebiges $N \geqq n$ gilt die Abschätzung

$$\varrho(u_\infty, u_n) \leqq \varrho(u_\infty, u_N) + \varrho(u_N, u_n) \leqq \varrho(u_\infty, u_N) + \sigma_N - \sigma_n.$$

Strebt also $N \to \infty$, so strebt $\varrho(u_\infty, u_N)$ nach $\Theta_{\mathfrak{M}}$, folglich ist

$$\varrho(u_\infty, u_n) \leqq \Theta_{\mathfrak{M}} + \sigma_\infty - \sigma_n,$$

w. z. b. w.

SATZ 3b. *Unter den Voraussetzungen des Satzes* 1b *ist die Abschätzung*

$$\varrho(u_\infty, u_n) \leqq \sigma_n - \sigma_\infty$$

gültig.

Der BEWEIS entschpricht dem des vorigen Satzes.

SATZ 3c. *Unter den Voraussetzungen des Satzes* 1c *sei* S_n^* *durch*

$$S_n^*\tau = Q\tau + P_n\tau$$

definiert; es sei ferner $\mathfrak{M} \ni \tau_0 \geqq \varrho(u_0, u_\infty)$ *so gewählt, daß auch*

$$\tau_1 = S_0^*\tau_0 \leqq \tau_0$$

gilt. Dann gilt die Abschätzung

$$\varrho(u_n, u_\infty) \leqq \tau_n = S_{n-1}^*\tau_{n-1},$$

und hier $\tau_n \to \Theta_{\mathfrak{M}}$, *falls* $n \to \infty$.

BEWEIS: 1°. Da hier $\tau_0 \geqq \varrho(u_0, u_\infty)$ und $\tau_1 < \tau_0$ gültig ist, kann man durch vollständige Induktion zeigen, daß τ_n monoton abnehmend ist. Ist dann $\tau_n - \tau_{n+1} \geqq \Theta$ für $n = N-1$ noch gültig, so muß

$$\tau_N - \tau_{N+1} = S^*_{N-1}\tau_{N-1} - S^*_N \tau_N = Q\tau_{N-1} + P_{N-1}\tau_{N-1} - Q\tau_N - P_N\tau_N \geqq \Theta$$

auch gelten, da Q monoton nichtabnehmend ist, bzw. nach den Voraussetzungen von βc). Da nun weiter $\tau_n \geqq \Theta$ auch feststeht, so muss τ_n nach τ_∞ streben, da mit der Folge S_n auch S^*_n gleichgradig vollstetig ist; da ferner $S^*_\infty = S_\infty$ auch gültig ist, soll τ_∞ nach γc) gleich $\Theta_{\mathfrak{M}}$ sein.

2°. Die Abschätzung

$$\varrho(u_n, u_\infty) \leqq \tau_n \tag{11}$$

ist auch leicht mit Hilfe vollständiger Induktion zu prüfen. (11) ist nämlich für $n = 0$ nach Voraussetzung erfüllt. Ist es schon für $n = N$ bewiesen, so ist

$$\varrho(u_{N+1}, u_\infty) = \varrho(T_N u_N, T_\infty u_\infty) \leqq Q\varrho(u_N, u_\infty) +$$

$$+ P_N[\varrho(u_N, u_\infty) + \varrho(u_\infty, w_0)] - P_\infty \varrho(u_\infty, w_0) \leqq$$

$$\leqq Q\tau_N + P_N[\tau_N - \Theta] - P_\infty \Theta = S^*_N \tau_N = \tau_{N+1},$$

und damit ist alles bewiesen.

SATZ 3d. *Unter den Bedingungen des Satzes* 1d *soll* S^*_n *durch*

$$S^*_n \tau = Q\tau + P_n \tau + \alpha_{n,n+1} R\varrho_r \quad (\text{bzw.} + R_{n,n+1}\varrho_r)$$

definiert sein; ferner soll $\mathfrak{M} \ni \tau_0 \geqq \varrho(u_0, u_\infty)$ *so gewählt werden, daß*

$$\tau_1 = S^*_0 \tau_0 \leqq \tau_0$$

auch feststeht.

Dann strebt die Folge $\tau_{n+1} = S^*_n \tau_n$ *nach* Θ, *und es gilt die Abschätzung*

$$\varrho(u_n, u_\infty) \leqq \tau_n. \tag{12}$$

BEWEIS: 1°. Wort für Wort ähnlich, wie im früheren Satze, man kann leicht zeigen, daß τ_n monoton abnehmend nach Θ strebt.

2°. (12) ist nach Voraussetzung für $n = 0$ erfüllt. Ist es schon für $n = N$ bewiesen, so ist

$$\varrho(u_{N+1}, u_\infty) = \varrho(T_N u_N, T_\infty u_\infty) \leqq Q\varrho(u_N, u_\infty) + P_N[\varrho(u_N, u_\infty) + \varrho(u_\infty, w_0)] -$$

$$- P_\infty \varrho(u_\infty, w_0) \leqq Q\tau_N + P_N[\tau_N + \Theta_{\mathfrak{M}}] - P_\infty \Theta_{\mathfrak{M}} + \alpha_{N,\infty} R[\varrho(u_\infty, w_0) - \Theta_{\mathfrak{M}}] \leqq$$

$$\leqq Q\tau_N + P_N\tau_N + \alpha_{N,N+1} R\varrho_r = S^*_N \tau_N = \tau_{N+1},$$

und damit ist alles bewiesen.

3. §. Anwendungen bei dem Picardschen bzw. Picard-Carathéodorischen Iterationsverfahren

Im weiteren wenden wir die Fixpunktsätze bei der Untersuchung einiger Iterationsverfahren an, welche wir bei der Lösung der Anfangswertaufgabe des Differentialgleichungssystems

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}); \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{13}$$

benützen.

A°. Es sei zunächst vorausgesetzt, dass $\mathbf{f}(t, \mathbf{x})$ bei beliebiger $\mathbf{x}(t) \in C(-\infty, \infty)$ meßbar und an jedem endlichen Intervall integrierbar ist, ferner in $\mathbf{x}$ eine Lipschitz-Bedingung (mit Exponent 1) erfüllt, d. h.

$$|\mathbf{f}(t, \mathbf{x}_2) - \mathbf{f}(t, \mathbf{x}_1)| \leqq L(t) \cdot |\mathbf{x}_2 - \mathbf{x}_1| \tag{14}$$

gültig ist, wo $L(t)$ eine meßbare und am jeden endlichen Intervall integrierbare Funktion ist.

Es sei ferner

$$\left| \int_{t_0}^{t} |\mathbf{f}(\tau, \mathbf{x}_0)| \, d\tau \right| = K(t); \quad \left| \int_{t_0}^{t} L(\tau) \, d\tau \right| = \lambda(t). \tag{15}$$

Wir werden im weiteren die Picardsche Iterationsfolge

$$\mathbf{x}_{n+1}(t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}[\tau, \mathbf{x}_n(\tau)] \, d\tau \tag{16}$$

untersuchen. Zu diesem Zweck sei $\mathfrak{M}$ der Raum $C(-\infty, \infty)$, wo wir die Halbordnung durch

$$f \leqq g, \quad \text{falls} \quad f(t) \leqq g(t) \qquad (-\infty < t < \infty)$$

definieren, den abstrakten Konvergenzbegriff aber mit demjenigen „gleichmäßige Konvergenz über allen endlichen Intervallen" identifizieren.

In $C(-\infty, \infty)$ führen wir folgenden Pseudoabstand ein:

$$\varrho(\mathbf{u}(t), \mathbf{v}(t)) = e^{-\lambda(t)} \cdot |\mathbf{u}(t) - \mathbf{v}(t)|. \tag{17}$$

Da wir jetzt immer denselben Operator anwenden, gilt

$$T_n \mathbf{u} \equiv T\mathbf{u} = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}[\tau, \mathbf{u}(\tau)] \, d\tau,$$

folglich ist

$$\varrho(T\mathbf{u}, T\mathbf{v}) = e^{-\lambda(t)} \cdot \left| \int_{t_0}^{t} \{\mathbf{f}[\tau, \mathbf{u}(\tau)] - \mathbf{f}[\tau, \mathbf{v}(\tau)]\} \, d\tau \right| \leqq \tag{18}$$

$$\leqq e^{-\lambda(t)} \cdot \left| \int_{t_0}^{t} L(\tau) \cdot e^{\lambda(\tau)} \cdot \varrho(\mathbf{u}, \mathbf{v}) \, d\tau \right|.$$

Betrachten wir nur das Gebiet $t \geqq t_0$, so gilt

$$\varrho(T\mathbf{u}, T\mathbf{v}) = \int_{t_0}^{t} L(\tau) \cdot e^{-[\lambda(t)-\lambda(\tau)]} \varrho(\mathbf{u}, \mathbf{v})\, d\tau, \tag{19}$$

folglich ist dann $P_n \equiv 0$, $Q\varrho = \int_{t_0}^{t} L(\tau) \cdot e^{-[\lambda(t)-\lambda(\tau)]} \varrho\, d\tau$, d. h. ein linearer Operator, und ist somit Satz 1—2—3 anwendbar.

Da nun ferner hier Q linear ist, kann man $w_0 = u_0$, und somit $\sigma_0 = \Theta$ wählen. Es gilt folglich $\sigma_1 = \tau$, ferner

$$\sigma_2 = Q\sigma_1 + \tau = Q\sigma_1 + \sigma_1 = (E+Q)\sigma_1,$$

usf., und allgemein

$$\sigma_{n+1} = Q\sigma_n + \tau = Q\sigma_n + \sigma_1 = (E + Q + \ldots + Q^n)\sigma_1. \tag{20}$$

Die Konvergenz der Folge σ_n ist also hier gleichbedeutend mit der Konvergenz der Reihe $\sum_{n=0}^{\infty} Q^n$; letztere ist aber in unserem Falle die Resolvente der Integralgleichung

$$\varrho(t) - \int_{t_0}^{t} L(\xi) e^{-[\lambda(t)-\lambda(\xi)]} \varrho(\xi)\, d\xi = \psi(t). \tag{21}$$

Letztere ist keine Gleichung des Typs Volterra, da ihr Kern,

$$\boldsymbol{K}_1(t, \xi) = \begin{cases} 0, & \text{falls} \quad t < \xi \\ L(\xi) e^{-[\lambda(t)-\lambda(\xi)]}, & \text{falls} \quad t \geqq \xi \end{cases} \tag{22}$$

nicht beschränkt ist; jedoch kann man die iterierten Kerne in geschlossener Form darstellen, und somit die Konvergenz der Reihe $\sum Q^n$ zeigen, da

$$\boldsymbol{K}_1(t, \xi) = \frac{d}{d\xi} e^{-[\lambda(t)-\lambda(\xi)]} \quad \text{und} \quad L(\xi) = -\frac{d}{d\xi}[\lambda(t) - \lambda(\xi)] \tag{23}$$

für $t \geqq \xi$ fast überall gültig ist und beide Derivierten integrierbar sind. (Es sei bemerkt, daß man in dem Falle, in dem die Konvergenz von σ_n gegen σ_∞ direkt zu zeigen ist, die Bedingung δ) nicht voraussetzen muß.) So nämlich

$$K_2(t, \xi) = 0, \quad \text{falls} \quad t < \xi, \quad \text{bzw.} = \int_{\xi}^{t} L(s) e^{-[\lambda(t)-\lambda(\xi)]}.$$

$$L(\xi) \cdot e^{-[\lambda(s)-\lambda(\xi)]}\, ds = L(\xi) \cdot [\lambda(t) - \lambda(\xi)] \cdot e^{-[\lambda(t)-\lambda(\xi)]}, \quad \text{falls} \quad t \geqq \xi,$$

bzw. auch allgemein, wenn

$$\boldsymbol{K}_n(t, \xi) = \begin{cases} 0, & \text{falls} \quad t < \xi \\ L(\xi) e^{-[\lambda(t)-\lambda(\xi)]} \cdot \dfrac{[\lambda(t) - \lambda(\xi)]^{n-1}}{(n-1)!}, & \text{falls} \quad t \geqq \xi \end{cases} \tag{24}$$

für $n=1, 2, \ldots, N$ noch gültig ist, so

$$K_{N+1}(t, \xi) = 0, \quad \text{falls} \quad t < \xi,$$

$$\text{bzw} = \int_{\xi}^{t} L(s) e^{-[\lambda(t)-\lambda(s)]} \cdot L(\xi) e^{-[\lambda(s)-\lambda(\xi)]} \cdot \frac{[\lambda(s)-\lambda(\xi)]^{N-1}}{(N-1)!}\, d\xi =$$

$$= L(\xi) e^{-[\lambda(t)-\lambda(\xi)]} \cdot \frac{1}{N!} [\lambda(t)-\lambda(\xi)]^{N}, \quad \text{falls} \quad t \geqq \xi$$

auch feststeht, d. h. (23) ist für jedes n gültig.

Ist nun $\mathbf{u}_0 = \mathbf{x}_0$ als Anfangsglied gewählt, so ist

$$\mathbf{u}_1 = \mathbf{x}_0 + \int_{0}^{t} \mathbf{f}[\xi, \mathbf{x}_0]\, d\xi,$$

und deshalb

$$\tau = \sigma_1 = \varrho(\mathbf{u}_0, \mathbf{u}_1) \leqq e^{-\lambda(t)} \cdot \left| \int_{t_0}^{t} |\mathbf{f}(\xi, \mathbf{x}_0)|\, d\xi \right| \leqq e^{-\lambda(t)} \cdot K(t), \tag{24}$$

ferner

$$\sigma_n = (E + Q + Q^2 + \ldots + Q^{n-1})\sigma_1 = \tag{25}$$

$$= \sigma_1 + \int_{t_0}^{t} L(\xi) \cdot e^{-[\lambda(t)-\lambda(\xi)]} \cdot \sum_{k=0}^{n-2} \frac{[\lambda(t)-\lambda(\xi)]^k}{k!}\, \sigma_1\, d\xi \to \sigma_1 + \int_{t_0}^{t} L(\xi)\sigma_1(\xi)\, d\xi = \sigma_\infty,$$

da $\sum_{k=0}^{n-2} \frac{[\lambda(t)-\lambda(\xi)]^k}{k!} \to e^{[\lambda(t)-\lambda(\xi)]}$, und zwar gleichmäßig an jedem endlichen Intervall.

Somit ist

$$\sigma_\infty - \sigma_n = \int_{t_0}^{t} L(\xi) \cdot \left\{ 1 - e^{-[\lambda(t)-\lambda(\xi)]} \sum_{k=0}^{n-2} \frac{[\lambda(t)-\lambda(\xi)]^k}{k!} \right\} \sigma_1(\xi)\, d\xi \leqq$$

$$\leqq K(t) e^{-\lambda(t)} \int_{t_0}^{t} L(\xi) \left\{ e^{[\lambda(t)-\lambda(\xi)]} - \sum_{k=0}^{n-2} \frac{[\lambda(t)-\lambda(\xi)]^k}{k!} \right\} d\xi =$$

$$= K(t) \cdot e^{-\lambda(t)} \int_{0}^{\lambda(t)} \left\{ e^{v} - \sum_{k=0}^{n-2} \frac{v^k}{k!} \right\} dv = K(t) \cdot e^{-\lambda(t)} \left\{ e^{\lambda(t)} - \sum_{k=0}^{n-1} \frac{[\lambda(t)]^k}{k!} \right\},$$

d. h., beachtend Satz 3. bzw. die Definition von ϱ:

$$|\mathbf{x}(t) - \mathbf{x}_n(t)| \leqq K(t) \left\{ e^{\lambda(t)} - \sum_{k=0}^{n-1} \frac{[\lambda(t)]^k}{k!} \right\}, \tag{26}$$

wo $\mathbf{x}(t)$ nach Satz 2 die einzige stetige sog. verallgemeinerte Lösung des Anfangswertproblems (13), $\mathbf{x}_n(t)$ aber das n-te Glied der Iterationsfolge (16) bezeichnet. (26) gibt somit eine Abschätzung über die Konvergenzschnelle des Picardschen

Iterationsverfahrens in bezug auf n und t, bzw. die betrachteten Voraussetzungen. Es sei noch bemerkt, daß im Falle, daß unsere Voraussetzungen nur in einem Bereich erfüllt sind, auch die Behauptungen für diesen Bereich gültig sind.

B°. Betrachten wir wieder das Problem in A°., setzen wir jedoch statt (14) nur voraus, daß

$$|\mathbf{f}(t, \mathbf{x}_2)-\mathbf{f}(t, \mathbf{x}_1)| \leqq g(t, |\mathbf{x}_2-\mathbf{x}_1|) \tag{27}$$

gültig ist, und hier g eine monoton wachsende, stetige Funktion bei fixiertem t des zweiten Argumentes, ferner eine meßbare Funktion bei fixiertem $|\mathbf{x}_2-\mathbf{x}_1|$ des ersten Argumentes ist, mit $g(t, 0)\equiv 0$, endlich, daß eine meßbare und auf jedem endlichen Intervall integrierbare Majorantenfunktion $m(t)$ bzw. $\varphi(v)$ existiert, so daß

$$0 \leqq g(t, v) \leqq m(t)\cdot\varphi(v); \qquad (-\infty < t < \infty) \tag{28}$$

und hier

$$\varphi(v) \leqq C(1+v^n) \tag{29}$$

mit entsprechendem n.

Um einen unserer Fixpunktsätze anzuwenden, wählen wir ebenso wie in A°, die Operatoren $T_n \equiv T$ $(n=0, 1, 2, \ldots)$, ferner den Raum $\mathbf{C}$, S bzw. $\mathfrak{M}$, den Pseudoabstand aber in Analogiam mit (17) durch

$$\varrho^*(\mathbf{u}, \mathbf{v}) = \left| \int_{t_0}^{t} g(\zeta, |\mathbf{u}-\mathbf{v}|)\, d\zeta \right| \tag{30}$$

oder durch

$$\varrho(\mathbf{u}, \mathbf{v}) = |\mathbf{u}(t)-\mathbf{v}(t)|. \tag{31}$$

In beiden Fällen bekommen wir dieselbe Formel, nämlich:

$$\varrho^*(T\mathbf{u}, T\mathbf{v}) = \left| \int_{t_0}^{t} g\left(t, \left| \int_{t_0}^{\zeta} \mathbf{f}(\xi, \mathbf{v}(\xi))\, d\xi - \int_{t_0}^{\zeta} \mathbf{f}(\xi, \mathbf{u}(\xi))\, d\xi \right|\right) d\zeta \right| \leqq \tag{32}$$

$$\leqq \left| \int_{t_0}^{t} g(\zeta, \varrho^*(\mathbf{u}, \mathbf{v}))\, d\zeta \right|$$

bzw.

$$\varrho(T\mathbf{u}, T\mathbf{v}) \leqq \left| \int_{t_0}^{t} g(\zeta, \varrho(\mathbf{u}, \mathbf{v}))\, d\zeta \right|. \tag{33}$$

Wenn wir wieder den Fall $t \geqq t_0$ betrachten und ϱ bzw. ϱ^* durch $\hat{\varrho}$ bezeichnen, so gilt

$$\hat{\varrho}(T\mathbf{u}, T\mathbf{v}) \leqq \int_{t_0}^{t} g(\xi, \hat{\varrho}(\mathbf{u}, \mathbf{v}))\, d\xi. \tag{34}$$

Im weiteren betrachten wir nur den nach A° noch interessanten Fall, wo $g(t, v)$ eine konkave Funktion des zweiten Argumentes ist; dann können wir aber nur den Satz 1c—2c—3c anwenden. So wird also r eine echte Teilmenge ($\mathfrak{M}$-Beschränktheit!) der Funktionen von $\mathbf{C}[t_0, \infty)$, die auch die Voraussetzung $\mathbf{v}(t_0) = \mathbf{x}_0$ erfüllen.

Nach Satz 1c. werden wir also auch voraussetzen, dass die Integralungleichung

$$\hat{\varrho}(t) \leqq \int_{t_0}^{t} g(\xi; \hat{\varrho}(\xi))\, d\xi \tag{35}$$

für $\hat{\varrho}(t_0) = 0$ nur die triviale stetige Lösung $\hat{\varrho}(t) \equiv 0$ besitzt. Es sei ferner ψ eine entsprechende, monoton wachsende, konkave Funktion mit $\psi(0) = 0$, für welche die Integralgleichung

$$\hat{\varrho}^*(t) = \int_{t_0}^{t} g(\xi; \psi(\hat{\varrho}^*(\xi)))\, d\xi \tag{36}$$

auch eine stetige nichttriviale Lösung für $\hat{\varrho}^*(t_0) = 0$ besitzt (wie gut bekannt ist, z. B. $\psi(v) = v^\alpha$ mit $0 < \alpha < 1$ unseren Voraussetzungen genügt). Bezeichne nun $\hat{\varrho}_f(t)$ das sog. „obere Integral" von (36) — wie es nämlich Carathéodori bewies, (36) besitzt ein solches stetiges oberes Integral für $t \geqq t_0$, mit $\varrho_f(t_0) = 0$.

Es sei ferner $\mathbf{k}(t) = \mathbf{x}_0(t) = \mathbf{u}_0(t)$ das Anfangselement der Picardschen Iteration, welches man so wählt, daß

$$\mathbf{k}(t) \in \mathbf{C}[t_0, \infty);\ \mathbf{k}(t_0) = \mathbf{x}_0;\ \hat{\varrho}(\mathbf{k}, T\mathbf{k}) \leqq \hat{\varrho}_f \tag{37}$$

erfüllt seien (so entspricht z. B. $\mathbf{k}(t) \equiv \mathbf{x}_0$ sicher diesen Voraussetzungen). $r \subset \mathbf{C}[t_0, \infty)$ kann man also folgendermassen wählen: für jedes Elementenpaar $\mathbf{u}, \mathbf{v} \in r$ seien folgende Bedingungen erfüllt:

$$\mathbf{u}(t_0) = \mathbf{v}(t_0) = \mathbf{x}_0;\quad \hat{\varrho}(\mathbf{u}, \mathbf{k}) \leqq \hat{\varrho}_f;\quad \hat{\varrho}(\mathbf{v}, \mathbf{k}) \leqq \hat{\varrho}_f;\quad \varrho(\mathbf{u}, \mathbf{v}) \leqq \hat{\varrho}_f.$$

SATZ 4. *Unter den oben angegebenen Bedingungen sind alle Bedingungen des Satzes* 1c—2c—3c *erfüllt, strebt also die Iterationsfolge*

$$\mathbf{x}_{n+1} = T\mathbf{x}_n = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{x}_n(\tau))\, d\tau \tag{38}$$

nach dem einzigen Fixpunkt $\mathbf{x}_\infty$ *von* T *in* r. *Die Konvergenzschnelle ist durch*

$$\hat{\varrho}(\mathbf{x}_\infty, \mathbf{x}_n) \leqq \tau_n \tag{39}$$

angegeben, wo $\tau_0 = \hat{\varrho}_f$, *und* $\tau_{n+1} = \int_{t_0}^{t} g(\xi; \tau_n(\xi))\, d\xi$ *definiert ist.*

C°. Für numerische Anwendungen scheint es am günstigsten, das Picardsche und das Carathéodorische Verfahren entsprechend zu mischen, d. h. $\mathbf{x}_{n+1}(t)$ mit

Hilfe einer Mischung des „Carathéodorischen" Term $\int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{x}_{n+1}(\tau-\vartheta))\,d\tau$, des „Picardschen" Term $\int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{x}_n(\tau))\,d\tau$, bzw. vielleicht eines „Picardschen kompensatorischen" Term $\int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{x}_n(\tau+\vartheta))\,d\tau$ darstellen, wo man ϑ so klein wählt, wie es numerisch möglich ist. Es ist leicht zu verstehen, daß man dann eine gute Konvergenzschnelle erreichen kann, wenn man in den ersten Schritten (dann liegt noch $\mathbf{x}_n(t)$ weit von $\mathbf{x}_\infty(t)$) den „Carathéodorischen" Term, später aber immer stärker den „Picardschen" Term beschwert.

Unter den Voraussetzungen von A° bzw. B° werden wir erst die Formel

$$\mathbf{x}_{n+1}(t) = \mathbf{x}_0 + \varepsilon_n \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{x}_{n+1}(\tau-\vartheta))\,d\tau + (1-\varepsilon_n) \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{x}_n(\tau))\,d\tau \tag{40}$$

betrachten (im weiteren setzen wir voraus, daß $\mathbf{x}_{n+1}(t) \equiv \mathbf{x}_0$ für $t_0 - \vartheta \leqq t \leqq t_0$ in (40) genommen ist).

Der Operator T_n ist somit durch

$$\boldsymbol{\omega}_n = T_n\mathbf{u} = \mathbf{x}_0 + \varepsilon_n \int_{t_0}^{t} \mathbf{f}(\tau, \boldsymbol{\omega}_n(\tau-\vartheta))\,d\tau + (1-\varepsilon_n) \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{u}(\tau))\,d\tau \tag{41}$$

definiert, der Pseudoabstand sei aber durch

$$\varrho(\mathbf{u}, \mathbf{v}) = \max_{t_0 \leqq \tau \leqq t} |\mathbf{u}(\tau) - \mathbf{v}(\tau)| \tag{42}$$

charakterisiert. Dann strebt T_n gleichmässig nach T, falls $\varepsilon_n \searrow 0$, und falls wir nur diejenige Teilmenge $r^* \subset C[t_0, \infty)$ betrachten, für welche $\mathbf{u} \in r^*$; $\mathbf{u}(\tau) \equiv \mathbf{x}_0$ für $t_0 - \vartheta \leqq \tau \leqq t_0$ feststeht. Dann ist ja auch

$$\varrho(\mathbf{u}(t-\vartheta), \mathbf{v}(t-\vartheta)) \leqq \varrho(\mathbf{u}(t), \mathbf{v}(t)) \tag{43}$$

gültig. Mit Hilfe dieser Ungleichung werden wir erst eine grobe Abschätzung für $\varrho(T_n\mathbf{u}, T_m\mathbf{v})$ angeben, und diese zur Verfeinerung selbst benützen. Zu diesem Zweck führen wir die folgenden Bezeichnungen ein:

$$\begin{aligned} T_n\mathbf{u} = \boldsymbol{\omega}_n &= \mathbf{x}_0 + \varepsilon_n \int_{t_0}^{t} \mathbf{f}(\tau, \boldsymbol{\omega}_n(\tau-\vartheta))\,d\tau + (1-\varepsilon_n) \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{u}(\tau))\,d\tau \\ T_m\mathbf{v} = \mathbf{v}_m &= \mathbf{x}_0 + \varepsilon_m \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{v}_m(\tau-\vartheta))\,d\tau + (1-\varepsilon_m) \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{v}(\tau))\,d\tau. \end{aligned} \tag{44}$$

Hier werden die zweiten Glieder an der rechten Seite auch mit Hilfe einer Transformierten von $\mathbf{u}$ bzw. $\mathbf{v}$ angegeben; z. B.:

$$\mathbf{f}(\tau, \omega_n(\tau-\vartheta)) = \mathbf{f}(\tau, W_n\mathbf{u}) =$$

$$= \begin{cases} \mathbf{f}(\tau, \mathbf{x}_0), & \text{falls} \quad t_0 \leqq \tau \leqq t_0+\vartheta \\ \mathbf{f}\left[\tau, \mathbf{x}_0+\varepsilon_n \int\limits_{t_0}^{\tau-\vartheta} \mathbf{f}(\xi, \mathbf{x}_0)\, d\xi + (1-\varepsilon_n) \int\limits_{t_0}^{\tau-\vartheta} \mathbf{f}(\xi; \mathbf{u}(\xi))\, d\xi\right], & \\ & \text{falls} \quad t_0+\vartheta \leqq \tau \leqq t_0+2\vartheta, \\ \mathbf{f}\left[\tau, \mathbf{x}_0+\varepsilon_n \left\{\int\limits_{t_0}^{t_0+\vartheta} \mathbf{f}(\xi; \mathbf{x}_0)\, d\xi + \int\limits_{t_0+\vartheta}^{\tau-\vartheta} \mathbf{f}\left\langle \xi; \mathbf{x}_0+\varepsilon_n \int\limits_{t_0}^{\xi-\vartheta} \mathbf{f}(\zeta, \mathbf{x}_0)\, d\zeta + \right.\right.\right. & \\ \left.\left.\left. + (1-\varepsilon_n) \int\limits_{t_0}^{\xi-\vartheta} \mathbf{f}(\zeta, \mathbf{u}(\zeta))\, d\zeta \right\rangle d\xi \right\} + (1-\varepsilon_n) \int\limits_{t_0}^{\tau-\vartheta} \mathbf{f}(\xi, \mathbf{u}(\xi))\, d\xi\right], & \\ & \text{falls} \quad t_0+2\vartheta \leqq \tau \leqq t_0+3\vartheta, \\ \text{usw.} & \end{cases}$$

Somit ist

$$\mathbf{\omega}_n(t) = \mathbf{x}_0 + \varepsilon_n \int\limits_{t_0}^{t} \mathbf{f}(\tau, W_n\mathbf{u})\, d\tau + (1-\varepsilon_n) \int\limits_{t_0}^{t} \mathbf{f}(\tau, \mathbf{u})\, d\tau$$

(45)

$$\mathbf{v}_m(t) = \mathbf{x}_0 + \varepsilon_m \int\limits_{t_0}^{t} \mathbf{f}(\tau, W_m\mathbf{v})\, d\tau + (1-\varepsilon_m) \int\limits_{t_0}^{t} \mathbf{f}(\tau, \mathbf{v})\, d\tau,$$

folglich — vorausgesetzt, daß $n \geqq m$, d. h. $\varepsilon_n \leqq \varepsilon_m$ gültig ist —

(46)
$$\varrho(T_n\mathbf{u}, T_m\mathbf{v}) = \max_{t_0 \leqq \tau \leqq t} |\mathbf{\omega}_n(\tau) - \mathbf{v}_m(\tau)| \leqq$$

$$\leqq (1-\varepsilon_n) \int\limits_{t_0}^{t} g(\tau, |\mathbf{u}-\mathbf{v}|)\, d\tau + \varepsilon_n \int\limits_{t_0}^{t} g(\tau, |W_n\mathbf{u} - W_m\mathbf{v}|)\, d\tau +$$

$$+ (\varepsilon_m - \varepsilon_n) \int\limits_{t_0}^{t} \{|\mathbf{f}(\tau, W_m\mathbf{v}) - \mathbf{f}(\tau, \mathbf{x}_0)| + |\mathbf{f}(\tau, \mathbf{v}(\tau)) - \mathbf{f}(\tau, \mathbf{x}_0)|\}\, d\tau.$$

Hier ist das erste Glied an der rechten Seite direkt mit $\varrho(\mathbf{u}, \mathbf{v})$ ausdrückbar; die anderen aber nicht. Wir beschäftigen uns also mit diesen Gliedern. Es sei also $\mathbf{w}_0 \equiv \mathbf{x}_0$. Dann ist

$$|\mathbf{f}(\tau, \mathbf{v}(\tau)) - \mathbf{f}(\tau, \mathbf{x}_0)| \leqq g(\tau, |\mathbf{v}(\tau) - \mathbf{w}_0(\tau)|) \leqq g(\tau, \varrho(\mathbf{v}, \mathbf{w}_0)),$$

d. h.

(47)
$$\int\limits_{t_0}^{t} |\mathbf{f}(\tau, \mathbf{v}(\tau)) - \mathbf{f}(\tau, \mathbf{x}_0)|\, d\tau \leqq \int\limits_{t_0}^{t} g(\tau, \varrho(\mathbf{v}, \mathbf{w}_0))\, d\tau$$

und ebenso

$$\int_{t_0}^{t} |\mathbf{f}(\tau, W_m\mathbf{v}) - \mathbf{f}(\tau, \mathbf{x}_0)|\, d\tau \leqq \int_{t_0}^{t} g\big(\tau, \varrho(W_m\mathbf{v}, \mathbf{w}_0)\big)\, d\tau, \tag{48}$$

da g eine monoton wachsende Funktion ihres zweiten Argumentes ist. Nun ist nach der zweiten Relation von (45)

$$\mathbf{v}_m(t) - \mathbf{x}_0 = \varepsilon_m \int_{t_0}^{t} [\mathbf{f}(\tau, W_m\mathbf{v}) - \mathbf{f}(\tau, \mathbf{x}_0)]\, d\tau - \varepsilon_m \int_{t_0}^{t} [\mathbf{f}(\tau, \mathbf{v}(\tau)) - \mathbf{f}(\tau, \mathbf{x}_0)]\, d\tau + \tag{49}$$

$$+ \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{v}(\tau))\, d\tau.$$

Führt man hier die Bezeichnung $\varrho(\mathbf{v}_m(t), \mathbf{w}_0(t)) = \varrho_m$ ein, so ist nach (43) auch $\varrho(W_m\mathbf{v}, \mathbf{w}_0) \leqq \varrho_m$ gültig. Setzt man diese in (49) ein, so folgt

$$\varrho_m \leqq \varepsilon_m \int_{t_0}^{t} \{g(\tau; \varrho_m) + g\big(\tau, \varrho(\mathbf{v}, \mathbf{w}_0)\big)\}\, d\tau + \int_{t_0}^{t} |\mathbf{f}(\tau, \mathbf{w}_0)|\, d\tau + \int_{t_0}^{t} g\big(\tau, \varrho(\mathbf{v}, \mathbf{w}_0)\big)\, d\tau. \tag{50}$$

Führt man noch die Bezeichnung $\varrho_0 = \varrho(\mathbf{v}, \mathbf{w}_0)$ ein und betrachtet man die Integralgleichung entsprechend der Formel (50):

$$\hat{\varrho}_m = \varepsilon_m \int_{t_0}^{t} \{g(\tau, \hat{\varrho}_m) + g(\tau, \varrho_0)\}\, d\tau + \int_{t_0}^{t} g(\tau, \varrho_0)\, d\tau + \int_{t_0}^{t} |\mathbf{f}(\tau, \mathbf{w}_0)|\, d\tau,$$

so kann man das maximale stetige Integral letzterer Gleichung $\hat{\varrho}_m = h_m(t, \varrho_0(t))$ für die Majorisierung von ϱ_m benützen. Es ist gleich zu sehen, daß $h_m(t, \varrho_0(t))$ eine stetige, monoton wachsende und konvexe bzw. konkave Funktion ihres zweiten Argumentes ist, je nachdem g konvex bzw. konkav im zweiten Argument ist; h_m ist ferner monoton abnehmend mit dem Index m, falls ε_m eine solche Folge ist. Es sei aber bemerkt, daß wegen des letzteren Gliedes im allgemeinen $h_m(t, 0) > 0$ für $t > t_0$ ist. Nach (50) gilt also die Abschätzung

$$\varrho(W_m\mathbf{v}, \mathbf{w}_0) \leqq \varrho_m \leqq h_m(t, \varrho_0)$$

und somit

$$g\big(\tau, \varrho(W_m\mathbf{v}, \mathbf{w}_0)\big) + g(\tau, \varrho_0) \leqq \chi_m\big(\tau, \varrho(\mathbf{v}, \mathbf{w}_0)\big) = g\big(\tau, h_m(\tau, \varrho_0)\big) + g(\tau, \varrho_0), \tag{51}$$

wo χ_m eine monoton wachsende, konvexe bzw. konkave Funktion des zweiten Argumentes ist — ebenso, wie g —, ferner eine monoton fallende Funktion des Indexes, endlich eine meßbare Funktion des ersten Argumentes.

Benützen wir denselben Gedankengang auch für (51), so bekommen wir für $\varrho_{n,m} = \varrho(\boldsymbol{\omega}_n, \mathbf{v}_m)$, bzw. für

$$\varrho(W_n\mathbf{u}, W_m\mathbf{v}) = \varrho\big(\boldsymbol{\omega}_n(t-\vartheta), \mathbf{v}_m(t-\vartheta)\big) \leqq \varrho_{n,m}$$

die Abschätzung

$$\varrho_{n,m}(t) \leqq (1-\varepsilon_n) \int_{t_0}^{t} g\big(\tau, \varrho(\mathbf{u}, \mathbf{v})\big)\, d\tau + \varepsilon_n \int_{t_0}^{t} g(\tau, \varrho_{n,m})\, d\tau + (\varepsilon_m - \varepsilon_n) \int_{t_0}^{t} \chi_m(\tau, \varrho_0)\, d\tau. \tag{52}$$

Setzt man hier 1 statt $1-\varepsilon_n$, ferner statt $\varepsilon_m-\varepsilon_n$, und beachtet man, daß zu g und χ_m eine solche Majorant $\psi_m(\tau, v)$ sich angeben läßt, daß ψ_m auch eine stetige, monoton wachsende und konvexe bzw. konkave Funktion des zweiten Argumentes sei, und

$$g(\tau, \varrho(\mathbf{u}, \mathbf{v})) + \chi_m(\tau, \varrho_0) \leqq \psi_m(\tau, \varrho(\mathbf{u}, \mathbf{v}) + \varrho_0)$$

feststeht, so kann man ϱ_{nm} in (52) mit dem oberen Integral der Gleichung

$$\hat{\varrho}_{n,m} = \varepsilon_n \int_{t_0}^{t} g(\tau, \hat{\varrho}_{n,m})\, d\tau + \int_{t_0}^{t} \psi_m(\tau, \varrho + \varrho_0)\, d\tau, \tag{53}$$

bezeichnet durch $\hat{\varrho}_{n,m} = \gamma_{n,m}(\tau, \varrho(\mathbf{u}, \mathbf{v}) + \varrho(\mathbf{v}, \mathbf{w}_0))$, wo γ eine ebensolche Funktion des zweiten Argumentes wie g ist, ferner monoton fallend mit den Indexen. Setzt man (51) und (53) in (46) ein, so folgt:

$$\varrho(T_n\mathbf{u}, T_m\mathbf{v}) \leqq (1-\varepsilon_n) \int_{t_0}^{t} g(\tau, \varrho(\mathbf{u}, \mathbf{v}))\, d\tau + \varepsilon_n \int_{t_0}^{t} \gamma_{n,m}(\tau, \varrho(\mathbf{u}, \mathbf{v}) + \varrho(\mathbf{v}, \mathbf{w}_0))\, d\tau + \tag{54}$$

$$+ (\varepsilon_m - \varepsilon_n) \int_{t_0}^{t} \chi_m(\tau, \varrho(\mathbf{v}, \mathbf{w}_0))\, d\tau.$$

Es sei aber zugleicht bemerkt, daß man bei der Majorisation von (50) bzw. (52) auch 1 statt ε_m, bzw. statt $1-\varepsilon_n$, ε_n, und ε_m schreiben kann, und dann bekommt man solche obere Integrale x bzw. γ, die unabhängig von m bzw. von n und m sind. Es gilt also auch die Abschätzung

$$\varrho(T_n\mathbf{u}, T_m\mathbf{v}) \leqq (1-\varepsilon_n) \int_{t_0}^{t} g(\tau, \varrho(\mathbf{u}, \mathbf{v}))\, d\tau + \varepsilon_n \int_{t_0}^{t} \gamma(\tau, \varrho(\mathbf{u}, \mathbf{v}) + \varrho(\mathbf{v}, \mathbf{w}_0))\, d\tau + \tag{54*}$$

$$+ (\varepsilon_m - \varepsilon_n) \int_{t_0}^{t} \chi(\tau, \varrho(\mathbf{v}, \mathbf{w}_0))\, d\tau.$$

SATZ 5. *Ist die Funktion $g(t, v)$ eine monoton nichtabnehmende und nichtkonkave stetige Funktion des zweiten Argumentes (und ist $g(\tau, v(\tau))$ auf jeden endlichen Intervall integrierbar, falls $v(\tau) \in C[t_0, \infty)$), strebt ferner ε_n monoton abnehmend nach 0 und genügt die Ungleichung $\varepsilon_n \leqq \frac{1}{2}$ $(n = 1, 2, \ldots)$, so strebt die Folge* (40) *neben einer jeden stetigen Anfangsfunktion $\mathbf{k}(t) = \mathbf{x}_0(t)$; $\mathbf{k}(t_0) = \mathbf{x}_0$ nach der einzigen verallgemeinerten stetigen Lösung des Anfangswertproblems* (13).

BEWEIS: Neben den angegebenen Bedingungen kann man (54*) auch so umschreiben:

$$\varrho(T_n\mathbf{u}, T_m\mathbf{v}) \leqq (1-\varepsilon_n) \int_{t_0}^{t} \{g(\tau, \varrho(\mathbf{u}, \mathbf{v}) + \varrho(\mathbf{v}, \mathbf{w}_0)) + \gamma(\tau, \varrho(\mathbf{u}, \mathbf{v}) + \varrho(\mathbf{v}, \mathbf{w}_0)) +$$

$$+ \chi(\tau, \varrho(\mathbf{u}, \mathbf{v}) + \varrho(\mathbf{v}, \mathbf{w}_0))\}\, d\tau - (1-\varepsilon_m) \int_{t_0}^{t} \{g(\tau, \varrho(\mathbf{v}, \mathbf{w}_0) + \gamma(\tau, \varrho(\mathbf{v}, \mathbf{w}_0)) +$$

$$+ \chi(\tau, \varrho(\mathbf{v}, \mathbf{w}_0))\}\, d\tau,$$

wo man die Abschätzungen $\varepsilon_n \leqq 1-\varepsilon_n$ und $\varepsilon_m-\varepsilon_n=(1-\varepsilon_n)-(1-\varepsilon_m)$ benützte. Man kann also Satz 1—2—3 anwenden, und zwar mit $Q=0$, $P_n=(1-\varepsilon_n)P$; $P\varrho=$

$$=\int_{t_0}^{t}\{g(\tau,\varrho)+\gamma(\tau,\varrho)+\chi(\tau,\varrho)\}d\tau.$$

SATZ 5b. *Ist die Funktion $g(t,v)$ eine solche Funktion, wie in Satz 5, jedoch nichtkonvex, ferner ε_n eine positive Zahlenfolge, für welche $\varepsilon_{n+1}\leqq\frac{1}{2}\varepsilon_n$ $(n=0,1,2,\ldots)$ gültig ist, besitzt endlich die Integralgleichung*

$$\varrho(t)=\int_{t_0}^{t} g(\tau,\varrho(\tau))\,d\tau$$

nur die triviale stetige Lösung, so strebt die Iterationsfolge (40) *nach der eindeutig definierten stetigen Lösung des Anfangswertproblems* (13), *falls nur das Anfangsglied $\mathbf{k}(t)=\mathbf{x}_0(t)$ mit $\mathbf{k}(t_0)=\mathbf{x}_0$ schnell genug in Absolutwert wächst.*

BEWEIS: Unter den angegebenen Bedingungen kann man $\varrho(T_n\mathbf{u}, T_m\mathbf{v})$ nach (54) wie folgt abschätzen:

$$\varrho(T_n\mathbf{u}, T_m\mathbf{v}) \leqq \int_{t_0}^{t} g(\tau,\varrho(\mathbf{u},\mathbf{v}))\,d\tau+$$

$$+\varepsilon_m\int_{t_0}^{t}\{\gamma_{m,m}(\tau,\varrho(\mathbf{u},\mathbf{v})+\varrho(\mathbf{v},\mathbf{w}_0))+\chi_m(\tau,\varrho(\mathbf{u},\mathbf{v})+\varrho(\mathbf{v},\mathbf{w}_0))\}\,d\tau-$$

$$-\varepsilon_n\int_{t_0}^{t}\{\gamma_{n,n}(\tau,\varrho(\mathbf{v},\mathbf{w}_0)+\chi_n(\tau,\varrho(\mathbf{v},\mathbf{w}_0))\}\,d\tau,$$

da $\varepsilon_n\leqq\varepsilon_m-\varepsilon_n$, falls $m>n$ gültig ist. Man kann somit Satz 1d—2d—3d anwenden, und zwar mit

$$Q\varrho=\int_{t_0}^{t} g(\tau,\varrho(\tau))\,d\tau;\quad P_n\varrho=\varepsilon_n\int_{t_0}^{t}\{\gamma_{n,n}(\tau,\varrho(\tau))+\chi_n(\tau,\varrho(\tau))\}\,d\tau.$$

Die Bedingung $\sigma_1<\sigma_0$ kann man durch entsprechende Wahl von $\mathbf{x}_0(t)$ bzw. $\varepsilon_1\leqq\frac{1}{2}\varepsilon_0$ immer sichern; die Bedingung $P_\infty\Theta=\Theta$ ist automatisch erfüllt, da $P_n\to 0$.

Es sei noch bemerkt, dass Satz 5 bzw. 5b auch dann gültig bleibt, wenn man in (40) eine allgemeinere Summe von Picardschen und Carathéodorischen Gliedern betrachtet. Für numerische Zwecke erwies sich als günstigste Iterationsfolge

$$\mathbf{x}_{n+1}(t)=\varepsilon_n\left\{\int_{t_0}^{t}\mathbf{f}(\tau,\mathbf{x}_{n+1}(\tau-\vartheta))\,d\tau+\int_{t_0}^{t}\mathbf{f}(\tau,\mathbf{x}_n(\tau+\vartheta))\,d\tau\right\}+$$

$$+(1-2\varepsilon_n)\int_{t_0}^{t}\mathbf{f}(\tau,\mathbf{x}_n(\tau))\,d\tau,\quad\text{mit}\quad\varepsilon_0=1;\quad\varepsilon_n=\frac{1}{2^n},$$

wo man ϑ so klein wählt, wie es bei den angewandten numerischen Integrationsmethoden bzw. Funktionswert-Auswertungen möglich ist.

4. §. Anwendungen bei Randwertaufgaben

Betrachten wir beispielweise die nichtlineare Randwertaufgabe

$$L[u]=f(\mathbf{x}, u) \quad \text{in} \quad B \tag{55}$$

$$R[u]=\varphi(\mathbf{x}) \quad \text{auf} \quad \Gamma, \tag{56}$$

wo B einen Bereich des n-dimensionalen Raumes X (mit $\mathbf{x}\in X$), ferner Γ den Rand dieses Bereiches bezeichnet, $L[u]$ einen linearen Differentialausdruck, $R[u]$ aber einen ebensolchen Randausdruck bedeutet. Es sei vorausgesetzt, dass zum (55)—(56) entsprechenden linearen Problem eine Greensche Funktion $G(\mathbf{x}, \mathbf{s})$ angegeben ist, mit Hilfe welcher man das Problem $L[u]=r(\mathbf{x})$; $R[u]=\varphi(\mathbf{x})$ in der Form

$$u(\mathbf{x}) = \psi(\mathbf{x})+\int_B G(\mathbf{x}, \mathbf{s})r(\mathbf{s})\, d\omega_{\mathbf{s}} \tag{57}$$

lösen kann, falls r integrierbar ist, wo ψ eine feste, nur von φ abhängende Funktion bedeutet. Falls nun G in $\bar{B}=B\cup\Gamma$ gleichmäßig L-stetig ist, d. h. falls man zu jedem $\varepsilon>0$ ein $\delta(\varepsilon)$ so angeben kann, daß

$$\int_B |G(\mathbf{x}^*, \mathbf{s})-G(\mathbf{x}, \mathbf{s})|\, d\omega_{\mathbf{s}} < \varepsilon$$

immer gültig ist, wenn nur $\mathbf{x}, \mathbf{x}^*\in\bar{B}$ und $|\mathbf{x}-\mathbf{x}^*|<\delta$ feststeht, ferner f in B nach u eine beschränkte Ableitung besitzt, so hat Schröder das Problem (55)—(56) gelöst (s. z. B. [5]). Wir betrachten nun den Fall, wo G wieder gleichmäßig L-stetig in $\bar{B}$ ist, ferner f in u nur die Bedingung

$$|f(\mathbf{x}, u_2)-f(\mathbf{x}, u_1)|\leqq g(\mathbf{x}, |u_2-u_1|) \tag{58}$$

genügt, und g im zweiten Argument stetig, monoton wachsend und konkav (bzw. nicht konvex) ist, ferner für jede $v(\mathbf{x})\in\mathbf{C}(\bar{B})$ $g(\mathbf{x}, v(\mathbf{x}))$ in B integrierbar ist.

Es sei dann $T_n\equiv T$ durch

$$Tv(\mathbf{x}) = \psi(\mathbf{x})+\int_B G(\mathbf{x}, \mathbf{s})f(\mathbf{s}, v(\mathbf{s}))\, d\omega_{\mathbf{s}}, \tag{59}$$

ferner der Pseudoabstand der in $\bar{B}$ stetigen Funktionen durch

$$\varrho(u, v)=|u(\mathbf{s})-v(\mathbf{s})| \tag{60}$$

definiert. Dann gilt

$$\varrho(Tu, Tv)\leqq\int_B |G(\mathbf{x}, \mathbf{s})|\cdot g(\mathbf{s}, \varrho(u, v))\, d\omega_{\mathbf{s}}. \tag{61}$$

Da g im zweiten Argumente konkav ist, so werden wir Satz 1c—2c—3c anwenden, mit $P_n\equiv 0$. Daher setzen wir $g(\mathbf{x}, 0)\equiv 0$ voraus. Somit sind die Voraussetzungen αc—βc —δc und εc erfüllt. Ferner ist $S_n\sigma\equiv S_\infty\sigma$ durch

$$S\sigma = \int_B |G(\mathbf{x}, \mathbf{s})|\cdot g(\mathbf{s}, \sigma(\mathbf{s}))\, d\omega_{\mathbf{s}} \tag{62}$$

definiert. Somit sind also unsere Sätze anwendbar, falls einerseits die Gleichung

$$\sigma = \int_B |G(\mathbf{x}, \mathbf{s})| \cdot g(\mathbf{s}, \sigma(\mathbf{s}))\, d\omega_{\mathbf{s}} \tag{63}$$

für $\sigma \geqq 0$ nur die triviale Lösung $\sigma \equiv 0$ besitzt, ferner falls man ein $u_0 \in C(\overline{B})$ so angeben kann, daß mit $\sigma_0 \geqq \varrho(u_0, u_1)$ auch $\sigma_1 \leqq \sigma_0$ erfüllt ist.

LITERATURVERZEICHNIS

[1] SCHRÖDER, J.: Das Iterationsverfahren bei allgemeinerem Abstandsbegriff, *Math. Zeitschr.* **66** (1956) 111—116.

[2] SCHRÖDER, J.: Anwendung von Fixpunktsätzen bei der numerischen Behandlung nichtlinearer Gleichungen in Halbgeordneten Räumen, *Arch. Rat. Mech. Anal.* **4** (1960) 177—192.

[3] SCHRÖDER, J.: Fehlerabschätzung bei linearen Gleichungssystemen mit dem Brouwerschen Fixpunktsatz, *Arch. Rat. Mech. Anal.* **3** (1959) 28—44.

[4] COLLATZ, L.: *Funktionalanalysis und numerische Mathematik*, Springer, Berlin—Göttingen, Heidelberg. 1964

[5] SCHRÖDER, J.: Nichtlineare Majoranten beim Verfahren der schrittweisen Näherung, *Arch. Math. J.* (1956) 471—484.

RECHENTECHNISCHES ZENTRUM DER UNGARISCHEN AKADEMIE DER WISSENSCHAFTEN, BUDAPEST

(Eingegangen: 12. Juni 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 115—117.

A REMARK CONCERNING THE RATIONAL APPROXIMATION TO $|x|$

by

G. FREUD

Let us denote in the sequel by $\pi_n(x)$ a polynomial of degree n at most, and let us agree, that the $\pi_n(x)$ need not be the same, even if they enter in the same formula, resp. the same equation.

We call an expression of the form

$$\frac{\pi_n(x)}{\pi_n(x)}$$

a rational function of degree n, and the set of all such functions we denote by R_n.

D. NEWMAN [2] constructed a sequence $r_n(x) \in R_n$ with the property

$$(1) \qquad ||x| - r_n(x)| \leqq 3e^{-\sqrt{n-1}} \quad \text{for} \quad x \in [-1, +1], \qquad n = 5, 6, \ldots$$

and he proved also that there does not exist any sequence $r_n^*(x) \in R_n$ for which

$$(2) \qquad ||x| - r_n^*(x)| \leqq \frac{1}{2} e^{-9\sqrt{n}} \qquad x \in [-1, +1]$$

would hold.

This fine and surprising result of D. NEWMAN was developed to a method of approximating certain classes of functions by rational functions. The first result of this kind is due to P. SZÜSZ and P. TURÁN [4], we mention here two of the more recent results:

A) If $f(x)$ is of bounded variation in $[0, 1]$ and belongs to the class $\mathrm{Lip}\,\alpha\,(0 < \alpha < 1)$, then there is a sequence $\varrho_n(f; x) \subset R_n$ so that

$$f(x) - \varrho_n(f; x) = O\left(\frac{\log^2 n}{n}\right);$$

the dependence of the estimate on the Lipschitz-exponent α enters only in the value of the constant of the O-estimate (see G. FREUD [1]).

B) Let us denote by $V^{(r)*}$ the class of functions $f(x)$ with the following properties:

a) $f(x)$ is continuous in $[0, 1]$,

b) There is a subdivision $0 = \xi_0 < \xi_1 < \ldots < \xi_s = 1$ of $[0, 1]$ so that $f(x)$ is $(r-1)$-times continuously differentiable in each $[\xi_k, \xi_{k+1}]$, and $f^{(r-1)}(x)$ is absolutely continuous in $[\xi_k, \xi_{k+1}]$.

c) The function $f^{(r)}(x)$ (defined almost everywhere) is equivalent to a function of bounded variation.

Now, let $f \in V_r^*$, then there exists a sequence $\varrho_n(f; x) \in R_n$ with

$$f(x) - \varrho_n(f; x) = O\left(\frac{\log^2 n}{n^{r+1}}\right)$$

(J. SZABADOS [3]).

Owing to this results, it seems to have some interest to extend NEWMAN's result to weighted approximation on the infinite interval. We hope that this extension could be a starting point of investigations concerning weighted rational approximation on an infinite interval of certain classes of functions.

THEOREM. α) *There is a sequence* $\varrho_n(x) \in R_n$ *for which*

$$\frac{1}{1+x^2}\,\big||x| - \varrho_n(x)\big| \leqq \frac{3}{2}\, e^{-\sqrt{\frac{n-4}{2}}}, \qquad -\infty < x < +\infty \tag{3}$$

is valid:

β) *There does not exist any sequence* $\varrho_n^* \in R_n$, *for which*

$$\frac{1}{1+x^2}\,\big||x| - \varrho_n^*(x)\big| \leqq \frac{1}{4}\, e^{-9\sqrt{n}}, \qquad -\infty < x < \infty. \tag{4}$$

PROOF. Part β): (2) is a consequence of (4), and we know that (2) is impossible.

Part α): We substitute in (1) $\dfrac{2x}{1+x^2}$ for x, v for n and observe, that $\left|\dfrac{2x}{1+x^2}\right| \leqq 1$ for all real x. In this way we obtain

$$\left|\frac{2|x|}{1+x^2} - r_v\left(\frac{2x}{1+x^2}\right)\right| \leqq 3e^{-\sqrt{v-1}}, \qquad -\infty < x < \infty. \tag{5}$$

Using the formula

$$\pi_v\left(\frac{2x}{1+x^2}\right) = \sum_{k=0}^{v} c_k \left(\frac{2x}{1+x^2}\right)^k = \frac{1}{(1+x^2)^v} \sum_{k=0}^{v} 2^k c_k x^k (1+x^2)^{v-k} = (1+x^2)^{-v}\pi_{2v}(x)$$

we have[1]

$$\sigma_v(x) = \frac{1+x^2}{2}\, r_v\left(\frac{2x}{1+x^2}\right) = \left(\frac{1+x^2}{2}\right) \frac{\pi_v\left(\frac{2x}{1+x^2}\right)}{\pi_v\left(\frac{2x}{1+x^2}\right)} =$$

$$= \frac{1+x^2}{2}\, \frac{(1+x^2)^{-v}\pi_{2v}(x)}{(1+x^2)^{-v}\pi_{2v}(x)} = \frac{(1+x^2)\pi_{2v}(x)}{2\pi_{2v}(x)} = \frac{\pi_{2v+2}(x)}{\pi_{2v}(x)} \in R_{2v+2},$$

[1] See the notation we agreed to use at the beginning of the paper.

and from (5)

$$\frac{1}{1+x^2}\,||x|-\sigma_\nu(x)| \leqq \frac{3}{2}\,e^{-\sqrt{\nu-1}}, \qquad -\infty<x<\infty.$$

Finally, we obtain (3), if we put $\nu=\dfrac{n-2}{2}$ for even n and $\nu=\dfrac{n-3}{2}$ for odd n, Q.e.d.

REFERENCES

[1] FREUD, G.: Über die Approximation reeller Funktionen durch rationale Funktionen, *Acta Math. Ac. Sci. Hung.* **17** (1966).
[2] NEWMAN, D.: Rational approximation to $|x|$, *Michigan Math. Journal* **11** (1964) 11—14.
[3] SZABADOS, J.: Generalizations of two theorems of G. Freud concerning the rational approximation, *Studia Sci. Math. Hung.* **1** (1966) (to appear).
[4] SZÜSZ, P. and TURÁN, P.: A konstruktív függvénytan egy újabb irányáról, *MTA III. Oszt. Közleményei* **16** (1966) 33—46.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received June 14, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 119—123.

REMARKS ON THE POISSON PROCESS

by

A. RÉNYI

The inhomogeneous Poisson process on the real line is usually characterized as a stochastic additive set function $\xi(E)$ defined for each bounded Borel subset E of the real line such that

a) the random variable $\xi(E)$ has for each bounded Borel set E a Poisson distribution, i. e.

$$\mathbf{P}(\xi(E) = n) = \frac{[\lambda(E)]^n \cdot e^{-\lambda(E)}}{n!} \qquad (n = 0, 1, \ldots) \tag{1}$$

where $\lambda(E)$ is a nonatomic measure on the real line such that $\lambda(E)$ is finite for each finite interval E, and

b) if $E_1, E_2, \ldots, E_n$ are mutually disjoint bounded Borel sets the random variables $\xi(E_1), \ldots, \xi(E_n)$ are independent.

If we put $\xi_t = \xi([0, t))$ for $t > 0$, $\xi_t = -\xi([t, 0))$ for $t < 0$, this means that ξ_t is a process with independent increments such that $\xi_t - \xi_s$ has a Poisson distribution with mean value $\Lambda(t) - \Lambda(s)$ where $\Lambda(t)$ is the λ-measure of the interval $[0, t)$ if $t > 0$ and $-\Lambda(t)$ is the λ-measure of the interval $[t, 0)$ if $t < 0$. D. SZÁSZ (oral communication) asked the question whether there exists a point process for which a) holds but b) does not hold.

We shall show in this note that such a process does not exist, i. e. the usual supposition about independence in the above characterisation of the Poisson process is unnecessary; by other words supposition b) is a consequence of the supposition a).

More exactly we prove the following

THEOREM 1. *Let J denote the family of all subsets of the real line which can be obtained as the union of a finite number of disjoint finite intervals $[a, b)$ closed to the right and open to the left. Let $\xi(E)$ be an additive stochastic set function defined for each $E \in J$, i. e. such that if E_1 and E_2 are disjoint one has $\xi(E_1 + E_2) = \xi(E_1) + \xi(E_2)$. Suppose that for each $E \in J$ $\xi(E)$ has a Poisson distribution with mean value $\lambda(E)$ where $\lambda(E)$ is a nonatomic measure on the Borel subsets of the real line, which is finite for each $E \in J$. Then it follows that if $E_1, \ldots, E_n$ are disjoint sets $(E_k \in J)$ the random variables $\xi(E_1), \ldots, \xi(E_n)$ are independent, i. e. $\xi(E)$ is a Poisson process.*

PROOF OF THEOREM 1. Let $A(E)$ denote the event $\xi(E) = 0$. If E is the union of the disjoint sets $E_j \in J$ $(j = 1, 2, \ldots, n)$ then[1] clearly $A(E) = A(E_1) \ldots A(E_n)$ because $\xi(E) = \sum_{j=1}^{n} \xi(E_j)$ and thus $\xi(E) = 0$ iff $\xi(E_j) = 0$ for $j = 1, 2, \ldots, n$.

[1] Here and in what follows the product of events denotes the joint occurrence of these events

But by supposition

$$\mathbf{P}(A(E)) = \mathbf{P}(\xi(E) = 0) = e^{-\lambda(E)} = \prod_{j=1}^{n} e^{-\lambda(E_j)} = \prod_{j=1}^{n} \mathbf{P}(A(E_j)). \tag{2}$$

Thus it follows that if the sets $E_1, \ldots, E_n$ are disjoint, the events $A(E_1), \ldots, A(E_n)$ are independent.

Now let $1_{A(E)}$ be the indicator of the event $A(E)$.

Let $E \in J$ and $F \in J$ be two disjoint sets. For any $\varepsilon > 0$ we can clearly decompose E into disjoint intervals E_i $(1 \leqq i \leqq n)$ and F into disjoint intervals F_j $(1 \leqq j \leqq m)$ such that

$$\max_i \lambda(E_i) < \varepsilon \quad \text{and} \quad \max_j \lambda(F_j) < \varepsilon.$$

Now evidently $\xi(E) \neq \sum_{i=1}^{n} 1_{A(E_i)}$ implies $\max_i \xi(E_i) \geqq 2$ and $\xi(F) \neq \sum_{j=1}^{m} 1_{A(F_j)}$ implies $\max_j \xi(F_j) \geqq 2$. On the other hand for any $B \in J$

$$\mathbf{P}(\xi(B) \geqq 2) = \sum_{k=2}^{\infty} \frac{\lambda(B)^k \cdot e^{-\lambda(B)}}{k!} \leqq \lambda^2(B). \tag{3}$$

Thus

$$\mathbf{P}\left(\xi(E) \neq \sum_{i=1}^{n} 1_{A(E_i)}\right) \leqq \sum_{i=1}^{n} \lambda^2(E_i) < \varepsilon\lambda(E) \tag{4a}$$

and

$$\mathbf{P}\left(\xi(F) \neq \sum_{j=1}^{m} 1_{A(F_j)}\right) \leqq \sum_{j=1}^{m} \lambda^2(F_j) < \varepsilon\lambda(F). \tag{4b}$$

This implies, as the sums $\sum_{i=1}^{n} 1_{A(E_1)}$ and $\sum_{j=1}^{m} 1_{A(F_j)}$ are independent that $\xi(E)$ and $\xi(F)$ are independent, too.

As a matter of fact it follows from (4a) and (4b) that for any n and m $(n, m = 0, 1, 2, \ldots)$

$$|\mathbf{P}(\xi(E) = n, \xi(F) = m) - \mathbf{P}(\xi(E) = n) \cdot \mathbf{P}(\xi(F) = m)| \leqq 2\varepsilon\lambda(E + F). \tag{5}$$

As $\varepsilon > 0$ can be chosen arbitrarily small, our statement follows. The independence of the variables $\xi(E_i)$ $(i = 1, 2, \ldots r)$ with disjoint E_i and $r > 2$ is proved in exactly the same way. Thus our theorem is proved.

REMARK 1. Note that to prove the independence of $\xi(E_i)$ $(i = 1, 2, \ldots, r)$ for $E_i E_j = \varnothing$ if $i \neq j$ we have not used the full supposition that for each $E \in J$ $\xi(E)$ has a Poisson distribution, only that

$$\mathbf{P}(\xi(E) = 0) = e^{-\lambda(E)} \tag{6a}$$

and

$$\mathbf{P}(\xi(E) \geqq 2) = o(\lambda(E)) \quad \text{if} \quad \lambda(E) \to 0 \tag{6b}$$

uniformly in E.

Thus even these suppositions imply that the process $\xi(E)$ is a process of independent increments. It is easy to show however that this together with (6a) and (6b) implies that $\xi(E)$ has a Poisson distribution.

Thus the following theorem is true.

THEOREM 2. *Let J denote the family of all subsets of the real line which can be obtained as the union of a finite number of disjoint finite intervals $[a, b)$. Let $\xi(E)$ be an additive stochastic set function defined for $E \in J$, i. e. such that if $E_1 \in J$ and $E_2 \in J$ are disjoint one has $\xi(E_1 + E_2) = \xi(E_1) + \xi(E_2)$. Suppose that $\xi(E)$ is for each $E \in J$ a nonnegative integer valued random variable such that*

$$\mathbf{P}\big(\xi(E) = 0\big) = e^{-\lambda(E)} \tag{7a}$$

and

$$\mathbf{P}\big(\xi(E) \geqq 2\big) \leqq \lambda(E) \cdot \delta\big(\lambda(E)\big) \tag{7b}$$

where $\delta(x)$ is an increasing positive function defined for $x > 0$ such that $\lim_{x \to 0} \delta(x) = 0$ and $\lambda(E)$ a nonatomic measure on J. Then it follows that $\xi(E)$ is a Poisson process, i. e. if E_i $(i = 1, 2, \ldots, r)$ are disjoint sets, $E_i \in J$ the random variables $\xi(E_i)$ $(i = 1, 2, \ldots, r)$ are independent, and (1) holds.

PROOF OF THEOREM 2. Put for $E \in J$

$$\varphi_E(u) = \mathbf{M}(e^{iu\xi(E)}) \qquad (-\infty < u < +\infty)$$

then clearly

$$|\varphi_E(u)| \geqq e^{-\lambda(E)} - (1 - e^{-\lambda(E)}) > 0 \tag{8}$$

if $\lambda(E) < \log 2$. Thus if $E = \sum_{i=1}^{r} E_i$, where $E_i \in I$ and $E_i E_j = \varnothing$ if $i \neq j$, then if $\lambda(E_i) < \log 2$ we have (as it follows from the proof of Theorem 1 that the random variables $\xi(E_i)$ $(i = 1, 2, \ldots, r)$ are independent)

$$\varphi_E(u) = \prod_{i=1}^{r} \varphi_{E_i}(u) \neq 0 \tag{9}$$

and therefore

$$\log \varphi_E(u) = \sum_{i=1}^{r} \log \varphi_{E_i}(u). \tag{10}$$

As however

$$\varphi_{E_i}(u) = e^{-\lambda(E_i)} + e^{iu}(1 - e^{-\lambda(E_i)}) + O\big(\lambda(E_i)\delta\big(\lambda(E_i)\big)\big) \tag{11}$$

we get

$$\log \varphi_{E_i}(u) = \lambda(E_i)(e^{iu} - 1) + O\big(\lambda(E_i)\big(\lambda(E_i) + \delta\big(\lambda(E_i)\big)\big)\big). \tag{12}$$

It follows that if $\lambda(E_i) < \varepsilon$ for $i = 1, 2, \ldots, r$

$$\log \varphi_E(u) = \lambda(E)(e^{iu} - 1) + O\big(\varepsilon + \delta(\varepsilon)\big) \tag{13}$$

that is, as $\varepsilon>0$ can be chosen arbitrarily small,

$$\varphi_E(u) = e^{\lambda(E)(e^{iu}-1)} \tag{14}$$

which implies that $\xi(E)$ has a Poisson distribution with mean $\lambda(E)$. Thus Theorem 2 follows from Theorem 1.

REMARK 2. The proof can be carried over without any change to the discussion of a Poisson process in more than one dimension or even in an abstract space. Thus we obtain the following

THEOREM 3. *Let X be any space, J a family of subsets of X and $\lambda(E)$ a non-negative finite valued set function defined on J such that*

1) *if $E_1 \in I$, $E_2 \in J$ and $E_1 E_2 = \varnothing$, then $E_1 + E_2 \in J$,*
2) *If $E_1 \in I$, $E_2 \in J$, $E_1 E_2 = \varnothing$ then $\lambda(E_1 + E_2) = \lambda(E_1) + \lambda(E_2)$,*
3) *There is a constant α with $0<\alpha<1$ such that for every $E \in J$ with $\lambda(E)>0$ there exists a subset F of E such that $F \in J$, $E-F \in J$ and $\alpha < \dfrac{\lambda(F)}{\lambda(E)} < 1-\alpha$. Let us suppose that a stochastic set function is defined on J i. e. to every $E \in J$ there corresponds a random variable $\xi(E)$ such that if $E_1 \in J$, $E_2 \in J$ and $E_1 E_2 = \varnothing$ we have $\xi(E_1 + E_2) = \xi(E_1) + \xi(E_2)$ and $\xi(E)$ has a Poisson distribution with mean value $\lambda(E)$.*

Then the random variables $\xi(E_i)$ $(i=1, 2, \ldots, r)$ are independent if the sets $E_i \in J$ $(i=1, \ldots, r)$ are disjoint, i. e. $\xi(E)$ is a Poisson process.

Note that condition 3) is not quite the same as that λ is nonatomic, because we did not suppose that J is a σ-algebra of sets.

REMARK 3. The question arises whether the condition that the process should be one with independent increments can be deduced from other suppositions for other processes of independent increments, too.

The most interesting case is that of the Wiener process. For this process one has the following (almost trivial) analogue of Theorem 1.

THEOREM 4. *Let ξ_t $(-\infty<t<+\infty)$ be a stochastic process such that $\xi_t - \xi_s$ is normally distributed with mean 0 and variance $(t-s)$ for $s<t$. Suppose further that if the intervals $[s_j, t_j)$ $(j=1, 2, \ldots, r)$ are disjoint, any linear combination $\sum_{j=1}^{r} b_j(\xi_{t_j} - \xi_{s_j})$ of the increments $\xi_{t_j} - \xi_{s_j}$ with real coefficients b_j is normally distributed. Then $\{\xi_t\}$ is the Wiener process, i. e. the random variables $\xi_{t_j} - \xi_{s_j}$ are independent if the intervals $[s_j, t_j)$ are disjoint.*

PROOF OF THEOREM 4. Clearly putting for $I_k = [s_k, t_k)$ $\xi(I_k) = \xi_{t_k} - \xi_{s_k}$ $(k=1, 2)$ if I_1 and I_2 are adjacent intervals $(s_1 < t_1 = s_2 < t_2)$ $\xi(I_1 + I_2) = \xi(I_1) + \xi(I_2)$ and thus

$$\mathbf{M}\big((\xi(I_1+I_2))^2\big) = t_2 - s_1 = t_2 - s_2 + t_1 - s_1 = \mathbf{M}\big(\xi^2(I_1)\big) + \mathbf{M}\big(\xi^2(I_2)\big)$$

and thus $\mathbf{M}\big(\xi(I_1)\xi(I_2)\big) = 0$, i. e. $\xi(I_1)$ and $\xi(I_2)$ are uncorrelated. Now let I_1 and I_2 be arbitrary disjoint intervals

$$I_1 = [s_1, t_1),\ I_2 = [s_2, t_2) \quad \text{where} \quad s_1 < t_1 < s_2 < t_2$$

and put $I_3=[t_1, s_2)$. Then, taking into account that $\mathbf{M}(\xi(I_1)\xi(I_3))=0$ and $\mathbf{M}(\xi(I_3)\xi(I_2))=0$, we get $\mathbf{M}(\xi^2(I_1+I_2+I_3)) = t_2-s_1=\mathbf{M}(\xi^2(I_1))+\mathbf{M}(\xi^2(I_2))+$ $+\mathbf{M}(\xi^2(I_3)+2\mathbf{M}(\xi(I_1)\xi(I_2))$.

Thus

$$\mathbf{M}(\xi(I_1)\xi(I_2))=0.$$

(We have used here the following elementary geometrical fact: if a, b, c are vectors in the 3-dimensional Euclidean space for which c is orthogonal both to b and to $a+b$, then c is orthogonal to a, too.) Thus $\xi(I_1)$ and $\xi(I_2)$ are uncorrelated if I_1, and I_2 are arbitrary disjoint intervals.

It follows that if $I_1, I_2, \ldots, I_r$ are disjoint intervals and I_j has length $|I_j|$, and $b_1, \ldots, b_r$ are arbitrary real constants, then

$$\mathbf{M}\left(\left(\sum_{j=1}^{r} b_j\xi(I_j)\right)^2\right) = \sum_{j=1}^{r} b_j^2 |I_j|.$$

Thus

$$\mathbf{M}\left(e^{iu\sum_{j=1}^{r} b_j\xi(I_j)}\right) = e^{-\frac{1}{2}u^2\sum_{j=1}^{r} b_j^2 |I_j|}$$

and thus for any real numbers $u_1, u_2, \ldots, u_r$

$$\mathbf{M}\left(e^{i\sum_{j=1}^{r} u_j\xi(I_j)}\right) = \prod_{j=1}^{r} \mathbf{M}(e^{iu_j\xi(I_j)})$$

i. e. the $\xi(I_j)$ $(j=1, 2, \ldots, r)$ are independent i.e. ξ_t is the Wiener process.

REMARK 4. Returning to the Poisson process, the question arises whether if in Theorem 1 instead of the condition that $\xi(E)$ has a Poisson distribution if E is any finite union of intervals, one supposes only that $\xi(I)$ has a Poisson distribution if I is any interval, does this ensure that the process is a Poisson process? We can prove only that in this case $\xi(I_1)$ and $\xi(I_2)$ are uncorrelated if I_1 and I_2 are disjoint intervals. The proof of this is essentially the same as the first step of the proof of Theorem 4.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received June 14, 1966.)

Remark added on August 22, 1966.

I have been informed by JAY GOLDMAN that the answer to the question in Remark 4 is: no. This has been shown by a counterexample by L. SHEPP; his example will be published in a forthcoming paper of J. GOLDMAN.

Remark added on March 15, 1967.

P. A. P. MORAN has obtained independently from L. SHEPP the same results. His paper will be published in this journal.

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 125—132.

ON RANDOM DETERMINANTS I

by

A. PRÉKOPA

1. Introduction

Consider the determinant with random entries

$$\Delta_n = \begin{vmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1n} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \xi_{n1} & \xi_{n2} & \cdots & \xi_{nn} \end{vmatrix} \tag{1. 1}$$

where we suppose that the random variables ξ_{ik}, $i, k = 1, \ldots, n$ are independent and identically distributed. (The reader will observe that certain conditions can be weakened without violating the validity of our subsequent statements.) We shall assume later on the existenceof the moments of the ξ_{ik}'s of order as high as it will be necessary. We are now interested in finding the moments $\mathbf{E}(\Delta_n^{2k})$, $k = 1, 2, \ldots$. The odd order moments of Δ_n are clearly equal to 0 as Δ_n has a symmetrical distribution with respect to 0. In fact interchanging two rows in Δ_n we obtain $-\Delta_n$ and this latter has the same probability distribution as Δ_n. Suppose that $\mathbf{E}(\xi_{ik}) = 0$, $\mathbf{E}(\xi_{ik}^2) = 1$, $i, k = 1, \ldots, n$. Then it is well known that[1] (first remarked in [4])

$$\mathbf{D}^2(\Delta_n) = \mathbf{E}(\Delta_n^2) = n!. \tag{1. 2}$$

The fourth moment of Δ_n was obtained by NYQUIST, RICE and RIORDAN [6] and the formula is the following

$$\mathbf{E}(\Delta_n^4) = \frac{(n!)^2}{2} \sum_{i=0}^{n} \frac{(n-i+1)(n-i+2)}{i!} (m_4 - 3)^i, \tag{1. 3}$$

where

$$m_4 = \mathbf{E}(\xi_{ik}^4), \; i, k = 1, \ldots, n.$$

In an earlier paper TURÁN and SZEKERES [1] (see also [2], [3]) investigated the sum of squares and the sum of the fourth powers of all determinants with entries $-1, 1$. Applying non-probabilistic arguments they obtained the formula (1. 2) for the arithmetic mean of all squares and a recursion formula for the arithmetic mean of the fourth powers. This recursion formula was not solved, however, but it is a special case of the recursion formula proved later in [6] for $\mathbf{E}(\Delta_n^4)$ which lead to (1. 3). We can therefore obtain from (1. 3) the explicit formula for the arithmetic

[1] $\mathbf{E}(\xi)$ denotes the expectation and $D(\xi)$ the dispersion of the random variable ξ.

mean of all fourth powers of the determinants with entries $-1, 1$ if we substitute $m_4 = 1$ in (1. 3).

There is only one type of probability distributions as the distribution of the ξ_{ik}'s for which all moments of Δ_n are known and this is the standard normal distribution. In this case we have

$$\mathbf{E}(\Delta_n^{2k}) = n! \frac{(n+2)!}{2!} \frac{(n+4)!}{4!} \cdots \frac{(n+2k-2)!}{(2k-2)!}. \tag{1. 4}$$

This result can be obtained in a well known way from the WISHART distribution (see e. g. [7]). Other proof is published in [6]. In a summary of a lecture FORSYTH and TUKEY [5] gave without proof a formula for the $2k$-th moment of the content of n random unit vectors uniformly distributed on the surface of the unit sphere in the n-dimensional space. Formula (1. 4) can simply be obtained from this and vice versa. The proof was never published. We shall give a direct proof for that without using any deeper tools as this case seems to be of particular interest.

2. Reformulation of the Problem. Moments of Permanents. New Proof of an Earlier Result

Together with the random determinant

$$\Delta_n = \sum_{(i_1, i_2, \ldots, i_n)} \pm \xi_{1i_1} \xi_{2i_2} \cdots \xi_{ni_n}$$

we shall investigate the random permanent

$$P_n = \sum_{(i_1, i_2, \ldots, i_n)} \xi_{1i_1} \xi_{2i_2} \cdots \xi_{ni_n}$$

of the same random matrix. We assume that the random variables ξ_{ij} are symmetrically distributed with respect to 0. Let us introduce the notation $m_{2k} = \mathbf{E}(\xi_{ij}^{2k})$. The problem of finding the $2k$-th moment of the random variable P_n can be reformulated in the following way. Consider all tables of $2k$ rows and n columns one row of which consists of a permutation of the elements $1, 2, \ldots, n$. The number of all such tables is $(n!)^{2k}$. A table is called regular if every number in every column has an even multiplicity. We assign a weight to each column and define the weight of a regular table as the product of the weights of the columns. The weight of a column is defined as

$$m_2^{j_1} m_4^{j_2} \ldots m_{2k}^{j_k}, \quad m_2 = 1$$

where $2j_1 + 4j_2 + \ldots + 2kj_k = 2k$ and j_1 is the number of different numbers with multiplicity 2, j_2 is the number of different numbers with multiplicity 4 in that column etc. If at least one column contains a number with an odd multiplicity then the weight of the table is 0 by definition. The sum of the weights of the tables is equal to $\mathbf{E}(P_n^{2k})$.

Let us now give a positive or a negative sign to each table according that the sum of inversions contained in the different rows is an even or an odd number. The sum of the signed weights is equal to $\mathbf{E}(\Delta_n^{2k})$.

The above assertions follow immediately from the definition of the permanent and the determinant taking into account the independence of the random variables

ξ_{ik}, $i, k = 1, \ldots, n$. We observe that P_n has also a symmetrical distribution with respect to 0. Now we prove the following

THEOREM **1.** $\mathbf{E}(P_n^2) = \mathbf{E}(\Delta_n^2)$, $\mathbf{E}(P_n^4) = \mathbf{E}(\Delta_n^4)$, $n = 1, 2, \ldots,$

$$\mathbf{E}(P_n^{2k}) = \mathbf{E}\Delta_n^{2k}), \quad \textit{for} \quad n = 1, 2;\ k = 1, 2, \ldots,$$

but if $\mathbf{P}(\xi_{ij} = 0) \neq 1$ *then*

$$\mathbf{E}(P_n^{2k}) \neq \mathbf{E}(\Delta_n^{2k}) \quad \textit{for} \quad k \geqq 3;\ n \geqq 3.$$

PROOF. The validity of the first equality is trivial. When proving the second one we give at the same time a proof for the formula (1. 3). We shall make use of the above reformulation of the moment-problem and consider all $4 \times n$ tables where each number in each column has an even multiplicity. Any multiplicity can be now just either 2 or 4. We may fix the permutation of the first row as 1 2 3 ... n and at the end multiply the result by $n!$

Consider together the first and the second rows:

$$1\ 2\ 3\ \ldots\ n,$$

$$j_1\, j_2\, j_3\, \cdots\, j_n.$$

This is conceivable as one permutation. Let $i_1, i_2, \ldots, i_n$ denote the number of cycles of lengths $1, 2, \ldots, n$, respectively. The $4 \times n$ table is regular if and only if in the third and fourth rows below each cycle with the same numbers in the same ordering is repeated what stands in the first and second rows but there are two possibilities. Below from the considered cycle the third row may contain the above standing part i. e. the first row while the fourth row contains the corresponding part of the second row or conversely.

These two possibilities can be used independently of each other below each cycle of the permutation defined by the first two rows. To illustrate the situation consider the first three numbers of the first and second rows, and suppose that they form the following cycle:

$$\begin{matrix} 1 & 2 & 3 \\ 2 & 3 & 1. \end{matrix}$$

Now in the first column we must have one more 1 and one more 2 to obtain a regular table. This can be done so that 1 stands in the third row and 2 in the fourths or conversely. This choice uniquely determines the other two elements in the third row and also in the fourth row. Therefore we have

$$\text{either} \quad \begin{matrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 1 & 2 & 3 \\ 2 & 3 & 1 \end{matrix} \qquad \text{or} \quad \begin{matrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 2 & 3 & 1 \\ 1 & 2 & 3. \end{matrix}$$

Having this structure of the $4 \times n$ regular tables we show that every such table has a positive sign. Applying a permutation for the n columns so that elements in the cycles in the first two rows be connected and stand after each other, a regular

table keeps the weight and the sign. In this new table from the point of view of the weight and sign it is immaterial whether the elements inside a cycle in the second row are repeated in the third or in the fourth row. In fact a cycle of an odd length has an even number of transpositions and a cycle of an even length has an odd number of transpositions therefore the internal number of transpositions remains the same. The external number of transpositions also remains the same thus the new table is not sensitive for such a change from the point of view of signed weight. But if the whole first row is placed in the third and the whole second row is placed in the fourth row then the table is clearly positive. This proves the second assertion of the theorem.

To prove that $\mathbf{E}(P_n^{2k}) \neq \mathbf{E}(\Delta_n^{2k})$ if $m_2 \neq 0$, $n \geqq 3$, $k \geqq 3$, it is enough to show that there are tables with negative weights. If $n = 3$ and $k = 3$ then the signed weight of the table

$$\begin{array}{ccc} 1 & 2 & 3 \\ 1 & 3 & 2 \\ 2 & 1 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \\ 3 & 2 & 1 \end{array}$$

is $-m_2^9 < 0$. For arbitrary $n \geqq 3$ and $k \geqq 3$ it sufficies to supply the above table by adding 4 5 6 ... n to each row in the same permutation and adding as many new rows as it is necessary containing the same permutations. The obtained table is surely negative. Finally it is easy to see that

$$\mathbf{E}(P_n^{2k}) = \mathbf{E}(\Delta_n^{2k}) \quad \text{for} \quad n = 1, 2;\ k = 1, 2, \ldots .$$

To derive formula (1. 3) we remark that the columns of a regular table can be subdivided into three categories. The first category contains columns which have four times the same number. The second one contains columns which have the same numbers in the first and second rows also in the third and fourth rows but these numbers are different. The remaining columns belong to cycles of length at least 2 of the first two rows and these columns form the third category. In order to obtain the number of tables (which is the sum of weights in this case) all columns of which belong to the third category we mention that if $d(n, k)$ is the number of permutations consisting of k cycles with lengths at least 2 then it is well known that (see e. g. [8])

$$d_n(t) = \sum_{k=0}^{n} d(n, k) t^k = \sum_{k=0}^{n} \binom{n}{k} t(t+1)\ldots(t+n-k-1)(-t)^k.$$

As each cycle can be repeated either in the third or in the fourth row, $d_n(2)$ gives the number of tables having columns belonging just into the third category.

If a table consists of columns of the second kind then the first two rows are identical and also the third and fourth rows. As there is no column containing four times the same number, the number of all $4 \times n$ tables is the sum of weights

and it is $d_n(1)$. Thus

$$\mathbf{E}(\Delta_n^4) = n! \sum_{i_1+i_2+i_3=n} \frac{n!}{i_1!\, i_2!\, i_3!} d_{i_1}(1)\, d_{i_2}(2) m_4^{i_3} =$$

$$= \frac{(n!)^2}{2} \sum_{k=0}^{n} \frac{(n-k+1)(n-k+2)}{k!} (m_4-3)^k.$$

We remark that the numbers $d(n, k)$ are the associated Stirling numbers of the first kind.

3. The Case of the Standard Normal Distribution

We suppose that the ξ_{ij}'s in (1. 1) have standard normal distribution and give a proof for the formula (1. 4). For this purpose first we prove the following

LEMMA. *Let $\zeta_1, \ldots, \zeta_k$ be n-dimensional independent random vectors with independent components having standard normal distribution. The k-dimensional content of the parallelotope determined by these vectors is the product of two independent random variables one of which has a χ-distribution with $n-k+1$ degrees of freedom and the other is distributed as the $k-1$-dimensional content of $k-1$ independent random vectors having independent and standard normally distributed components.*

PROOF. Let $\Delta_n^{(k)}$ denote the content of the k random vectors. Then

$$\Delta_n^{(k)} = \alpha_k \Delta_n^{(k-1)}$$

where $\Delta_n^{(k-1)}$ is the $k-1$-dimensional content of the parallelotope determined by $\xi_1, \ldots, \xi_{k-1}$ and α_k is the distance of ζ_k from the subspace spanned by $\zeta_1, \ldots, \zeta_{k-1}$. In view of the spherical symmetry of the distribution of ζ_i, α_k and $\Delta_n^{(k-1)}$ are independent of each other. α_k is clearly a χ-variable with $n-k+1$-degrees of freedom as the subspace of the first $k-1$ vectors can be fixed as the set of those points $(x_1, x_2, \ldots, x_n)$ for which $x_k = x_{k+1} = \ldots = x_n = 0$. This completes the proof.

THEOREM 2. *If the ξ_{ij}'s have standard normal distribution then the random variable* (1. 1) *can be written as the product of n independent χ-variables:*

$$\Delta_n = \chi_1 \chi_2 \cdots \chi_n$$

where α_k has $n-k+1$ degrees of freedom.

PROOF. The theorem follows from a subsequent application of the idea of the proof in the preceding Lemma.

As the k-th moment of a χ^2-variable with i-degrees of freedom is equal to

$$(i+2k-2)(i+2k-4) \ldots (i+2)i,$$

it follows that

$$\mathbf{E}(\Delta_n^{2k}) = \prod_{i=1}^{n} (i+2k-2)(i+2k-4)\ldots(i+2)i = n! \frac{(n+2)!}{2!} \frac{(n+4)!}{4!} \cdots \frac{(n+2k-2)!}{(2k-2)!}$$

which proves (1. 4). We remark that the moments of the content of n random vectors uniformly distributed on the surface of the unit sphere in the n-dimensional space can be obtained from this because

$$\Delta_n = \chi_1 \chi_2 \cdots \chi_n \begin{vmatrix} \frac{\xi_{11}}{\chi_1} & \cdots & \frac{\xi_{1n}}{\chi_n} \\ \cdots\cdots & \cdots & \cdots\cdots \\ \frac{\xi_{n1}}{\chi_1} & \cdots & \frac{\xi_{nn}}{\chi_n} \end{vmatrix}$$

where

$$\chi_i = \sqrt{\xi_{i1}^2 + \ldots + \xi_{in}^2}, \qquad i = 1, \ldots, n$$

and the $n+1$ factors in the product as well as the rows of the determinant are independent.

4. Polynomials Associated with Random Determinants, Generalization of the Formula (1.3)

Let us define the polynomials $f_n(m_1, m_2, \ldots, m_k)$, $k, n = 1, 2, \ldots$ as the sum of signed weights of all $k \times n$ tables where in each row we write one permutation of the numbers $1, 2, \ldots, n$, the weight of a table is the product of the weights of the columns and the weight of a column is $m_1^{i_1} m_2^{i_2} \ldots m_k^{i_k}$ where i_j is the number of different numbers with multiplicity j in the column. The sign is the total sum of the transpositions in the k rows. The non-signed sum of weights will be denoted by $g_n(m_1, m_2, \ldots, m_k)$. The variables $m_1, m_2, \ldots, m_k$ can be real or complex. Considering a random determinant (1. 1) where the random entries are independent, identically (but not necessarily symmetrically) distributed having finite moments up to order k, and these moments are $m_1, m_2, \ldots, m_k$, then

$$f_n(m_1, m_2, \ldots, m_k) = \mathbf{E}(\Delta_n^k), \tag{4. 1}$$

while

$$g_n(m_1, m_2, \ldots, m_k) = \mathbf{E}(P_n^k). \tag{4. 2}$$

As Δ_n has a symmetrical distribution with respect to 0, $f_n(m_1, m_2, \ldots, m_k)$ vanishes if $m_1, m_2, \ldots, m_k$ are moments of a probability distribution and k is odd. This implies that $f_n(m_1, m_2, \ldots, m_k)$ vanishes for all values of the variables $m_1, m_2, \ldots, m_k$ if k is an odd number. The same holds for g_n if the entries have symmetrical distribution with respect to 0. The polynomials f_n, g_n will be called polynomials associated with random determinants, random permanents, respectively. Both f_n and g_n are clearly homogeneous polynomials of their variables. We mention also the following

THEOREM 3. For fixed k and $m_1, m_2, \ldots, m_{k-1}$, the polynomials $g_n/n!$ are Appel polynomials of the variable m_k. The same holds for $f_n/n!$ if k is an even number.

PROOF. Note that polynomials $y_1(x), y_2(x), \ldots$ are called Appel polynomials if $y_n'(x) = n y_{n-1}(x)$, $n = 1, 2, \ldots$. To prove this property of the above polynomials

consider the $k \times n$ tables. Each table contains a certain number of columns consisting of k times the same number. If the number of such columns is j then they can be selected in $\binom{n}{j}$ different ways. Thus g_n has the form

$$g_n(m_1, m_2, \ldots, m_k) = n! \sum_{j=0}^{n} \binom{n}{j} m_k^j \, d_{n-j}(m_1, \ldots, m_{k-1}). \tag{4. 3}$$

f_n has a similar form but we have to remark that if k is even then any particular choice of the j columns consisting of k times the same numbers the remaining columns form a $k \times (n-j)$ table of the same sign. Thus

$$f_n(m_1, m_2, \ldots, m_k) = n! \sum_{j=0}^{n} \binom{n}{j} m_k^j c_{n-j}(m_1, \ldots, m_{k-1}). \tag{4. 4}$$

Our assertions follow immediately from (4. 3) and (4. 4).

If the random variables ξ_{ij} in (1. 1) have a symmetrical distribution then $m_1 = m_3 = m_5 = \ldots = 0$. If moreover we take into account that $m_2 = 1$ then we have polynomials $g_n(m_4, \ldots, m_{2k}), f_n(m_4, \ldots, m_{2k})$. Now we generalize the formula (1. 3) and express it in

THEOREM 4. *If the random variables ξ_{ij} in* (1. 1) *have a symmetrical distribution and this has a finite moment of order $2k$ moreover the moments of order $2, 4, \ldots, 2k-2$ are the same as those of the standard normal distribution,*

$$m_{2j} = \frac{(2j)!}{j!\, 2^j}, \qquad j = 1, 2, \ldots, k-1 \tag{4. 5}$$

while m_{2k} is arbitrary, then

$$\mathbf{E}(\Delta_n^{2k}) = (n!)^2 \sum_{j=0}^{n} \frac{1}{j!} \left(m_{2k} - \frac{(2k)!}{k!\, 2^k} \right)^j \frac{M_{n-j}^{(2k)}}{[(n-j)!]^2} \tag{4. 6}$$

where $M_n^{(2k)}$ stands for the $2k$-th moment of Δ_n the entries of which have the standard normal distribution, i.e $M_n^{(2k)}$ is given by (1. 4).

PROOF. From Theorem 3 we know that

$$\frac{d}{dm_{2k}} \frac{\mathbf{E}(\Delta_n^{2k})}{n!} = n \frac{\mathbf{E}(\Delta_{n-1}^{(2k)})}{(n-1)!} \tag{4. 7}$$

where we have the initial conditions

$$\mathbf{E}(\Delta_n^{2k}) = M_n^{(2k)} \quad \text{for} \quad m_{2k} = \frac{(2k)!}{2^k k!}. \tag{4. 8}$$

The sequence of polynomials $\mathbf{E}(\Delta_1^{2k}), \mathbf{E}(\Delta_2^{2k}), \ldots$ is uniquely determined by (4. 7) and (4. 8). But (4. 6) satisfies these conditions hence our theorem is proved.

REFERENCES

[1] SZEKERES, G. and TURÁN P.: Egy szélsőértékfeladat a determináns elméletben (On an extremal problem in the theory of determinants, in Hungarian with German abstract), *Matematikai és Term. tud. Értesítő* **56** (1937) 796—804.
[2] TURÁN, P.: Determinánsokra vonatkozó szélsőértékfeladatok (On extremal problems concerning determinants, in Hungarian with English abstract), *Matematikai és Term. tud. Értesítő* **59** (1940) 95—105.
[3] TURÁN, P.: On a problem in the theory of determinants, *Acta Math. Sinica* **5** (1955) 417—423.
[4] FORTET, R.: Random determinants, *J. Research, Nat. Bur. Standards* **17** (1951) 465—470.
[5] FORSYTH, G. E. and TUKEY, J. W.: The extent of n random unit vectors (Abstract), *Bull. Amer. Math. Soc.* **58** (1952) 502.
[6] NYQUIST, H., RICE, S. O., and RIORDAN, J.: The distribution of random determinants, *Quart. Applied Math.* **12** (1954) 97—104.
[7] WILKS, S. S.: *Mathematical Statistics*, Wiley, New York, London, 1962.
[8] RIORDAN, J.: *An Introduction to Combinatorial Analysis*, Wiley, New York, London, 1958.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received July 20, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 133—151.

GENERALIZATION OF LINNIK'S ASYMPTOTIC FORMULA FOR THE ADDITIVE PROBLEM OF DIVISORS TO GAUSSIAN NUMBERS

by

J. PERGEL

Let us denote by $\tau_k(z)$ the number of ways of writing the Gaussian number z in the form $z=\xi_1\,\xi_2 \ldots \xi_k$, where $\xi_1, \ldots, \xi_k$ are Gaussian numbers. $\tau_2(z)=\tau(z)$ is the number of Gaussian divisors of the Gaussian number z.

Let Ω be a star region, such that if its boundary is represented by the function $|\varrho|=F(\arg \varrho)$ where $F(\cdot)$ is a mod 2π periodic function. Let us suppose that $F(\cdot)$ is absolutely continuous. If α is any complex number let us denote by $\alpha\Omega$ the set of complex numbers $\{\alpha z;\ z\in\Omega\}$.

We want to get asymptotic formula for the expression

$$\sum_{z\in x\Omega} \tau(z+l)\tau_k(z) \qquad (x>0, \quad x\to\infty). \tag{1}$$

Where l is a fixed Gaussian number, the z are the Gaussian numbers of $x\Omega$, and $\tau(0)=$ $=\tau_k(0)=0$.

If $f(x_1, \ldots, x_k)$ is any function of Gaussian numbers, then let us denote by $N\{f(x_1, \ldots, x_n)=0,\ x_1, \ldots, x_n\in H\}$ the number of solutions of the equation $f(x_1, \ldots, x_n)=0$ in Gaussian integers $x_1, \ldots, x_n$, such that $\{x_1, \ldots, x_n\}\in H$ where H is any set of n-tuples of complex numbers. Then

$$\sum_{z\in x\Omega} \tau(z+l)\tau_k(z) = N\{\xi\eta-\vartheta_1 \ldots \vartheta_k = l,\ \vartheta_1 \ldots \vartheta_k \in x\Omega\}. \tag{2}$$

In this paper we use both of the forms of (1).

The analogon of this problem for natural numbers was solved by LINNIK [1] and BREDIKHIN [2] with the aid of LINNIK's "dispersion method". Dealing with (1) we also use a version of this "dispersion method" for Gaussian numbers. We deduce the following asymptotic formula for (1)

$$\sum_{z\in x\Omega} \tau(z+l)\tau_k(z) = \frac{A_k B_k(l)}{(k-1)!}|\Omega| x^2 (\log x)^k + O\left(x^2 (\log x)^{k-1} (\log\log x)\right)^c \tag{3}$$

where $|\Omega|$ is the area of Ω and c is a suitable constant,

$$A_k = \sum_{q\neq 0} \frac{\mu(q)}{(N(q))^2} \prod_{p|q} N(p)\left\{1-\left(1-\frac{1}{N(p)}\right)^{k-1}\right\} \tag{4}$$

$$B_k(l) = \prod_{p|l}\left\{1+\frac{1}{N(p)}\left(1-\frac{1}{N(p)}\right)^{k-2}\right\}\cdot\left\{1+\left(1-\frac{1}{N(p)}\right)^k \sum_{t=1}^{\alpha(p)} \sum_{m=t}^{\infty} \frac{S_k(m)}{(N(p))^m} + \right. \tag{5}$$

$$\left. + \frac{1}{N(p)}\left(1-\frac{1}{N(p)}\right)^{k-2} \frac{S_k(\alpha(p))}{N(p^{\alpha(p)})}\right\}.$$

Where q runs through the Gaussian numbers, $N(q)$ is the norm of q, p denotes prime number, $\alpha(p)$ is the greatest natural number, such that $p^{\alpha(p)}|l$ and

$$S_k(m) = \sum_{\beta_1+\ldots+\beta_k=m,\ \beta_j\geqq 0} 1$$

is the number of partitions of m into nonnegative summands.

1. Using the dispersion method, we regard in spite of the equation

$$\xi\eta - \vartheta_1\ldots\vartheta_k = l, \quad \vartheta_1\ldots\vartheta_k \in x\Omega \tag{6}$$

the equation

$$\xi\eta - \vartheta_1\ldots\vartheta_k = a_j l, \quad \vartheta_1\ldots\vartheta_k \in x\Omega \tag{7}$$

where the numbers a_j are quasi primes, that is if p is a prime divisor of a_i, then we have

$$|p| \geqq \exp(\log\log x)^{3/2}. \tag{8}$$

We suppose, moreover, that

$$\frac{1}{2}x^{1-\varepsilon_0} \leqq |a_j| \leqq x^{1-\varepsilon_0} \quad \left(0 < \varepsilon_0 \leqq \frac{1}{100}\right). \tag{9}$$

Let $a_1, \ldots, a_H$ be the set of all quasi prime numbers, with the properties (8) and (9). The numbers

$$q_0 = l, \quad q_1 = a, l, \ldots \quad q_H = a_H l$$

we call coherent numbers.

We want to bring equations (6) and (7) to another form, more suitable for the dispersion method.

First we show that the terms in (1), for which $\tau_k(z) \geqq (\log x)^K$ have estimation $o\big(x(\log x)^{-K/2}\big)$ even if we substitute g_j for l, where K is a suitable constant.

First we remark that if we denote by $d_k(n)$ the number of decompositions of the natural number n, into products of k natural numbers, then we have the estimation

$$\sum_{n\leqq x} \big(d_k(n)\big)^l = O\big(x(\log x)^{r(k,l)}\big) \tag{A}$$

where $r(k, l)$ is a suitable function of k and l but independent of x (see LINNIK [1]).

From this for $\tau_k(z)$ we get

$$\sum_{|z|\leqq x} \big(\tau_k(z)\big)^l \leqq \sum_{n\leqq x^2} \sum_{N(z)=n} \big(d_k(n)\big)^l \leqq \sum_{n\leqq x^2} \big(d_k(n)\big)^{l+1} = o\big(x^2(\log x)^{r(k,\,l+1)}\big). \tag{B}$$

Then

$$\sum_{z\in x\Omega} \big(\tau_k(z)\tau(z+q_j)\big)^2 \leqq \sum_{|z|\leqq x\,\sup|\Omega|} \big(\tau_k(z)\big)^4 = o\big(x^2(\log x)^{r(k,\,5)}\big). \tag{10}$$

From (10) we get

$$\sum_{z\in x\Omega}{}' \tau_k(z)\tau(z+q_j) = o\big(x^2(\log x)^{-K/2}\big)$$

where prime denotes that z_k runs through Gaussian numbers for which $\tau_k(z) \geqq (\log x)^K$, and $K = 2r(k, 5)$.

2. We write equations (6) and (7) in the form

$$\xi\eta - z = q_j \qquad z \in x\Omega,\ j = 0, 1, 2, \ldots, H \tag{11}$$

where each z is considered $\tau_k(z)$ times. As we have seen, we may neglect the solutions, for which

$$\tau_k(z) \geqq (\log x)^K.$$

It is easy to see that we may neglect the solutions for which $|\xi| = |\eta|$. (They have estimation $Bx^{1+\varepsilon_a}$, where ε_a can be made arbitrarily small.)

So if we suppose that $|\xi| < |\eta|$, and then consider the double number of the solutions, we get an error $o(x^{1+\varepsilon_a})$. Now we show that the number of solutions for which $|(\xi, q_j)| > |(\xi, l)|$ may be neglected.

Indeed, the number of such solutions may be estimated so:

$$B(\log x)^K \sum_{p|a_j} \sum_{|\xi_1| < \frac{1}{|p|}\sqrt{x}} \sum_{\substack{z_1 \equiv q_j' \pmod{\xi_1} \\ |z_1| < \frac{2x}{|p|}}} 1 = \tag{12}$$

$$= Bx^2(\log x)^{K+1} \sum_{p|a_j} \frac{1}{N(p)} = Bx^2 \exp\left(-\sqrt{\log x}\right).$$

Now we consider the solutions of (11) for which z has a prime divisor v with the property

$$\exp(\log x)^{\varepsilon_1} \leqq |v| \leqq x^{\mu} \tag{13}$$

for arbitrary $0 < \varepsilon_1 < 1$, $0 < \mu < 1$. More exact values of ε_1 and μ will be chosen later.

If (13) is not fulfilled, we have the following possibilities

I. z has only prime divisors p with $|p| < \exp(\log x)^{\varepsilon_1}$.

II. z has only prime divisors p with $|p| > x^{\mu}$.

III. z has only prime divisors for which I. or II. is fulfilled.

We show, that the number of solutions for which I., II. or III. is fulfilled, has an estimation $Bx^2(\log x)^{\varepsilon_b}$, where

$$0 < \varepsilon_b < 1.$$

If I. is fulfilled, then we have an estimation for the number of solutions

$$B(\log x)^K \sum_{|z| < x - q_j}' \tau(z + q_j) = B(\log x)^K \left(\sum_{|z| \leqq x}' 1\right)^{1/2} \left(\sum_{|m| < x} \tau^2(m)\right)^{1/2}. \tag{14}$$

Here prime denotes, that for z I. is fulfilled.

We have

$$\sum_{|z| \leqq x}' 1 = \sum_{n \leqq x^2}' \sum_{\alpha^2 + \beta^2 = n} 1 < \sum_{n \leqq x^2}' d(n). \tag{15}$$

The set $\{1, x^2\}$ of natural numbers we divide into two subsets: the set for which $d(n) \leqq \exp\left(\frac{1}{2}\sqrt{\log x}\right)$ and for which $d(n) > \exp\left(\frac{1}{2}\sqrt{\log x}\right)$. We denote by $\sum'^*$

the sum for the first, and by $\sum'^{**}$ the sum for the second subset. We have

$$\sum'^{*} d(n) \leqq \exp\left(\frac{1}{2}\sqrt{\log x}\right) \sum_{n \leqq x^2}' 1 = Bx^2 \exp\left(-\frac{1}{2}\sqrt{\log x}\right) \tag{16}$$

and, as by (A)

$$\sum_{n \leqq x^2}'^{**} d(n) = Bx^2 \exp\left(-\frac{1}{4}\sqrt{\log x}\right). \tag{17}$$

We have for (14)

$$Bx^2 \exp\left(-\frac{1}{8}\sqrt{\log x}\right). \tag{18}$$

If II. is fulfilled, we must use the version of BRUN—TITCHMARSH's theorem for Gaussian numbers. Let us consider the prime numbers, occurring in the set $\{A\xi + C,\ |\xi| < G\}$, where A, C, ξ are Gaussian numbers, $(A, C) = 1$. If $A\xi + C$ is prime, then $N(A\xi + C)$ is a prime of the form $4h+1$, or the quadrat of a prime of the form $4h+3$.

We get for the primes of the first form, using the sieve method of A. SELBERG, the following estimation

$$B \frac{G^2}{\varphi(N(A)) \log \frac{G}{N(A)}}. \tag{19}$$

For the primes of the second form we get the estimation

$$BG^{1+\varepsilon_G} \frac{1}{\varphi(N(A)) \log \frac{G}{N(A)}} \tag{20}$$

if we take into mind that the congruence $t^2 \equiv a \pmod{m}$ in natural numbers has at most 2^{ν} solutions, where ν is the number of distinct prime divisors of m.

Now if II. is fulfilled, we get for the number of solutions of (7) the estimation

$$B \sum_{|\xi| \leqq \sqrt{x}} \sum_{\substack{z \equiv q_j \pmod{\xi} \\ |z| < 2x}}' 1 = B \frac{x^2}{\log x} \sum_{|\xi| < \sqrt{x}} \frac{1}{\varphi(N(\xi))} = Bx^2. \tag{21}$$

If III. is fulfilled, then we take $z = z_1 z_2$, where

$$z_1 = \prod_{p} p^{\alpha},\ |p| < \exp(\log x)^{\varepsilon_1}, \quad z_2 = p_1 \ldots p_s,\ |p_j| > x^{\mu}.$$

Let us first suppose that

$$|z_1| < \exp(\log x)^{2\varepsilon_1}.$$

Then we have the following estimation for the number of the equation (7)

$$B \sum_{|\xi|<\sqrt{x}} \sum_{\substack{z_1 z_2 \equiv q_j \pmod{\xi} \\ |z_1 z_2|<2x}}' =$$

(22)

$$= B \sum_{|\xi|<\sqrt{x}} \sum_{|z_1|<\exp(\log x)^{2\varepsilon}}' \tau_k(z_1) \sum_{\substack{z_2 \equiv \frac{q_j}{(z_1,\xi)}\left(\frac{z_1}{(z_1,\xi)}\right)' \left(\mathrm{mod}\ \frac{\xi}{(z_1,\xi)}\right) \\ |z_2|<\frac{2x}{|z_1|}}} 1 .$$

Using the version for Gaussian numbers of the BRUN—TITCHMARSH theorem, we get for (22)

$$Bx^2 (\log x)^{\varepsilon_G}.$$

If $|z_1| > \exp(\log x)^{2\varepsilon_1}$ then we have for the number of solutions of the equation (7)

$$B(\log x)^K \sum \tau(z_1 z_2 + q_j) = Bx^2 (\log x)^{K+1} \sum_{\exp(\log x)^{2\varepsilon}<|z_1|<2x}' \frac{1}{N(z_1)} =$$

$$= Bx^2 \exp\left(-\frac{1}{4}(\log\log x)^2\right).$$

It is easy to prove, that for such z, for which z is divisible by the quadrate of a prime of the form (13), the number of the solutions has the estimation

$$Bx^2 \exp\left(-\frac{1}{4}\sqrt{\log x}\right). \tag{23}$$

So we can suppose that every prime divisor of z of the form (13) is simple. So if ν_1 and ν_2 are two such prime divisors of z, then we have also $|\nu_1| \neq |\nu_2|$.

If we choose the least of these primes, and denote this by ν we can write the equation (7) in the form

$$\xi\eta - \nu D' = q_j;\ \nu D' \in x\Omega \tag{24}$$

where ν is a prime of the form (13) and if ν_2 is a prime divisor of D', for which $|\nu_2| \geqq \exp(\log x)^{\varepsilon_1}$, then $|\nu_2| > |\nu|$.

There is no other dependence between ν and D'. Every solution of (24) must be multiplied by

$$\tau_k'(D') = k\tau_k(D') \leqq (\log x)^K.$$

3. So ν and D' are dependent but we can get rid off this dependence in the following way: we divide the region $\{\exp(\log x)^{\varepsilon_1} \leqq |z| \leqq x^{\mu}\}$ into subregions of the form

$$\{\nu_0 \leqq |z| \leqq \nu_0 + \nu_0',\ \varphi_1 \leqq \arg z < \varphi_2\} = \theta(\nu_0, \varphi_1, \varphi_2)$$

where $\nu_0' = \dfrac{\nu_0}{(\log x)^{K_1}}$, $K_1 > 100$ and $\varphi_2 - \varphi_1 = \dfrac{2\pi}{(\log x)^{K_2}}$, $K_2 > 100$, K_1, K_2 are natural numbers.

If $\nu \in \theta(\nu_0, \varphi_1, \varphi_2)$ we substitute the dependence between ν and D' by the suppositions

(25)

a) $$D' \in \frac{x}{\nu_0} e^{-i\varphi_1} \Omega$$

b) if ν_2 prime, $\nu_2 | D'$ and $|\nu_2| \geqq \exp(\log x)^{\varepsilon_1}$, then $|\nu_2| > \nu_0$.

Then we can estimate the error in the following way: this error has an upper bound

(26) $$B(\log x)^K \sum_{|\xi| < c\sqrt{x}} \sum_{\nu \in \theta} \sum_{\substack{D'\nu \equiv q_j \pmod{\xi} \\ D' \in \Omega_\theta}} 1$$

Where the region Ω_θ we get if we remove from the region $x(\nu_0 - \nu_0')^{-1}\Omega$ the region $x(\nu_0 + \nu_0')^{-1}\Omega$. For (26) we get $B\frac{x^2}{\nu_0^2}(\log x)^{K-K_1}\pi(\theta) \sum_{|\xi| < c\sqrt{x}} \frac{1}{N(\xi)}$ where $\pi(\theta)$ denotes the prime numbers in $\theta(\nu_0, \varphi_1, \varphi_2)$. This last estimation equals to $B\frac{x^2}{\nu_0^2}(\log x)^{K-K_1+1}\pi(\theta)$. For $\pi(\theta)$ we can get by using SELBERG's method

$$B\frac{\varphi_2 - \varphi_1}{\pi} \frac{\nu_0 \nu_0'}{\log|\nu_0 \nu_0'|}.$$

Substituting this, and summarizing for $\nu_0, \varphi_1, \varphi_2$, we get for (26)

$$Bx^2(\log x)^{-\frac{K_1}{2}}$$

if we choose K_1 big enough.

If $\nu_0 \leqq x^\mu < \nu_0 + \nu_0'$, the number of solutions of (24) has an estimation

$$Bx^2(\log x)^{-\frac{K_1}{2}}.$$

4. Now we see easily that the number of solutions of (24) for which

$$|D'| < \frac{x}{\nu_0(\log \nu_0)^{K_3}} \qquad K_3 > 100$$

has an estimation

$$Bx^2(\log x)^{-\frac{K_3}{3}}$$

The region $\left\{D' \in \frac{x}{\nu_0} e^{-i\varphi_1}\Omega;\ |D'| \geqq \frac{x}{\nu_0(\log \nu_0)^{K_3}}\right\}$ we divide into subregions of the form

(27) $$\{D' : D_1 < |D'| \leqq D_1 + D_2,\quad \psi_1 < \arg D' \leqq \psi_2\}$$

$$D_2 = \frac{D_1}{(\log x)^{K_4}}, \quad \psi_2 - \psi_1 = \frac{\pi}{[\log x]^{K_4}}, \qquad K_4 > 100.$$

The subregions of the form (27) may not be complete, but the number of solutions of (24) with D' in such regions may be neglected.

5. Let us denote a region of the form (27) by $\Delta(D_1, \psi_1)$. We have got the following problem: to find the number of solutions of (24), where $\nu \in \theta(\nu_0, \varphi_1)$, $D' \in \Delta(D_1, \psi_1)$ and D' satisfies the supposition (25) b).

The next step is to prove, that for different q_{j_1} and q_{j_2} this number is almost the same. For this purpose let us examine the following expression:

$$V(q_{j_1}, q_{j_2}) = \sum_{D' \in \Delta(D_1, \psi_1)} \left(\sum_{\nu \in \theta(\nu_0, \varphi_1)} \tau(q_{j_1} + \nu D') - \sum_{\nu \in \theta(\nu_0, \varphi_1)} \tau(q_{j_2} + \nu D') \right)^2$$

FUNDAMENTAL LEMMA.

$$V(q_{j_1}, q_{j_2}) = B D_2 \nu_0'^4 (\log x)^{-\frac{K_1}{3}}. \tag{28}$$

6. We prove this lemma in more steps. These steps are analogous to those of LINNIK [1] and BREDIKHIN [2]. It is easy to see, that

$$V(q_{j_1}, q_{j_2}) = V_1 - 2V_2 + V_3 + Bx^2 \exp\left(-(\log\log x)^2\right).$$

Here V_1 is equal to the number of the solutions of the equation

$$\nu_1 \xi\eta - \nu_2 \zeta\vartheta = q_{j_1}(\nu_1 - \nu_2) \tag{29}$$

where $\frac{\xi\eta - q_{j_1}}{\nu_2} \in \Delta(D_1, \psi_1)$, $|\xi| < |\eta|$, $|\zeta| < |\vartheta|$ and V_2 is equal to the number of the solutions of the equation

$$\nu_1 \xi\eta - \nu_2 \zeta\vartheta = q_{j_2}(\nu_1 - \nu_2) \tag{30}$$

where $\frac{\xi\eta - q_{j_2}}{\nu_2} \in \Delta(D_1, \psi_1)$, $|\xi| < |\eta|$, $|\zeta| < |\vartheta|$ and V_3 is equal to the number of the solutions of the equation

$$\nu_1 \xi\eta - \nu_2 \zeta\vartheta = q_{j_1}\nu_1 - q_{j_2}\nu_2 \tag{31}$$

where

$$\frac{\xi\eta - q_{j_1}}{\nu_2} \in \Delta(D_1, \psi_1)\ |\xi| < |\eta|,\ |\zeta| < |\vartheta|.$$

7. In the following we examine the equation

$$\nu_1 \xi\eta - \nu_2 \zeta\vartheta = q_a \nu_1 - q_b \nu_2 = M \tag{32}$$

with the suppositions

$$\text{a)}\ \ \frac{\xi\eta - q_a}{\nu_2} \in \Delta(D_1, \psi_1); \qquad \text{b)}\ \ |\xi| < |\eta|; \qquad \text{c)}\ \ |\zeta| < |\vartheta|$$

q_a and q_b are coherent numbers.

Now let us first suppose that $(\xi, \zeta) = 1$. Then we can write (32) in the form

$$\eta \equiv M\nu_1'\xi' \pmod{\nu_2\zeta} \quad \text{where} \quad \nu_1'\xi'\nu_1\xi \equiv 1 \pmod{\nu_2\zeta} \tag{33}$$

with the suppositions (32) a), b). The supposition (32) c) we can write in the form

$$|\vartheta| = \left| \frac{\nu_1(\xi\eta - q_a) + q_b\nu_2}{\nu_2\zeta} \right| > |\zeta|. \tag{33 c)}$$

We can write (33) c) in the form

$$|\zeta| < \sqrt{\left| v_1 \frac{\xi\eta - q_a}{v_2} + q_b \right|}. \tag{34}$$

As in [1] we can substitute this supposition by

$$|\zeta| < \sqrt{D_1 |v_1|}\,. \tag{35}$$

The error term we get by this substitution is

$$BD_2^2 (\log x)^{-K_5} \tag{35'}$$

where K_5 is as large as we want, depending on K_1, K_2, K_3, K_4.
From (32) b) we get

$$\left| \frac{\xi\eta - q_a}{v_2} \right| > \frac{|\xi|^2}{|v_2|} - \left| \frac{q_a}{v_2} \right|. \tag{36}$$

Using again the fact that $\frac{\xi\eta - q_a}{v_2} \in \Delta(D_1, \psi_1)$, we may substitute (36) by

$$|\xi| < \sqrt{D_1 |v_2|} \tag{37}$$

with the error term

$$BD_2^2 (\log x)^{-K_5}. \tag{37'}$$

From (32) a) we get for η the range

$$\eta \in \frac{v_2 \Delta(D_1, \psi_1) + q_a}{\xi}. \tag{38}$$

For fixed v_1, v_2, ξ and ζ we have for the number of solutions of (33) with the suppositions (35), (37) and (38)

$$\frac{|\Delta(D_1, \psi_1)|}{N(\xi \cdot \zeta)} + B \frac{D_2}{|\xi\zeta|}. \tag{39}$$

In the following we take v_1 and v_2 fixed. For the set $\{\xi, \zeta : N(\xi \cdot \zeta) < D_2^2 x^{-\varepsilon_3}\}$ we get the number of solutions of (33)

$$|\Delta(D_1, \psi_1)| \sum_{U_1} \frac{1}{N(\xi \cdot \zeta)} + BD_2^2 x^{-\varepsilon_4} \tag{40}$$

where U_1 denotes the set of the suppositions (35), (37), (38)

$$N(\xi\zeta) < D_2^2 x^{-\varepsilon_3} \quad \text{and} \quad (\xi, \zeta) = 1 \quad \text{and} \quad \varepsilon_4 = \tfrac{1}{2}\varepsilon_3$$

Now let us suppose that

$$N(\xi\zeta) \geqq D_2^2 x^{-\varepsilon_3}. \tag{41}$$

Then

$$|\xi| > x^{1-\mu-1/2-\varepsilon_5} > x^{1/6-\varepsilon_5} \quad \text{if} \quad \mu < \tfrac{1}{3}, \quad \varepsilon_5 = 2\varepsilon_3. \tag{42}$$

So for fixed ζ we have the range for ξ

$$\frac{D_2 x^{-\frac{\varepsilon_3}{2}}}{|\zeta|} \leqq |\xi| \leqq \sqrt{D_1 |v_2|}. \tag{43}$$

Let us divide this region into subregions of the form

$$\begin{gathered} \alpha < \arg \xi \leqq \alpha + \alpha' \\ \beta < \arg \left(\xi - A(\xi_0, \alpha)\right) \leqq \beta + \beta' \\ |A(\xi_0, \alpha) - \xi_0 e^{i\alpha}| \geqq \xi_0 \end{gathered} \tag{44}$$

$\alpha, \alpha+\alpha', \beta, \beta+\beta'$

have tangents of the form $\frac{j}{[x^{1/4}]}$.

For fixed ξ_0 and α let us substitute the range for η by the region

$$\frac{1}{\xi_0} \exp(-\alpha i)(v_2 \Delta(D_1, \psi_1) + q_a). \tag{45}$$

This region differs from that of (38) by a set of measure $Bx^2 \xi_0^{-2-\varepsilon_6}$. For the error term of (32) we get

$$BD_2^2 v_0'^4 x^{-\varepsilon_7} \qquad \varepsilon_7 = \varepsilon_7(\varepsilon_6). \tag{46}$$

Now let us divide the region (45) into two subregions. The first is the union of the squares of the lattice Λ generated by the vectors $v_2\zeta$ and $i\vartheta_2\zeta$, and contained completely by the region (45). Let us denote this union by $Y_1(\xi_0)$.

Let us denote the remaining part by $Y_2(\xi_0)$. We divide $Y_2(\xi_0)$ too into subregions so that one subregion is the part of $Y_2(\xi_0)$ contained by a given square of the lattice Λ. Let us denote these subregions by $Y_{2j}(\xi_0)$. $(j = 1, 2 \ldots R)$ ($R =$ the number of the subregions.) These subregions or their complements to the square are convex. Obviously

$$R = B \frac{D_2}{|\zeta| \xi_0}. \tag{47}$$

Let us denote by $|Y_{2j}(\xi_0)|$ the area of $Y_{2j}(\xi_0)$. For those $Y_{2j}(\xi_0)$ for which

$$\frac{|Y_{2j}(\xi_0)|}{N(v_2\zeta)} < \frac{x^{2-\varepsilon_8}}{N(v_2\zeta)} \xi_0^{-1} \tag{48}$$

the number of solutions of (32) has an estimation

$$BD_2^2 v_0'^4 x^{-\varepsilon_9} \qquad \varepsilon_9 = \varepsilon_9(\varepsilon_8). \tag{49}$$

So we consider the case

$$\frac{|Y_{2j}(\xi_0)|}{N(v_2\zeta)} \geqq \frac{x^{2-\varepsilon_8}}{N(v_2\zeta)} \xi_0^{-1}. \tag{50}$$

Now we need the following

LEMMA: *Let Γ be a convex region in the plane, contained by the unity square $\{(0,0), (0,1), (1,0), (1,1)\}$, $r>0$ natural number, γ a real number, $0<\gamma<\frac{1}{2}$, $\gamma<\inf d$, where d is the distance of two parallel straights surrounding Γ; $1-\gamma \geqq \sup_{x,\, y\in\Gamma} d(x, y)$, where $d(x, y)$ is the distance of the points x, y. Then there exists a function $\psi(x, y)$, for which*

1. *$\psi(x, y)=1$ if $(x, y)\in\Gamma$ and the distance of (x, y) from the complementary of Γ is $\geqq\frac{1}{2}\gamma$.*

2. *$\psi(x, y)=0$ if $(x, y)\notin\Gamma$ and the distance of (x, y) from Γ is $>\frac{1}{2}\gamma$.*

3. *$0\leqq\psi(x, y)\leqq 1$ if the distance of (x, y) from Γ and from the complementary of Γ is $\leqq\frac{1}{2}\gamma$.*

4. *$\psi(x, y)$ can be expanded into Fourier series*

$$\psi(x, y = |\Gamma| + \sum_{\substack{m,\, n \\ m^2+n^2\neq 0}}\sum a_{mn} \exp(2\pi i(mx+ny)$$

such that

$$\text{a)}\quad |a_{mn}| \leqq \frac{B}{\max(m, 1)\max(n, 1)}, \qquad \text{b)}\quad |a_{mn}| \leqq B|\Gamma|,$$

$$\text{c)}\quad |a_{mn}| \leqq B(\max(m, 1)\max(n, 1))^{-1}\left(\frac{r^2}{\max(m, 1)\max(n, 1)\gamma^2}\right)^r.$$

PROOF: Let $\psi_0(x, y)=0$, if (x, y) is an outer point of Γ, $\psi_0(x, y)=1$ if (x, y) is an inner point of Γ, and $\psi_0(x, y)=\frac{\alpha}{2\pi}$ if (x, y) is a boundary point of Γ, and α is the infimum of the measures of the angles, containing Γ, and whose vertex is in (x, y). $\psi_0(x, y)$ satisfies the conditions 1., 2., 3. Then we know from the theory of the Fourier series that $\psi_0(x, y)$ can be expanded into Fourier series

$$\psi_0(x, y) = |\Gamma| + \sum_{\substack{m,\, n \\ m^2+n^2\neq 0}}\sum a_{mn}^{(0)} \exp(2\pi i(mx+ny))$$

for which the coefficients a_{mn} satisfy a) and b). Now we choose $\delta=\frac{\gamma}{2r}$ and define the functions $\psi_\varrho(x, y)$ $(\varrho=1, 2, \ldots r)$ recursively by

$$\psi_\varrho(x, y) = \frac{1}{\delta^2}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta}\psi(x+\xi, y+\eta)\,d\xi\,d\eta.$$

Then $\psi_\varrho(x, y)$ for $1<\varrho\leqq r$ satisfies conditions 1., 2., 3. It can be expanded into Fourier series

$$\psi_\varrho(x, y) = |\Gamma| + \sum_{\substack{m,\, n \\ m^2+n^2\neq 0}}\sum a_{mn}^{(\varrho)} \exp(2\pi i(mx+ny))$$

and it is evident that the coefficients $a_{mn}^{(\varrho)}$ satisfy the conditions a) and b). We have the recursive relations

$$a_{mn}^{(\varrho)} = a_{mn}^{(\varrho-1)} \cdot \frac{\sin 2\pi m\delta \sin 2\pi n\delta}{\pi mn\delta^2}.$$

If we take $\varrho = r$, condition c) will be satisfied, too. If we have two mod 1 periodic functions $F_i(x, y)$, $i = 1, 2$ of integral x and y, and we want to estimate the number $N\{(F_1, F_2) \in \Gamma | (x, y) \in S\}$, where S is a set of Gaussian numbers, from above, then, denoting by Γ_1 the set of points, whose distance from Γ is not greater than $\gamma_1 = \frac{1}{2}\gamma$ and we determine the function $\psi_1(x, y)$ belonging to Γ_1 and γ_1 and some r and then examine the sum

$$\sum_{(x, y) \in S} \psi_1(F_1(x, y), F_2(x, y)) \tag{51}$$

which is on upper bound for $N\{(F_1, F_2) \in \Gamma | (x, y) \in S\}$. We get a lower estimate, if we take the set of points $\Gamma_2 \subset \Gamma$, whose distance from the complementary of Γ is not less than $\gamma_1 = \frac{1}{2}\gamma$.

If we want to determine the number of the solutions of (33) where $\eta \in Y_{2k}(\xi_0)$ by this method, we must examine the sums of the form

$$\sum_{\xi \in (\xi_0)} \exp\left(2\pi i \frac{\bar{Q}A\xi' + Q\bar{A}\bar{\xi}'}{2N(Q)}\right) \tag{52}$$

where (ξ_0) is the set of Gaussian numbers ξ in (44) such that

$$(\xi, Q) = 1.$$

Let us denote the sum in (52) by $G(A)$, and the sum

$$\sum_{\substack{\xi \pmod{Q} \\ (\xi, Q) = 1}} \exp\left(2\pi i \frac{\bar{Q}A\xi + Q\bar{A}\bar{\xi} + \bar{Q}B\xi' + Q\bar{B}\bar{\xi}'}{2N(Q)}\right) \tag{53}$$

by $g(A, B)$.

Then

$$G(A) = \frac{1}{N(Q)} \sum_{B \pmod{Q}} \sum_{S \in [\xi_0]} g(B, A) \exp\left(2\pi i \frac{-\bar{Q}BS - Q\bar{B}\bar{S}}{2N(Q)}\right). \tag{54}$$

We denote by $[\xi_0]$ the region (44) without the assumption

$$(S, Q) = 1.$$

From (54) we get the estimation

$$|G(A)| \leqq \frac{1}{N(Q)} \max_{B, A} |g(B, A)| \sum_{B \pmod{Q}} \left| \sum_{S \in [\xi_0]} \exp\left(2\pi i \frac{-\bar{Q}BS - Q\bar{B}\bar{S}}{2N(Q)}\right)\right|. \tag{55}$$

Taking into account the conditions (44) we get

$$\sum_{B \pmod{Q}} \left| \sum_{S \in [\xi_0]} \exp\left(2\pi i \frac{-\bar{Q}BS - Q\bar{B}\bar{S}}{2N(Q)}\right)\right| = B_\varepsilon |Q|^\varepsilon. \tag{56}$$

Let us now examine $g(B, A)$. If Q is a Gaussian prime number (that is $N(Q)$ is a natural prime number of the form $4j+1$) then every residue class (mod Q) contains a rational element. So it follows immediately from WEYL's estimation that in this case (53) has an estimation

$$B|Q| \min\left(\sqrt{(N(A), N(Q))}, \sqrt{(N(B), N(Q))}\right). \tag{57}$$

If Q is a rational prime of the form $4j+3$, then (53) can be written in the form

$$\sum_{S} \sum_{t \pmod{Q}} \exp \frac{2\pi i \left(\left(A\xi_S^{(0)} + \overline{A\xi_S^{(0)}} \right) t + \left(B\xi_S^{(0)\prime} + \overline{B\xi_S^{(0)\prime}} \right) t' \right)}{2Q} \tag{58}$$

where $t=1, 2, \ldots, Q-1$, and $\xi_S^{(0)}$ $(S=1, 2, \ldots Q-1)$ are such Gaussian numbers that every residue class of Gaussian numbers has a well determined representative of the form $\xi_S^{(0)} t$.

If $(A, Q)=1$, then there exists a uniquely determined $\xi_{S_1}^{(0)}$ with

$$A\xi_{S_1}^{(0)} + \overline{A\xi_{S_1}^{(0)}} \equiv 0 \pmod{Q}. \tag{59}$$

So in this case (58) has an estimation

$$B|Q|^{3/2} \min \left(\sqrt{(A, Q)}, \sqrt{(B, Q)} \right). \tag{60}$$

In the general case we can get the estimation in the same way as in [9]

$$|g(A, B)| = B_\varepsilon |Q|^{3/2+\varepsilon}. \tag{61}$$

We get from (54), (56) and (60)

$$|G(A)| = B_\varepsilon |Q|^{3/2+\varepsilon}. \tag{62}$$

So for fixed $v_1, v_2, \zeta, \xi_0, \alpha$ and $Y_{2j}(\xi_0, \alpha)$ we get the number of solutions of (33) in the same way as in [1]

$$\frac{|Y_{2j}(\xi_0, \alpha)|}{N(v_2 \zeta)} (1 + Bx^{-\varepsilon_{10}}) \left(\sum_{\xi \in (\xi_0)} 1 \right) + R_\zeta \tag{63}$$

where

$$R_\zeta = B|v_2 \zeta|^{3/2 + \frac{\varepsilon_{10}}{6}} |(M, \zeta)|.$$

It can be shown in the same way as in [1] that the number of the solutions of (32) for which

$$|(M, \zeta)| > x^{\varepsilon_{10}}$$

has an estimation

$$BD_2^2 v_0'^4 x^{-\frac{\varepsilon_{10}}{2}}. \tag{64}$$

If $|(M, \zeta)| < x^{\frac{\varepsilon_{10}}{10}}$ and $\mu = \frac{1}{24}$, then if $\varepsilon_1, \ldots, \varepsilon_{10}$ are little enough, we have

$$R_\zeta = Bx^{7/8} \tag{65}$$

$$\frac{|Y_{2j}(\xi_0, \alpha)|}{N(v_2 \zeta)} \left(\sum_{\xi \in (\xi_0)} 1 \right) \geqq \xi_0^{2(1-2\varepsilon_6)} \frac{x^{2-\varepsilon_8}}{N(v_2 \zeta)} \xi_0^{-2} \geqq x^{\frac{23}{24} - 4\varepsilon_6 - \varepsilon_8} > x^{11/12}. \tag{66}$$

So we get for (63)

$$\frac{|Y_{2j}(\xi_0 \alpha)}{N(v_2 \zeta)} (1 + Bx^{-\varepsilon_{11}}) \sum_{\xi \in (\xi_0)} 1. \tag{67}$$

We get the number of the solutions for the whole range of η

$$\left(\frac{|\Delta(D_1,\psi_1)|}{\xi_0^2 N(\zeta)}+Bx^{-\varepsilon_{11}}\frac{D_2}{\xi_0|\zeta|}\min\left(1,\frac{D_2}{\xi_0|\zeta|}\right)\right)\sum_{\xi\in(\xi_0)}1. \tag{68}$$

As $\sum_{\xi_0|\zeta|<D_2}\frac{1}{\xi_0|\zeta|}=BD_2$ and $\sum_{|\zeta|<\sqrt{2x}}\frac{1}{N(\zeta)}=B\log x$, we get for the error term in (68) if we summarize for ξ_0, ν_1, ν_2 and ζ

$$BD_2^2 v_0'^4 x^{-\varepsilon_{12}} \tag{69}$$

if ε_{12} is little enough.

Taking this into account, we get for the number of the solutions of (32) from (40) and (66) for $(\xi,\zeta)=1$

$$4|\Delta(D_1,\psi_1)|\sum_{\nu_1,\nu_2\in\theta(\nu_0,\varphi_1)}\sum_{\substack{\zeta\\ U_2}}\left(\sum_{\substack{\xi\\ U_3}}\frac{1}{N(\xi\zeta)}+\sum_{(\xi_0)}\sum_{\substack{\xi\in(\xi_0)\\ U_4}}\frac{1}{\xi_0^2 N(\zeta)}\right)+BD_2^2(\log x)^{-K_5} \tag{70}$$

where U_2, U_3, U_4 denote the conditions

$$\begin{aligned}
&U_2: |(M,\zeta)|<x^{\varepsilon_{10}}, \quad |\zeta|<\sqrt{D_1|\nu_1|}\\
&U_3: |\xi|<\sqrt{D_1|\nu_2|}, \quad (\xi,\zeta)=1, \qquad N(\xi\zeta)<D_2^2 x^{-\varepsilon_3}\\
&U_4: |\xi|<\sqrt{D_1|\nu_2|}, \quad (\xi,\zeta)=1, \quad N(\xi\zeta)\geqq D_2^2 x^{-\varepsilon_3}.
\end{aligned} \tag{71}$$

The error term in (67) is composed from the error terms (35), (37), (40), (46), (49), (64), (69).

Now let us suppose that $|\delta|=|(\xi,\zeta)|>1$. For fixed δ we get in the same way as in the case $|\delta|=1$ for the number of the solutions of (32)

$$4\frac{\Delta(D_1,\psi_1)}{N(\delta)}\sum_{\nu_1,\nu_2\in\nu_0}{}'\sum_{\substack{\zeta\\ U_5}}\left(\sum_{\substack{\xi\\ U_6}}\frac{1}{N(\xi\zeta)}+\sum_{(\xi_0)}\sum_{\substack{\xi\in(\xi_0)\\ U_7}}\frac{1}{\xi_0^2 N(\zeta)}\right)+B\frac{D_2^2 v_0'^4}{N(\delta)}(\log x)^{-K_5}. \tag{72}$$

In (72) we must summarize for ν_1, ν_2 for which

$$M=q_a\nu_1-q_b\nu_2\equiv 0 \pmod{\delta}. \tag{73}$$

U_5, U_6, U_7 denote the conditions

$$\begin{aligned}
&U_5: \left|\left(\frac{M}{\delta},\frac{\zeta}{\delta}\right)\right|<x^{\varepsilon_{10}}, \quad |\zeta|<\sqrt{D_1|\nu_1|}\\
&U_6: |\xi|<\sqrt{D_1|\nu_2|}, \quad (\xi,\zeta)=\delta, \quad N(\xi\zeta)\leqq N(\delta)D_2^2 x^{-\varepsilon_3}\\
&U_7: |\xi|<\sqrt{D_1|\nu_2|}, \quad |\xi,\zeta|=\delta, \quad N(\xi\zeta)>N(\delta)D_2^2 x^{-\varepsilon_3}.
\end{aligned} \tag{74}$$

In the same way as in [1] we can make sure that the number of the solutions for which

$$|\delta|>(\log x)^{K_6}$$

has an estimation

$$BD_2^2 v_0'^4 (\log x)^{-\frac{K_6}{2}}. \tag{75}$$

So we can examine the case $|\delta| \leqq (\log x)^{K_6}$ only.

If we substitute the conditions U_5, U_6, U_7 by the conditions

$$U_8: \left|\left(\frac{M}{\delta}, \frac{\zeta}{\delta}\right)\right| < x^{\varepsilon_{10}}, \quad |\zeta| < \sqrt{D_1 v_0}$$

$$U_9: |\xi| < \sqrt{D_1 v_0}, \quad (\xi, \zeta) = \delta, \quad N(\xi\zeta) < N(\delta) D_2^2 x^{-\varepsilon_3} \tag{76}$$

$$U_{10}: |\xi| < \sqrt{D_1 v_0}, \quad (\xi, \zeta) = \delta, \quad N(\xi\zeta) \geqq N(\delta) D_2^2 x^{-\varepsilon_3}$$

then in the same way as in [1] we get the error term

$$B \frac{D_2^2}{N(\delta)} v_0'^2 (\log x)^{-K_7}. \tag{77}$$

So we get for (72)

$$4 \frac{|\Delta(D_1, \psi_1)|}{N(\delta)} \pi(q_a, q_b) \sum_{\substack{\zeta \\ U_8}} \left(\sum_{\substack{\xi \\ U_9}} \frac{1}{N(\xi\zeta)} + \sum_{(\xi_0)} \sum_{\xi \in (\xi_0)} \frac{1}{\xi_0^2 N(\zeta)} \right) + \tag{78}$$

$$+ B \frac{D_2^2 v_0'^4}{N(\delta)} (\log x)^{-K_7}$$

where $\pi(q_a, q_b)$ is the number of the solutions of the congruence

$$q_a v_1 - q_o v_2 \equiv 0 \pmod{\delta}. \tag{79}$$

With primes v_1, $v_2 \in \theta(v_0, \varphi_1)$. To determine $\pi(q_a, q_b)$ we need the following

LEMMA: *Let δ be Gaussian number $x > 0$, $|\delta| < (\log x)^K$, $K > 0$, l Gaussian $|l| < 2\delta$, $(\delta, l) = 1$, $0 < \alpha < 2\pi$, $\alpha' \geqq \frac{2\pi}{(\log x)^K}$, $K > 0$ fixed. Then the number of the Gaussian primes P for which*

$$P \equiv l \pmod{\delta} \tag{80}$$

$$\alpha \leqq \arg p \leqq \alpha + \alpha', \quad |p| < x$$

is equal to

$$\frac{\alpha}{2\pi} \frac{1}{2\varphi(N(\delta))} \operatorname{li}(x^2) + Bx^2 \exp\left(-c\sqrt{\log x}\right) \tag{81}$$

where c is an absolute constant and $\varphi(\cdot)$ is the Euler's function.

PROOF: For Gaussian z let $\lambda_j(z) = \left(\frac{z}{|z|}\right)^j$ $(j = 0, 1, \ldots)$, $\chi(z)$ a character mod δ and $\lambda_j(z)\chi(z) = H_j(z)$. Let further

$$\Lambda(z) = \begin{cases} \log |p| & \text{if } z = p^n \text{ where } p \text{ is a Gaussian complex prime} \\ 2 \log |p| & \text{if } z = p_1^n \text{ where } p \text{ is a Gaussian rational prime} \\ 0 & \text{in the remaining cases.} \end{cases}$$

Then we get from the results of HECKE [6], FOGELS [5] and KÖRNYEI [7] that

$$\sum_{|z|<x} H_j(z)\Lambda(z) = E_0 x^2 - E_1 \frac{x^{2\beta_1}}{\beta_1} + O\left(x^2 \exp(-c_0)\sqrt{\log x}\right) \tag{82}$$

where $E_0 = \begin{cases} 1 & \text{if } j=0, \quad \chi=\chi_0 \text{ the principal character} \\ 0 & \text{in the other cases.} \end{cases}$

$E_1 = \begin{cases} 1 & \text{if } j=0, \chi \text{ is a real character with exceptional real root } \beta_1 \\ 0 & \text{in the other cases.} \end{cases}$

As we know $0<\beta_1 \leqq 1 - \frac{\tau(\varepsilon)}{|\delta|\varepsilon}$. We get then, if we define

$$\psi(x, j, \delta, l,) = \sum_{\substack{|z|<x \\ z \equiv l \pmod{\delta}}} \lambda_j(z)\Lambda(z) \tag{83}$$

that

$$\psi(x, j, \delta, l) = \begin{cases} Bx^2 \exp\left(-c\sqrt{\log x}\right) & \text{if } j > 0 \\ \dfrac{x^2}{\varphi(N(\delta))} + Bx^2 \exp\left(-c\sqrt{\log x}\right) & \text{if } j = 0. \end{cases} \tag{84}$$

Now let us define

$$\psi(x, \alpha, \alpha', \delta, l) = \sum_{|z|<x} \sum_{z \equiv l \pmod{\delta}} \sum_{\alpha \leqq \arg z \leqq \alpha + \alpha'} \Lambda(z). \tag{85}$$

Then using VINOGRADOV's method (see [1] p. 42) with $\Delta = \exp\left(-c\sqrt{\log x}\right)$ and $r = [\exp(c\sqrt{\log x})]$ we get

$$\psi(x, \alpha, \alpha', \delta, l) = \frac{\alpha'}{2\pi} \frac{x^2}{\varphi(N(\delta))} + Bx^2 \exp\left(-c\sqrt{\log x}\right). \tag{86}$$

From this follows the Lemma.

From this lemma it follows that

$$\pi(q_a, q_b) = \frac{1}{2\varphi(N(\delta))[\log x]^{2K_1}} \left(\operatorname{li}\left((v_0+v_0')^2\right) - \operatorname{li}(v_0^2)\right)^2 + \tag{87}$$

$$+ Bv_0'^4 \exp(-2cv_0).$$

Substituting this into (78) we get for this expression

$$L + B\frac{D_2^2 v_0'^4}{N(\delta)} (\log x)^{-K_7} \tag{88}$$

where L does not depend on the coherent numbers. From this fact, summarizing for $N(\delta) < (\log x)^{K_6}$ we get the Fundamental Lemma.

It follows from the Fundamental Lemma that

$$\begin{aligned} &|N\{\xi\eta - vD' = q_{j_1};\ v \in \theta(v_0, \varphi_1),\ D' \in \Delta(D_1, \psi_1)\} - \\ &- N\{\xi\eta - vD' = q_{j_2};\ v \in \theta(v_0, \varphi_1),\ D' \in \Delta(D_1, \psi_1)\}| = \\ &= BD_2^2 v_0'^2 (\log x)^{-K_8}. \end{aligned} \tag{89}$$

Summarizing for the regions $\theta(v_0, \varphi_1)$ and $\Delta(D_1, \psi_1)$

$$|N\{\xi\eta - \vartheta_1 \dots \vartheta_k = q_{j_1}, \vartheta_1 \dots \vartheta_k \in x\Omega\} - \tag{90}$$

$$-N\{\xi\eta - \vartheta_1 \dots \vartheta_k = q_{j_2}, \vartheta_1 \dots \vartheta_k \in x\Omega\}| = B_\varepsilon x^2 (\log x)^\varepsilon.$$

And so we get

$$N\{\xi\eta - \vartheta_1 \dots \vartheta_k = l, \vartheta_1 \dots \vartheta_k \in x\Omega\} = \tag{91}$$

$$= \frac{1}{H} N\{\xi\eta - \vartheta_1 \dots \vartheta_k - q_j = 0, \vartheta_1 \dots \vartheta_k \in x\Omega, j = 0, 1, \dots, H\} + B_\varepsilon x^2 (\log x)^\varepsilon.$$

Let us now examine the equation

$$\xi\eta - \vartheta_1 \dots \vartheta_k - q_j = 0; \quad \vartheta_1 \dots \vartheta_k \in x\Omega, \; j = 0, \dots, H. \tag{92}$$

We divide the region $x\Omega$ into squares whose sides are parallel to real and imaginary axes and are of the length $x^{1-2\varepsilon_0}$. Then the incomplete squares give $BHx^{2-2\varepsilon_0}$ solutions of (92).

Let us denote these squares by I_a $(a = 1, 2, \dots)$. Let us extend the lattice of these squares to the whole plane. The set of those a for which $I_a \subset x\Omega$ we denote by T. Then let us denote

$$L(a, b) = N\{\xi\eta - \vartheta_1 \dots \vartheta_k - q_j = 0; \quad \xi\eta \in I_a, \vartheta_1 \dots \vartheta_k \in I_b, j = 0, \dots, H\}. \tag{93}$$

Let us denote by $T_1(b)$, $b \in T$ the set of natural numbers a for which $\frac{1}{2}|l|x^{1-\varepsilon_0} < |z_1 - z_2| < lx^{1-\varepsilon_0}$ for every $z_1 \in I_a$, $z_2 \in I_b$. Then for fixed $b \in T$ the number of such a for which $a \notin T_1(b)$ but I_a has at least one element z_1 and I_b, an element z_2 such that $\frac{1}{2} lx^{1-\varepsilon_0} < |z_1 - z_2| < lx^{1-\varepsilon_0}$ is

$$Bx^{\varepsilon_0}. \tag{94}$$

The number of solutions for such a, b is

$$Bx^{4-7\varepsilon_0}. \tag{95}$$

That is, we may suppose that $b \in T$, $a \in T_1(b)$. For such a, b $L(a, b)$ has a more simple form

$$L(a, b) = N\{\xi\eta - \vartheta_1 \dots \vartheta_k - a_j l = 0, \; \xi\eta \in I_a, \vartheta_1 \dots \vartheta_k \in I_b, \; a_j \in W\} \tag{94}$$

where W is the set of the quasiprime numbers, that is, the numbers every prime divisor of which is $\geqq \exp(\log\log x)^{3/2}$. We can write $L(a, b)$ in the form

$$L(a, b) = \sum_{\substack{\lambda \in \Lambda_x \\ |\lambda| < 2x}} \mu(\lambda) N\{\xi\eta - \vartheta_1 \dots \vartheta_k - l\lambda t = 0\} \tag{95}$$

where Λ_x is the set of the Gaussian numbers, every prime divisor of which $< \exp(\log\log x)^{3/2}$.

In the same way as in [1] we can prove that for the set of λ for which $\exp(\log\log x)^4 \leqq |\lambda| \leqq 2x$ we have the estimation

$$B_c x^{4-2\varepsilon_0} (\log x)^{-c}. \tag{96}$$

So we can suppose that $|\lambda| < x_1 = \exp(\log\log x)^4$. Now let us examine

$$N\{\xi\eta - \vartheta_1 \dots \vartheta_k - l\lambda t = 0,\ \xi\eta \in I_a,\ \vartheta_1 \dots \vartheta_k \in I_b\}$$

for fixed λ. If we put

$$S_j^{(a)}(A, l\lambda) = \sum_{\xi_1 \dots \xi_j \in I_a} \exp\left(2\pi i \frac{\overline{l\lambda} A\xi_1 \dots \xi_j + l\lambda \overline{A\xi_1 \dots \xi_j}}{2N(l\lambda)}\right) \tag{98}$$

then have

$$N\{\xi\eta - \vartheta_1 \dots \vartheta_k - l\lambda t = 0,\ \xi\eta \in I_a,\ \vartheta_1 \dots \vartheta_k \in I_b\} = \tag{99}$$

$$= \frac{1}{N(l\lambda)} \sum_{A \pmod{l\lambda}} S_2^{(a)}(A, l\lambda) S_k^{(a)}(-A, l\lambda).$$

If $\delta = (A, l\lambda)$, $A = A_1\delta$, $l\lambda = \lambda_1\delta$, then in the same way as in [1] we can get

$$S_j^{(a)}(A, l\lambda) = S_j^{(a)}(A_1, \lambda_1) = e_{\lambda_1}^{(j)} S_j^{(a)}(0, 1) + B\frac{\tau_j(\lambda_1)\log|\lambda_1|}{N(\lambda_1)\log x} = \tag{100}$$

$$= e_{\lambda_1}^{(j)} \sum_{\xi_1 \dots \xi_j \in I_a} 1 + B\frac{\tau_j(\lambda_1)\log|\lambda_1|}{N(\lambda_1)\log x} \quad j \geqq 2$$

where

$$e_{\lambda_1}^{(j)} = \prod_{p|\lambda_1} \left(1 - \left(1 - \frac{1}{N(p)}\right)^{j-1}\right) \quad \text{especially} \quad e_{\lambda_1}^{(2)} = \frac{1}{N(\lambda_1)}.$$

So we can write (99) in the form

$$\frac{1}{N(l\lambda)} \sum_{\lambda_1|\lambda} \frac{\varphi(N(\lambda_1)) e_{\lambda_1}^{(k)}}{N(\lambda_1)} \sum_{\substack{\xi\eta \in I_a \\ \vartheta_1 \dots \vartheta_k \in I_b}} 1 + \tag{101}$$

$$+ B\frac{S_2^{(a)}(0, 1) S_k^{(b)}(0, 1)}{N(l\lambda)} \sum \left(\frac{\log\log x}{\log x} \frac{\tau_k(\lambda_1)\log|\lambda_1|}{N(\lambda_1)} + \frac{\tau_k(\lambda_1)\tau(\lambda_1)\log|\lambda_1|}{N(\lambda_1)(\log x)^2}\right).$$

Summarizing the error term in (101) we get

$$BS_2^{(a)}(0, 1) S_k^{(a)}(0, 1) \frac{(\log\log x)^K}{\log x}. \tag{102}$$

Let us now denote $\frac{1}{N(l\lambda)} \sum_{\lambda_1|2\lambda} \frac{\varphi(N|\lambda_1|) e_{\lambda_1}^{(k)}}{N(\lambda_1)}$ by $C(l\lambda)$. For $L(a, b)$ we get from (95)

$$L(a, b) = \left(\sum_{\substack{\lambda \in A_x \\ |\lambda| < x_1}} \mu(\lambda) C(l\lambda)\right) \sum_{\substack{\xi\eta \in I_a \\ \vartheta_1 \dots \vartheta_k \in I_b}} 1 + BS_2^{(a)}(0, 1)\, S_k^{(a)}(0, 1) \frac{(\log\log x)^K}{\log x}. \tag{103}$$

Now if we summarize for a, b

$$\sum_{\substack{a \in T_1(b) \\ b \in T}} \sum_{\substack{\xi\eta \in I_a \\ \vartheta_1 \dots \vartheta_k \in I_l}} 1 = \sum_{\substack{\frac{1}{2} lx^{1-\varepsilon_0} \leqq |\xi\eta - \vartheta_1 \dots \vartheta_k| \leqq lx^{1-\varepsilon_0} \\ \vartheta_1 \dots \vartheta_k \in x\Omega}} 1 + Bx^{4-7\varepsilon_0}. \tag{104}$$

For fixed $\vartheta_1 \dots \vartheta_k$ we have

$$\sum_{\frac{1}{2} lx^{1-\varepsilon_0} \leqq |\xi\eta - \vartheta_1 \dots \vartheta_k| \leqq lx^{1-\varepsilon_0}} 1 = E(\log x + B) \tag{105}$$

where E is the area of the ring $\{z: \frac{1}{2} lx^{1+\varepsilon_0} \leqq |z| \leqq x^{1-\varepsilon_0}\}$. We have

$$\sum_{\vartheta_1 \dots \vartheta_k \in x\Omega} 1 = |\Omega| x^2 \frac{(\log x)^{k-1}}{(k-1)!} + Bx^2 (\log x)^{k-2}. \tag{106}$$

From (105) and (106) we get for (104)

$$\sum_{\substack{a \in T_1(b) \\ b \in T}} \sum_{\substack{\xi\eta \in I_a \\ \vartheta_1 \dots \vartheta_k \in I_b}} 1 = \Omega E \frac{x^2 (\log x)^k}{(k-1)!} + BEx^2 (\log x)^{k-1}. \tag{107}$$

We get for the number of the solutions of (90)

$$|\Omega| E \left(\sum_{\substack{\lambda \in \Lambda_x \\ |\lambda| < x_1}} \mu(\lambda) C(l\lambda) \right) \frac{x^2 (\log x)^k}{(k-1)!} + \tag{108}$$

$$+ BE \left(\sum_{\substack{\lambda \in \Lambda_x \\ |\lambda| < x_1}} \mu(\lambda) C(l\lambda) \right) x^2 (\log x)^{k-1} (\log \log x)^{K_4}.$$

Let us now examine the sum

$$\sum_{\substack{\lambda \in \Lambda_x \\ |\lambda| < x_1}} \mu(\lambda) C(l\lambda) = \sum_{\substack{\lambda \in \Lambda_x \\ |\lambda| < x_1}} \frac{\mu(\lambda)}{N(l\lambda)} \sum_{\lambda_1 | l\lambda} \frac{\varphi((N\lambda_1)) e_{\lambda_1}^{(k)}}{N(\lambda_1)}. \tag{109}$$

We have for this sum

$$\sum_{\substack{|\lambda_1| < x_1 \\ \lambda_1 \in \Lambda_x}} \sum_{l_1 l_2 = l} \frac{\mu(\lambda_1) e_{l_1 \lambda_1}^{(k)}}{N(l\lambda_1)} \prod_{p | l_1 \lambda_1} \left(1 - \frac{1}{N(p)}\right) \sum_{\substack{(\lambda_1, \lambda_2) = 1 \\ |\lambda_1 \lambda_2| < x_1 \\ \lambda_2 \in \Lambda_x}} \frac{\mu(\lambda_2)}{N(\lambda_2)} =$$

$$= \sum_{\substack{|\lambda_1| < x_1 \\ \lambda_1 \in \Lambda_x}} \sum_{l_1 l_2 = l} \frac{\mu(\lambda_1) e_{l_1 \lambda_1}^{(k)}}{N(l\lambda_1)} \prod_{p | l_1} \left(1 - \frac{1}{N(p)}\right) \prod_{p \nmid \lambda_1} \left(1 - \frac{1}{N(p)}\right) \sum_{\substack{(\lambda_1, \lambda_2) = 1 \\ |\lambda_1 \lambda_2| < x_1 \\ \lambda_2 \in \Lambda_x}} \frac{\mu(\lambda_2)}{N(\lambda_2)}. \tag{110}$$

From (110) we get for (109) in the same way as in [1]

$$\sum_{\substack{|t| < x_1 \\ t \in \Lambda_x}} \frac{\mu(t)}{N(t)} \cdot \sum_{l_1 l_2 = l} \sum_{\substack{|\lambda_1| < x_1 \\ \lambda_1 \in \Lambda_x}} \frac{\mu(\lambda_1) e_{l_1 \lambda_1}^{(k)}}{N(l\lambda_1)} \prod_{\substack{p | l_1 \\ p \nmid \lambda_1}} \left(1 - \frac{1}{N(p)}\right) + B(\log x)^{-5}. \tag{111}$$

We can extend the inner sum in (111) for every Gaussian $\lambda \neq 0$ with the error term $B(\log x)^{-5}$.

We have

$$0 < \sum_{l_1 l_2 = l} \sum_{\lambda} \frac{\mu(\lambda) e_{l_1 \lambda}^{(k)}}{N(l\lambda)} \prod_{\substack{p | l_1 \\ p \nmid \lambda}} \left(1 - \frac{1}{N(p)}\right) < \infty. \tag{112}$$

Let us denote the sum in (112) by A_k'. Then we get for (108)

$$|\Omega| E \sum_{\substack{|t|<x_1 \\ t\in A_x}} \frac{\mu(t)}{N(t)} A_k' \frac{x^2(\log x)^k}{(k-1)!} + BE \sum_{|t|<x_1} \frac{\mu(t)}{N(t)} x^2 (\log x)^{k-1} (\log\log x)^K. \tag{113}$$

As we know (see [1] lemma 1. 5. 2)

$$E \sum_{\substack{|t|<x_1 \\ t\in A_x}} \frac{\mu(t)}{N(t)} = N(l)H + B\frac{H}{(\log x)^2}.$$

From this the theorem follows if we write $A_k' N(l)$ instead of $A_k B_k(l)$.

Using the formulas

$$\sum_{l_1|l} \sum_{\lambda} \frac{\mu(\lambda) e_{l_1\lambda}^{(k)}}{N(\lambda)} \prod_{\substack{p|l_1 \\ p\nmid\lambda}} \left(1 - \frac{1}{N(p)}\right) = \sum_{l_1|l} \sum_{\delta|l_1} e_\delta^{(k)} \prod_{p|\delta} \left(1 - \frac{1}{N(p)}\right) \sum_{\substack{\lambda \\ (\delta_1\lambda)=1}} \frac{\mu(\lambda) e_\lambda^{(k)}}{N(\lambda)} \tag{114}$$

and

$$\sum_{\substack{\lambda \\ (\delta_1\lambda)=1}} \frac{\mu(\lambda) e_\lambda^{(k)}}{N(\lambda)} = \left(\sum_{\lambda} \frac{\mu(\lambda) e_\lambda^{(k)}}{N(\lambda)}\right) \prod_{p|\delta} \left(1 - \frac{1 - \left(1 - \frac{1}{N(p)}\right)^{k-1}}{N(p)}\right)^{-1}$$

we get

$$A_k' N(l) = A_k B_k(l) \tag{115}$$

and the theorem is proved.

I wish to express my thanks to Prof. LINNIK, Prof. CHUDAKOV, Mr. SKUBENKO and Mr. KÖRNYEI for their valuable advices.

REFERENCES

[1] Линник, Ю. В.: *Дисперсионный метод в бинарных аддитивных задачах*, Изд. Ленинградского Унив., 1961.

[2] Бредихин, Б. М.: Бинарные аддитивные проблемы неопределенного типа, *Изв. Акад. Наук СССР Сер. Мат.* **27** (1963) 439—462 и 777—794.

[3] Виноградов, А. И. и Линник, Ю. В.: Оценка суммы числа делителей в коротком отрезке арифметической прогрессии, *Успехи Мат. Наук* **12** (1957) № **4** (76), 277—280.

[4] Виноградов, А. И.: О числах с малыми простыми делителями, *Докл. Акад. Наук СССР* **109** (1956) 683—686.

[5] FOGELS, E.: On the abstract theory of primes, I., *Acta Arith.* **10** (1964) 137—182.

[6] HECKE, E.: Eine neue Art von Zetafunktionen und ihre Beziehungen zur Verteilung der Primzahlen, II, *Math. Z.* **6** (1920) 11—51.

[7] KÖRNYEI, I.: Eine Bemerkung zur Theorie der durch quadratische Formen darstellbaren Primzahlen, *Annales Un. Sci. Budapest. Eötvös Sect. Math.* **5** (1962) 95—108.

[8] PRACHAR, K.: *Primzahlenverteilung*, Berlin—Göttingen—Heidelberg, 1957.

[9] Малышев, А. В.: Обобщение суммы Клостермана и их оценки, *Вестник Ленинград. Унив. Сер. Мат. Мех. Астроном.* **13** (1960) 59—75.

COMPUTING CENTER OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received November 2, 1964)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 153—157

ON ANGLE-MODULATION PROCESSES

by

S. CSIBI

1. Posing the Problem

By angle-modulation we mean the following transformation of a real-valued random process, $\mu=\{\mu_t, t\in T\}$ into $\xi=\{\xi_t, t\in T\}$:

$$\xi_t=\cos(\Omega t+\mu_t+\Phi),$$

for all $t\in T$, where $T=(-\infty, \infty)$. $\Omega>0$ is a constant, Φ is a real-valued random variable, and μ and Φ are independent.

We shall call ξ angle-modulation process.

Obviously ξ is wide-sense stationary, provided $\nu=\{\exp i\mu_t, t\in T\}$ is stationary in the wide sense, and Φ is uniformly distributed in $(0, 2\pi]$. (E. g. ν is wide-sense stationary if μ is stationary in the strict sense.) In what follows we shall assume that ν as well as Φ meet these conditions.

More distinctly, we shall assume that ν is stationary in the wide sense and continuous in the mean. Accordingly [4]:

$$\nu_t=\int_{-\infty}^{\infty} e^{i\lambda t} z_\nu(d\lambda). \tag{1}$$

Here $z_\nu(A)$ is a random spectral measure defined for any Borel set, A on the real line, and (1) is defined as a limit in the mean.

We shall adopt the notations, $\Theta_t=\int_{-\infty}^{\infty}\exp(i\lambda t)\cdot z_\Theta(d\lambda)$ and $B_\Theta(\tau)=\mathbf{M}\Theta_t\overline{\Theta}_0$, for any random process, Θ which is stationary in the wide sense and continuous in the mean. (Here $z_\Theta(A)$ is a random spectral measure, and $\overline{\Theta}_t$ is the complex conjugate of Θ_t.) The spectral distribution function, F_Θ is defined by $F_\Theta(-\infty)=0$, and $F_\Theta(\lambda_2)-F_\Theta(\lambda_1)=\mathbf{M}|z_\Theta(A')|^2$, where $A'=(\lambda_1, \lambda_2]$ for any $\lambda_1<\lambda_2$. ($\mathbf{M}$ denotes the expectation.)

Observe that $z_\Theta(A'')=\overline{z_\Theta(A')}$ if Θ_t is real-valued. $(A''=[-\lambda_2, -\lambda_1)$ for any $\lambda_1<\lambda_2$.)

In communication theory it is a question of considerable interest to study the perturbation of angle-modulation processes. A prototype of this sort of problems is the perturbation of ξ due to a superimposed real-valued process, $\eta=\{\eta_t, t\in T\}$; i.e. the particular situation when, for any t, only

$$\zeta_t=\xi_t+\eta_t \tag{2}$$

is available for observation (or for further processing), instead of ξ_t, and η, μ and Φ are independent.

Within the scope of this paper we shall confine ourselves specifically to the study of independent additive perturbations such as given by (2). However, about η we shall only assume that it is stationary in the wide sense and continuous in the mean, its spectral distribution function, F_η being continuous at the origin.

Accordingly:

$$\eta_t = \int_{-\infty}^{\infty} e^{i\lambda t} z_\eta(d\lambda), \tag{3}$$

$$F_\eta(0+) = F_\eta(0-). \tag{4}$$

Notice that we have introduced (4) just for simplicity. In actual communication situations we may usually assume $F_\eta(0+) - F_\eta(0-) = 0$, and therefore restriction (4) is of no relevance.

From (3) and (4) it follows that

$$\zeta_t = \cos\psi_t + \int_0^{\infty} e^{i\lambda t} z_\eta(d\lambda) + \int_{-\infty}^{0} e^{i\lambda t} z_\eta(d\lambda) = |\chi_t| \cos(\psi_t + \varepsilon_t), \tag{5}$$

where $\psi_t = \Omega t + \mu_t + \Phi$,

$$\chi_t = 1 + 2\int_0^{\infty} e^{i(\lambda t - \psi_t)} z_\eta(d\lambda),$$

and

$$\varepsilon_t = \arg \chi_t = \operatorname{Im} \log \left[1 + 2\int_0^{\infty} e^{i(\lambda t - \psi_t)} z_\eta(d\lambda)\right]. \tag{6}$$

Frequently our interest is not a statistical inference from ζ on ξ but rather to describe $\varepsilon = \{\varepsilon_t, t \in T\}$, i. e. the deterioration of μ due to η, given F_v and Ω. More distinctly, our aim is to describe ε in terms of either the correlation or the spectral distribution function ([1]—[3], [5], [6]).

It is, of course, a question of considerable interest in what detail has one to specify η in order to be able to describe ε in terms of second moments, given F_v and Ω. We shall be concerned with this question specifically under the asymptotic condition $\mathbf{M}\eta_t^2 \to 0$.

Additive perturbations of angle-modulation processes have been frequently investigated under various particular assumptions [1]—[3], e.g. for a perturbation which is an (undistorted) angle-modulation process. However, more general situations are also of interest, e.g. when η is a moving average of an angle-modulation process [5].

The subject of this paper is a study of additive independent perturbations of an arbitrary, wide-sense stationary, angle-modulation process, described in terms of F_v and Ω, due to any η which is stationary in the wide-sense, and for which (3) and (4) hold.

2. Relations between Second-Moment Properties

First let us describe the asymptotic relation between η and ε by the following

LEMMA

$$\tilde{\varepsilon}_t = 2\,\mathrm{Im}\int_0^\infty e^{i(\lambda t-\psi_t)}z_\eta(d\lambda) \tag{7}$$

is an asymptotic approximation to ε_t *in the following sense:*

$$\lim_{M\eta_t^2\to 0}\mathrm{p}\,\frac{\varepsilon_t}{\tilde{\varepsilon}_t} = 1, \quad \text{for} \quad |\tilde{\varepsilon}_t| > 0, \tag{8}$$

and

$$\varepsilon_t = 0, \quad \textit{for} \quad \tilde{\varepsilon}_t = 0, \tag{9}$$

for any $t\in T$.

PROOF (9) follows immediately from (6) and (7). Therefore we may confine ourselves to $|\tilde{\varepsilon}_t|>0$.

Let

$$\varkappa_t = 2\int_0^\infty e^{i(\lambda t-\psi_t)}z_\eta(d\lambda).$$

Observe that

$$\mathbf{M}|\varkappa_t|^2 = 2\int_0^\infty dF_\eta(\lambda) = \mathbf{M}\eta_t^2,$$

and consider Čebyšev's inequality, i. e.:

$$\mathbf{P}(|\varkappa_t|^2 > C) \leqq \frac{\mathbf{M}\eta_t^2}{C}, \tag{10}$$

for any $C>0$.

From (6), (7) and the Taylor series of the logarithm we obtain:

$$\varepsilon_t-\tilde{\varepsilon}_t = \mathrm{Im}\sum_{n=2}^\infty (-1)^{2n}\frac{\varkappa_t^n}{n} = \tilde{\varepsilon}_t|\varkappa_t|\sum_{n=2}^\infty (-1)^{2n}\frac{|\varkappa_t|^{n-2}}{n}\frac{\sin n\alpha_t}{\sin\alpha_t}, \tag{11}$$

for any $t\in T$, and $|\varkappa_t|\leqq C<1$. (Here $\alpha_t=\arg\varkappa_t$, and $\tilde{\varepsilon}=\mathrm{Im}\,\varkappa_t$.)

Observe that $|\sin n\alpha_t/\sin\alpha_t|\leqq n$. Then, from (11) we obtain

$$\left|\frac{\varepsilon_t}{\tilde{\varepsilon}_t}-1\right| \leqq \frac{|\varkappa_t|}{1-C}, \tag{12}$$

for any $|\varkappa_t|\leqq C<1$, and $|\tilde{\varepsilon}_t|>0$.

Finally, from (10) and (12) it follows that

$$\mathbf{P}\left(\left|\frac{\varepsilon_t}{\tilde{\varepsilon}_t}-1\right| > \frac{C}{1-C}, |\tilde{\varepsilon}_t|>0\right) \leqq \mathbf{P}(|\varkappa_t| > C) \leqq \frac{\mathbf{M}\eta_t^2}{C},$$

for any $t\in T$, and $0<C<1$, that completes the proof.

REMARK 1. Since the transformation from η into $\tilde{\varepsilon}$ is the limit of linear combinations of time-invariant linear transformations, including coefficients independent of η but depending on ν, $\tilde{\varepsilon}$ is wide-sense stationary, provided ν and η are stationary in the wide-sense. Given F_ν and Ω, the spectral distribution function, $F_{\tilde{\varepsilon}}$ is obviously determined by F_η.

This relation is described in more detail by the following

THEOREM. I. *Let η and ν be stationary in the wide sense, then $\tilde{\varepsilon}$ is also wide-sense stationary, viz.*

$$B_{\tilde{\varepsilon}}(\tau) = 2 \operatorname{Re} B_\nu(\tau) \int_0^\infty e^{i(\Omega-\lambda)\tau} dF_\eta(\lambda). \tag{13}$$

II. *Specifically, if B_ν is real-valued, i.e. $B_\nu(\tau) = \overline{B_\nu(\tau)}$, for all τ, then*

$$F_{\tilde{\varepsilon}}(\lambda) = \int_{-\infty}^{\infty} F_\nu(\lambda - \omega)\, d\gamma_\eta(\omega),$$

for all $\lambda \in (-\infty, \infty)$. Here

$$\gamma_\eta(\lambda) = \sigma_\eta(\Omega + \lambda) - \sigma_\eta(\Omega - \lambda), \tag{15}$$

$$\sigma_\eta(\lambda_2) - \sigma_\eta(\lambda_1) = \begin{cases} F_\eta(\lambda_2) - F_\eta(\lambda_1), & \text{for} \quad \lambda_2 \geqq \lambda_1 \geqq 0, \\ 0 \qquad\qquad\quad , & \text{for} \quad 0 > \lambda_2 \geqq \lambda_1, \end{cases} \tag{16}$$

and $\sigma_\eta(-\infty) = 0$.

PROOF (13) follows from (7), observing the independence of η, μ and Φ, and the following well known properties of an integral with respect to a random measure:

$$\mathbf{M}\left[\int_0^\infty \varphi_1(\lambda) z_\eta(d\lambda) \int_0^\infty \varphi_2(\omega) \overline{z_\eta(d\omega)}\right] = \int_0^\infty \varphi_1(\lambda)\varphi_2(\lambda) dF_\eta(\lambda),$$

and

$$\mathbf{M}\left[\int_0^\infty \varphi_1(\lambda) z_\eta(d\lambda) \int_0^\infty \varphi_2(\omega) z_\eta(d\omega)\right] = 0,$$

for any $\varphi_i : \int_0^\infty |\varphi_i(\lambda)|^2 dF_\eta(\lambda) < \infty$, and $i = 1, 2$.

Next let B_ν be real-valued, and define σ_η by (16). Then for $\omega = \pm(\Omega - \lambda)$ it follows that

$$\int_0^\infty e^{\pm i(\Omega-\lambda)\tau} dF_\eta(\lambda) = \int_{-\infty}^{\infty} e^{i\omega\tau} d\beta_\eta(\omega), \tag{17}$$

where $\beta_\eta(\omega) = \mp \sigma_\eta(\Omega \mp \omega)$.

From (13) and (17) we obtain

$$B_{\tilde{\varepsilon}}(\tau) = B_\nu(\tau) \int_{-\infty}^{\infty} e^{i\lambda\tau} d\gamma_\eta(\lambda), \tag{18}$$

where γ_η is defined by (15).

Observe that $\gamma_\eta(\lambda_2)-\gamma_\eta(\lambda_1) = \gamma(-\lambda_1)-\gamma(-\lambda_2)\geqq 0$, for any $\lambda_1 \leqq \lambda_2$.

(14) follows from (18), the relation between B_θ and F_θ and the convolution theorem.

REMARK 2. The relation between the second-moment properties of the deterioration at the output and the interferring process at the input, described previously, is of particular relevance when studying angle-modulation channels. Obviously, if the deterioration at the output may be completely described by F_η (given F_ν and Ω), one has merely to examine the spectral distribution function of the interference sources and the attenuation (or gain) functions along all linear and time invariant interference paths between such channels, respectively, the phase characteristics being irrelevant. This situation simplifies the analysis as well as the experimentation considerably.

As a matter of fact the present note was also motivated by a previous study in this field [5].

Notice that the described relation between the second-moment properties of the interferring processes at the input and the output holds specifically in the asymptotic situation. However, the exact relation between ε and η is nonlinear, and it is well known, that for nonlinear transformations the knowledge of the correlation (or spectral distribution) function at the input is, in general, insufficient for specifying the correlation function at the output.

REMARK 3. Observe that $B_\nu(\tau)=\overline{B_\nu(\tau)}$, for all τ, if, for instance, μ is a stationary Gaussian process. It is well known that angle-modulation channels carrying multichannel telephony may be successfully studied by considering such μ. However, this is also the case when μ is generated by scanning a picture (or some similar procedure) and one may suppose B_ν to remain unaltered when reversing the sequence of the scanning. (Since, from $B_\nu(\tau)=B_\nu(-\tau)$ and the stationarity of ν it follows that: $B_\nu(\tau)=\overline{B_\nu(\tau)}$.)

REFERENCES

[1] BENETT, W. R., CURTIS, H. E. and RICE, S. O.: Interchannel Interference in FM and PM Systems Under Noise Loading Conditions, *Bell System Tech. J.*, **34** (1955) May.

[2] BORODIČ, S. V.: Computation of Allowable Radio Interference in Multichannel Radio-Relay Systems, *Telecommunication* No. 1, 1962.

[3] MEDHURST, R. C., Mrs. HICKS, E. M. and GROSSET, W.: Distortion in Frequency Division Multiplex FM Systems Due to an Interferring Carrier, *Proc. I. E. E.*, **105**, Pt. B., (1958), May.

[4] РОЗАНОВ, Ю. А.: *Стационарные случайные процессы*, Физматгиз, 1963., гл. I.

[5] CSIBI, S.: Noise in FM Radio Systems Due to Radio Interference, Dissertation, Hung. Acad. Sci., 1960. Ch. 3. (in Hungarian).

[6] CSIBI, S.: Direct Interference Between Closely Spaced FDM-FM Radio Channels, *Proc. 2nd Coll. Microwave Communication*, Budapest 1962, Publ. House, Hung. Acad. Sci., 1963.

RESEARCH INSTITUTE FOR TELECOMMUNICATION, BUDAPEST

(Received February 3, 1965.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 159—161.

A REMARK TO A PAPER OF L. SCHMETTERER

by

A. KRÁMLI

In the present note we shall generalize a theorem of L. SCHMETTERER [1]. The new form of this theorem, — and naturally also its proof — doesn't involve the notion of Gâteaux differential, the proof is somewhat simpler.

Let R be a set and S a σ-algebra of its subsets. Let $\mathfrak{P}$ be a non-void family of probability measures P over S and g a real function defined on $\mathfrak{P}$. A real S-measurable function h on R is an unbiased estimate of g if and only if $\int_R h dP = g(P)$ for every $P \in \mathfrak{P}$.

Let us denote by H_g the set of all unbiased estimates of g.

Let us make correspond to every $P \in \mathfrak{P}$ a Banach space B_P of measurable functions on R with the norm N_P. Suppose that $c \in B_P$ for every $P \in \mathfrak{P}$, where c is an arbitrary constant real valued function on R. We shall make use of the definitions given by SCHMETTERER.

An estimate $h_0 \in H_g \cap \bigcap_{P \in \mathfrak{P}} B_P$ is uniformly N_P-minimal if and only if $N_P(h_0 - g(P)) \leqq N_P(h - g(P))$ for every $h \in H_g \cap \bigcap_{P \in \mathfrak{P}} B_P$ and $P \in \mathfrak{P}$. (To simplify the writing further we shall denote $\bigcap_{P \in \mathfrak{P}} B_P$ by B.)

Further: An estimate $h_0 \in H_g \cap B_{P_0}$ is locally N_{P_0}-minimal if and only if $N_{P_0}(h_0 - g(P_0)) \leqq N_{P_0}(h - g(P_0))$ for every $h \in H_g \cap B_{P_0}$.

Denote by V the set of all measurable functions v on R such that $\int_R v dP = 0$ for every $P \in \mathfrak{P}$ (the unbiased estimates of the identically zero mapping). Obviously V is a linear manifold and every $h \in H_g$ assumes the form $h = h_0 + v$, where h_0 is a fixed element of H_g and $v \in V$.

THEOREM of SCHMETTERER: *When every B_P is smooth (namely the norm N_P is a Gâteux differentiable function on B_P) the estimate $h_0 \in H_g \cap B$ is uniformly N_P-minimal if and only if the Gâteaux differential L_P of the norm N_P vanishes at $h_0 - g(P)$ for every vector $v \in V \cap B$ and every $P \in \mathfrak{P}$ $(L_P(h_0 - g(P), v) = 0)$.*

From the HAHN—BANACH theorem there follows the existence of a linear functional l_P on B_P such that:

$$(*) \qquad l_P(h_0 - g(P)) = N_P(h_0 - g(P)) \quad \text{and} \quad |l_P(h - g(P))| \leqq N_P(h - g(P))$$

for every $h \in B_P$.

Now we give SCHMETTERER's condition without applying the notion of the Gâteaux differential, and this condition will be valid in non-smooth Banach spaces, too.

THEOREM: *An estimate $h_0 \in H_g \cap B$ is uniformly N_P-minimal if and only if for every $P \in \mathfrak{P}$ there exists a linear functional l_P of property $(*)$ and such that $l_P(v) = 0$ for all $v \in V \cap B$.*

REMARK: It is well known that every convex subset of a Banach space has a supporting hyperplane at every point of its boundary. A hyperplane may be always defined by an equation of the form $l(x) = c$, where $l(x)$ is a linear functional and c is a real number. Moreover this functional is uniquely determined by the hyperplane, apart from a multiplicative constant. The relations $(*)$ mean that the equation $l_P(x) = N_P(h_0 - g(P))$ defines a supporting hyperplane of the set $\{x : x \in B_P, N_P(x) \leqq \leqq N_P(h_0 - g(P))\}$ at the point $h_0 - g(P)$.

If the norm N_P is Gâteaux differentiable, then there exists one and only one such supporting hyperplane, and it is defined by the equation $L_P(h_0 - g(P), x - (h_0 - g(P))) = 0$ and — as L_P is linear — by $L_P(h_0 - g(P), x) = L_P(h_0 - g(P), h_0 - g(P)) = c$.

So $l_P(x)$ and $L_P(h_0 - g(P), x)$ may differ in a non-zero multiplicative constant only. Consequently, our theorem includes the SCHMETTERER's one.

PROOF:

a) Necessity: Let h_0 be a uniformly N_P-minimal unbiased estimate of g. Denote by V_P the closure of the linear manifold $V \cap B$ in the sense of the norm N_P.

Let $l_P^{(0)}$ be the identically vanishing functional defined on V_P. We can extend this functional to a functional l'_P over the Banach space $\langle V_P, h_0 - g(P) \rangle$ generated by V_P and $h_0 - g(P)$ so that $l'_P(h_0 - g(P)) = N_P(h_0 - g(P))$ and $|l'_P(h)| \leqq N_P(h)$ if $h \in \langle V_P, h_0 - g(P) \rangle$. In fact, it follows from the well-known proof of the HAHN—BANACH theorem (its idea is due to HELLY; see e. g. [2]) that $l'_P(h_0 - g(P))$ may have an arbitrary value between the two numbers

$$m = \sup_{v \in V_P} \{-N_P(v + (h_0 - g(P)) - l_P(v)\} \quad \text{and} \quad M = \inf_{v \in V_P} \{N_P(v + (h_0 - g(P)) - l_P(v)\}.$$

Since h_0 is a minimal estimate we have

$$-N_P(v + (h_0 - g(P))) \leqq N_P(h_0 - g(P)) \leqq N_P(v + (h_0 - g(P)))$$

for every $v \in V_P$. Considering that $l(v) = 0 (v \in V_P)$ we can write $m \leqq N_P(h_0 - g(P)) \leqq M$, and this fact proves the existence of the desired extension of the functional l'_P. Then there exists an extension of l'_P over the whole B_P to a functional l_P of property $(*)$, and such that $l_P(v) = 0$ for every $v \in V \cap B$.

Applying this extension procedure to every $P \in \mathfrak{P}$, we get the proof of the necessity of the condition.

b) The *sufficiency* is obvious, since, by the condition, for every $P \in \mathfrak{P}$ there exists a linear functional l_P of property $(*)$, so that $l_P(v) = 0$ $(v \in V \cap B)$ and consequently:

$$N_P(h_0 - g(P)) = l_P(h_0 - g(P)) = l_P(h_0 + v - g(P)) \leqq N_P(h_0 + v - g(P)) =$$

$$= N_P(h - g(P))$$

for every $h \in H_g \cap B$.

An analogous theorem is valid for locally N_P minimal unbiased estimates.

REFERENCES

[1] SCHMETTERER, L.: Über eine allgemeine Theorie der erwartungstreuen Schätzungen. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961) 295.

[2] RIESZ, F. et SZŐKEFALVI-NAGY, B.: *Leçons d'Analyse Fonctionnelle,* Budapest, 1953, p. 112.

JÓZSEF A. UNIVERSITY, SZEGED

(Received June 16, 1965.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 163—166.

ON *R*-PRODUCTS OF AUTOMATA, III

by

F. GÉCSEG

In [2], we dealt with the quasi-direct product and quasi-superposition of automata from the point of view of the metrical completeness. It was shown that there exists a finite system of finite automata which is metrically complete with respect to the quasi-superposition and that there exists no finite system of finite automata which is metrically complete with respect to the quasi-direct product.

In this part of the work we shall prove the existence of an infinite minimal system which is metrically complete with respect to the quasi-direct product. Furthermore, we shall study the question of when an automaton-map φ can be induced in length k by a quasi-direct product of automata such that the number of states of these automata is smaller than the weight of φ.

For the notions and notations not defined in this paper, see [1] and [2].

Before studying the above mentioned questions, we shall introduce a notion. Let $\mathfrak{A}=\langle \mathbf{A}_i, i=1, 2, \ldots\rangle$ be a system which is metrically complete with respect to the quasi-direct product. We say that $\mathfrak{A}$ is *minimal* if for arbitrary $\mathbf{A}_i \in \mathfrak{A}$, the set $\mathfrak{A}'=\mathfrak{A}\setminus\mathbf{A}_i$ is not metrically complete with respect to quasi-direct product.

We shall prove the following

THEOREM 1. *There exists a system of finite automata which is metrically complete with respect to quasi-direct product and minimal.*

PROOF. Let k be an arbitrary natural number and $X=\langle x_1, \ldots, x_n\rangle$ $(n>1)$ an arbitrary finite set. Denote by $\mathbf{A}^{(n,k)}=\mathbf{A}^{(n,k)}(X, A^{(n,k)}, a_0, \delta_k)$ the following Medvedev automaton:

$$A^{(n,k)}=\langle a_0, a_1, \ldots, a_{n+k}\rangle \tag{1}$$

$$\delta_k(a_i, x_j)=\begin{cases} a_{i+1}, & \text{if } i<k \\ a_{k+j}, & \text{if } i=k \\ a_i, & \text{if } i>k. \end{cases} \tag{2}$$

We show that the set $\mathfrak{A}_n$ of automata $\mathbf{A}^{(n,k)}$ $(k=0, 1, 2, \ldots)$ is metrically complete system with respect to quasi-direct product and minimal. First we shall prove that the set $\mathfrak{A}_n$ is metrically complete with respect to the quasi-direct product. Since the system of all k-free $(k=1, 2, \ldots)$ MEDVEDEV automata with $n(\geqq 2)$ as number of inputs is metrically complete (see [1]) with respect to the quasi-direct product, so it is sufficient to show that each k-free automaton $\mathbf{S}=\mathbf{S}(X, S, s_0, \delta_S)$ is k-isomorphic to a quasi-direct product of automata from $\mathfrak{A}_n$. If $k=1$, then $\mathbf{S}$ is k-isomorphic to $\mathbf{A}^{(n,0)}$. Let us suppose that our assertion is already proved for $k-1(\geqq 1)$ and let $\mathbf{B}=\mathbf{B}(X, B, b_0, \delta)$ be a quasi-direct product of automata from $\mathfrak{A}_n$ which

is $k-1$-isomorphic to $\mathbf{S}$. Let ϑ denote this $k-1$-isomorphism. We shall prove that the quasi-direct product $\mathbf{B}'=\mathbf{B}'(X, B', b_0', \delta')$ of automata $\mathbf{B}$ and $\mathbf{A}^{(n,k-1)}$ for which

$$B'=B\times A^{(n,k-1)}$$

and

$$\delta'((b,a),x)=(\delta(b,x),\ \delta_{k-1}(a,x))\ \ (b\in B,\ a\in A^{(n,k-1)})$$

hold is k-isomorphic to $\mathbf{S}$.

Let $\vartheta'(:S^{(k)}\to B'^{(k)})$ be the following map:

$$\vartheta'(s_0p)=(\vartheta(s_0p),\ a_0p)\ (l(p)<k,\ p\in F(X)),$$

$$\vartheta'(s_0px)=(\vartheta(s_0p),\ a_0px)\ (l(p)=k-1,\ p\in F(X)).$$

It is obvious that ϑ' is a $k-1$-isomorphism on the set $S^{(k-1)}$. So we have still to prove that for arbitrary $p\in F(X)$ with $l(p)=k-1$ and for arbitrary $x_i, x_j\in X$ $i\neq j$ implies $\vartheta'(s_0px_i)\neq\vartheta'(s_0px_j)$. This is true, because $a_0p=a_{k-1}$, so

$$\vartheta'(s_0px_i)=(\vartheta(s_0p),\ a_{k-1}x_i)=(\vartheta(s_0p),\ a_{k-1+i})$$

and

$$\vartheta'(s_0px_j)=(\vartheta(s_0p),\ a_{k-1}x_j)=(\vartheta(s_0p),\ a_{k-1+j}).$$

But $i\neq j$ and (2) imply $a_{k-1+i}\neq a_{k-1+j}$, that is $\vartheta'(s_0px_i)\neq\vartheta'(s_0px_j)$. So we proved that $\mathfrak{A}_n$ is metrically complete with respect to quasi-direct product.

Now we show that the system $\mathfrak{A}_n$ is minimal i.e. the set $\mathfrak{A}'=\mathfrak{A}_n\backslash\mathbf{A}^{(n,k)}$ for arbitrary $\mathbf{A}^{(n,k)}\in\mathfrak{A}_n$ is not metrically complete with respect to the quasi-direct product. To do this it is sufficient to prove that the map φ induced by arbitrary l-free $(l>k)$ automaton $\mathbf{A}^{(l)}=\mathbf{A}^{(l)}(X, A^{(l)}, a_0^{(l)}, \delta^{(l)})$ can be induced in length l by no quasi-direct product $\mathbf{A}=\mathbf{A}(X, A, a_0, \delta)$ of automata from $\mathfrak{A}_n'$, that is an arbitrary quasi-direct product $\mathbf{A}=\mathbf{A}(X, A, a_0, \delta)$ is not l-free. Indeed, if $l(p)=k$ then for the state a_0p $(p\in F(X))$ of automaton $\mathbf{A}$ and for arbitrary inputs $x_i, x_j\in X$ the equation $a_0px_i=a_0px_j$ holds, i.e. the automaton $\mathbf{A}$ is not l-free if $l>k$. This ends the proof of the Theorem 1.

We note that *a system which is metrically complete with respect to quasi-direct product need not have a minimal subsystem which is metrically complete with respect to the quasi-direct product.* Indeed, let $\mathfrak{A}$ be the set of all finite k-free automata $\mathbf{A}_k=\mathbf{A}_k(X, A_k, a_{k_0}, \delta_k)$ $(k=1, 2, \ldots)$ with fixed set X $(\overline{X}>1)$ of inputs. It can be seen easily that $\mathfrak{A}$ is metrically complete with respect to quasi-direct product and $\mathfrak{A}$ has no minimal subsystem which is metrically complete with respect to quasi-direct product.

Let $\mathbf{A}=\mathbf{A}(X, A, a_0, Y, \delta, \lambda)$ and $\mathbf{B}=\mathbf{B}(X, B, b_0, Y, \delta', \lambda')$ be arbitrary automata. Then a map $\vartheta: A^{(k)}\to B^{(k)}$ is said to be *k-homomorphism*, if

$$\left.\begin{array}{ll} \text{a)} & \vartheta(a_0)=b_0,\ \vartheta(\delta(a,x))=\delta'(\vartheta(a),x) \\ \text{and} & \\ \text{b)} & \lambda(a,x)=\lambda'(\vartheta(a),x) \end{array}\right\}\ (x\in X,\ a\in A^{(k-1)})$$

holds[1].

[1] The set $A^{(n)}$ will consist of the states of automaton $\mathbf{A}$, the heights of which do not exceed n (see [1]).

Let $\varphi: F(X) \to F(Y)$ be an automaton-map and let $\mathbf{S} = \mathbf{S}(X, S, s_0, Y, \delta_S, \lambda_S)$ be a k-free automaton, which induces φ in length k. A k-congruent partition π of $\mathbf{S}$ is called *natural partition* if $s \equiv s'(\pi)$ $(s, s' \in S^{(k)})$ implies $\lambda_S(s, p) = \lambda_S(s', p)$ $(p \in F(X))$ whenever $l(p) \leqq k - \max(h(s), h(s'))$.

It is clear that if ϑ is a k-homomorphism from a k-free automaton $\mathbf{S}$ into an automaton $\mathbf{A}$, then the following k-congruent partition π is natural:

$$s \equiv s'(\pi) \Leftrightarrow \vartheta(s) = \vartheta(s') \qquad (s, s' \in S^{(k)}).$$

The number of classes of a k-congruent partition π will be denoted by $|\pi|$.

The number of states of a minimal automaton $\mathbf{A}$ which induces φ is called *weight* of φ and it is denoted by $w(\varphi)$.

We have the following

THEOREM 2. *An automaton-map φ can be induced in length k by a quasi-direct product of Medvedev automata $\mathbf{A}_i$ $(i = 1, \ldots, r)$ such that $\bar{A}_i < w(\varphi)$ $(1 \leqq i \leqq r)$ if and only if there exists a k-free automaton $\mathbf{S}$ inducing φ in length k, which has k-congruent partitions $\pi_1, \ldots, \pi_r$ and the natural partition π satisfying the following conditions:*

(I) $\pi_i > \pi$ *for each* i $(1 \leqq i \leqq r)$,

(II) $|\pi_i| < w(\varphi)$ *for each* i $(1 \leqq i \leqq r)$,

(III) $\bigcap_{i=1}^{r} \pi_i = \pi$.

PROOF. To prove the necessity of our conditions suppose that there exists a quasi-direct product $\mathbf{A} = \mathbf{A}(X, A, a_0, Y, \delta, \lambda)$ of Medvedev automata $\mathbf{A}_i = \mathbf{A}_i(X_i, A_i, a_{i_0}, \delta_i)$ $(i = 1, \ldots, r)$ with $\bar{A}_i < w(\varphi)$ $(1 \leqq i \leqq r)$, such that $\mathbf{A}$ induces $\varphi: F(X) \to F(Y)$ in length k.

Let $S = S(X, S, s_0, Y, \delta_S, \lambda_S)$ be arbitrary k-free automaton inducing φ in length k and let ϑ denote a map of $S^{(k)}$ onto $A^{(k)}$ for which

$$\vartheta(s_0) = a_0 \tag{3}$$

and

$$\vartheta(s_0 p) = a_0 p \qquad (p \in F(X),\ l(p) \leqq k) \tag{4}$$

hold. It can be seen easily that ϑ is k-homomorphism. Let us consider the partition π of $S^{(n)}$ for which

$$s \equiv s'(\pi) \Leftrightarrow \vartheta(s) = \vartheta(s') \qquad (s, s' \in S^{(k)})$$

holds. Because ϑ is k-homomorphism so π is natural k-congruent partition.

For arbitrary i $(1 \leqq i \leqq r)$ let π_i' denote the partition of $A^{(k)}$ for which

$$a(=(a_1, \ldots, a_i, \ldots, a_r)) \equiv ((a_1', \ldots, a_i', \ldots, a_r') =) a'(\pi_i') \Leftrightarrow a_i = a_i'$$

holds. It is clear that $|\pi_i'| = \bar{A}_i$ for each i $(1 \leqq i \leqq r)$. Let π_i $(1 \leqq i \leqq r)$ be the following partition of $S^{(k)}$:

$$s \equiv s'(\pi_i) \Leftrightarrow \vartheta(s) \equiv \vartheta(s')\,(\pi_i') \qquad (s, s' \in S^{(k)}).$$

It is easy to show, that the k-congruent partitions $\pi, \pi_1, \ldots, \pi_r$ satisfy the conditions I—III.

Conversely, suppose that the conditions I—III hold. We shall construct for each $\pi_i\,(1 \leqq i \leqq r)$ an automaton $\mathbf{A}_i = \mathbf{A}_i(X, A_i, a_{i_0}, \delta_i)$ in the following way:

$$A_i = \langle \pi_i(s) | s \in S^{(k)} \rangle,$$

$$a_{i_0} = \pi_i(s_0),$$

$$\delta_i(\pi_i(s), x) = \begin{cases} \pi_i(\delta_S(s', x)) & \text{if there exists an } s' \in S^{(k-1)} \text{ such that } s \equiv s'(\pi_i) \\ \text{arbitrary } a_i \in A_i & \text{in the contrary case.} \end{cases}$$

It is clear that the number of states of an arbitrary automaton $\mathbf{A}_i$ $(1 \leqq i \leqq r)$ is not greater than $|\pi_i| (< w(\varphi))$.

Let $\mathbf{A} = \mathbf{A}(X, A, a_0, Y, \delta, \lambda)$ be a quasi-direct product of automata $\mathbf{A}_i\,(i = 1, \ldots, r)$ for which

$$\delta((a_1, \ldots, a_r), x) = (\delta_1(a_1, x), \ldots, \delta_r(a_r, x)) \tag{5}$$

and

$$\lambda((a_1, \ldots, a_r), x) = \begin{cases} \lambda_S(s, x) \text{ if there exists an } s \in \bigcap_{i=1}^{r} a_i \text{ such that } l(s) < k \\ \text{arbitrary } y \in Y \text{ in the opposite case.} \end{cases} \tag{6}$$

We shall prove by induction on length of words $p (\in F(X), l(p) \leqq k)$ that the automaton $\mathbf{A}$ induces φ in length k, i.e.

$$\varphi(p) = \lambda(a_0, p) \tag{7}$$

whenever $l(p) \leqq k$. Let $l(p) = 1$, i.e. $p = x \in X$, then $\lambda(a_0, x) = \lambda_S(s_0, x) = \varphi(x)$. Suppose that (7) holds whenever $l(p) \leqq j-1 (< k)$, we want to point out (7) if $l(p) = j$. p can be written as $p'x\,(l(p') = j-1)$. We have

$$\lambda(a_0, p'x) = \lambda(a_0, p') \cdot \lambda(a_0 p', x)$$

and

$$\lambda_S(s_0, p'x) = \lambda_S(s_0, p') \cdot \lambda_S(s_0 p', x).$$

The first factors on the right sides are equal by the induction hypothesis. But $s_0 p \in \bigcap_{i=1}^{r} a_i$, where $(a_1, \ldots, a_r) = (\pi_1(s_0 p), \ldots, \pi_r(s_0, p)) = a_0 p$, so by (6) $\lambda(a_0 p, x) = \lambda_S(s_0 p, x)$. This completes the proof of the Theorem 2.

REFERENCES

[1] GÉCSEG, F.: On R-products of automata, I, *Studia Sci. Math. Hung.* **1** (1966) 437—441.

[2] GÉCSEG, F.: On R-products of automata, II, *Studia Sci. Math. Hung.* **1** (1966) 443—447.

JÓZSEF A. UNIVERSITY, SZEGED

(Received June 15, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 167—183.

ON A POLYNOMIAL OF INTERPOLATION[1]

by

R. B. SAXENA

1. Very recently Professor Géza FREUD [1] has considered an interpolation process with roots as the zeros of Čebyšev polynomial of first kind $T_n(x) = \cos n(\arccos x)$ which gives directly a proof of JACKSON's Theorem. Later M. SALLAY [2] has solved the same problem with abscissas as the zeros of the orthogonal polynomial with weight function which is positive on $[-1, 1]$ and satisfies a Lipschitz condition of order 1. In this paper we solve the same problem with another interpolation process, constructed on the roots of Čebyšev polynomial of second kind

$$U_n(x) = \frac{\sin(n+1)\theta}{\sin\theta}, \quad \cos\theta = x,$$

which is a modified form of the process considered by Professor FREUD.

2. Let

$$x_{kn} = \cos\frac{k\pi}{n+1}, \qquad k = 1, 2, \ldots, n$$

be the zeros of

$$U_n(x) = \frac{\sin(n+1)\theta}{\sin\theta}, \quad x = \cos\theta, \qquad n = 1, 2, \ldots$$

the Čebyšev polynomial of second kind. For the fundamental polynomial of Lagrange interpolation we have the expression

$$l_{kn}(x) = \frac{(-1)^{k+1}(1-x_{kn}^2)}{n+1} \cdot \frac{U_n(x)}{x - x_{kn}}. \tag{2.1}$$

Further let

$$v_{kn}(x) \overset{\text{def}}{=} 1 - \frac{3x_{kn}(x-x_{kn})}{1-x_{kn}^2} \tag{2.2}$$

and

$$\psi_n(t, u) \overset{\text{def}}{=} \frac{2}{n+1} \sum_{k=1}^{n-1}{}' U_r'(t)\, U_r(u). \tag{2.3}$$

We then denote

$$\lambda_{kn}(x) \overset{\text{def}}{=} \left(\frac{1-x^2}{1-x_{kn}^2}\right)^2 [v_{kn}(x) l_{kn}^4(x) + 2(x-x_{kn}) l_{kn}^3(x)(1-x_{kn}^2)\psi_n(x_{kn}, x)] \tag{2.4}$$

[1] This research has been supported by the National Research Council (N. R. C.) Grant MCA-41 to the Department of Mathematics, University of Alberta, Edmonton, (Canada).

the fundamental polynomials of degree $4n+1$ of our interpolation process. Let $f(x)$ be an arbitrary function defined for $-1 \leqq x \leqq 1$, then we consider the following interpolation polynomial of degree $4n+2$:[2]

$$A_n(f,x) = \frac{1+x}{2} f(1) + \frac{1-x}{2} f(-1) + \tag{2.5}$$
$$+ \sum_{k=1}^{n}{}' \left[f(x_{kn}) - \left\{ \frac{1+x}{2} f(1) + \frac{1-x}{2} f(-1) \right\} \right] \lambda_{kn}(x).$$

We shall prove the following

THEOREM. *Let $f(x)$ be a continuous function defined in $[-1, 1]$; then for the sequence of polynomials $A_n(f, x)$ in (2. 5) we have*

$$|A_n(f,x) - f(x)| \leqq 414\omega\left(f; \frac{1}{n}\right)$$

uniformly in $[-1, 1]$ where $\omega(f, \delta)$ is the modulus of continuity of $f(x)$.

3. For the proof of our theorem we shall need a number of auxiliary Lemmas. In this section we shall estimate the quantity

$$\left| 1 - \sum_{k=1}^{n} \lambda_{kn}(x) \right|.$$

LEMMA 3. 1. *If*

$$\Phi_{n-1}(t,u) = \frac{2}{n+1}\left[1 + \sum_{r=1}^{n-1} U_r(t)\, U_r(u)\right] \tag{3.1}$$

then[3]

$$\sum_{k=1}^{n} \lambda_k(x) = [(1-x^2)\, \Phi_{n-1}(x,x)]^2. \tag{3.2}$$

PROOF. We know the following CHRISTOFFEL—DARBOUX formula for $U_n(x)$

$$\frac{U_n(x)}{x - x_k} = (-1)^{k+1} 2\left[1 + \sum_{r=1}^{n-1} U_r(x_k)\, U_r(x)\right]$$

and on making use of (2. 1) we have

$$l_k(x) = \frac{2}{n+1}(1-x_k^2)\left[1 + \sum_{r=1}^{n-1} U_r(x_k)\, U_r(x)\right], \tag{3.3}$$

i.e. owing to (3. 1)

$$l_k(x) = (1-x_k^2)\, \Phi_{n-1}(x_k, x). \tag{3.4}$$

[2] The polynomials $A_n(f, x)$ take the values $f(x_k)$ at $x = x_k$ $(k = 0, 1, 2, \ldots, n+1)$ where $x_0 = 1$, $x_{n+1} = -1$.

[3] For simplicity in writing we shall denote the suffixes kn by k and thus write x_k for x_{kn}, $l_k(x)$ for $l_{kn}(x)$ etc.

Now the HERMITE—FEJÉR polynomial of degree $2n+3$ for any function $\varphi(x)$ constructed on the points

$$x_k = \cos\frac{k\pi}{n+1}, \qquad k = 0, 1, 2, \ldots, n+1 \quad [x_0 = 1,\ x_{n+1} = -1]$$

is given by

$$H_n(\varphi, x) = \varphi(1)\left(\frac{1+x}{2}\right)^2\left[1+\frac{2n^2+2n+3}{3}(1-x)\right]\left(\frac{U_n(x)}{n+1}\right)^2 +$$

$$+\varphi(-1)\left(\frac{1-x}{2}\right)^2\left[1+\frac{2n^2+2n+3}{3}(1+x)\right]\left(\frac{U_n(x)}{n+1}\right)^2 +$$

$$+\sum_{k=1}^{n}\varphi(x_k)\left[1+\frac{x_k(x-x_k)}{1-x_k^2}\right]l_k^2(x)\left(\frac{1-x^2}{1-x_k^2}\right)^2+\varphi'(1)\frac{(1+x)(x^2-1)}{4(n+1)^2}U_n^2(x)+$$

$$+\varphi'(-1)\frac{(1-x)(1-x^2)}{4(n+1)^2}U_n^2(x)+\sum_{k=1}^{n}\varphi'(x_k)\left(\frac{1-x^2}{1-x_k^2}\right)^2(x-x_k)l_k^2(x).$$

From this we have for any arbitrary polynomial $P_{2n+2}(x)$ of degree $(2n+2)$ the identity:

$$(3.5)\qquad P_{2n+2}(x) \equiv P_{2n+2}(1)\left[1+\frac{2n^2+2n+3}{3}(1-x)\right]\left(\frac{1+x}{2}\right)^2\left(\frac{U_n(x)}{n+1}\right)^2 +$$

$$+P_{2n+2}(-1)\left[1+\frac{2n^2+2n+3}{3}(1+x)\right]\left(\frac{1-x}{2}\right)^2\left(\frac{U_n(x)}{n+1}\right)^2 +$$

$$+\sum_{k=1}^{n}P_{2n+2}(x_k)\left[1+\frac{x_k(x-x_k)}{1-x_k^2}\right]\left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^2(x)+P'_{2n+2}(1)\frac{(1+x)(x^2-1)}{4(n+1)^2}U_n^2(x)+$$

$$+P'_{2n+2}(-1)\frac{(1-x)(1-x^2)}{4(n+1)^2}U_n^2(x)+\sum_{k=1}^{n}P'_{2n+2}(x_k)\left(\frac{1-x^2}{1-x_k^2}\right)^2(x-x_k)l_k^2(x).$$

Let us take

$$P_{2n+2}(x) = [(1-x^2)\Phi_{n-1}(x, \xi)]^2$$

so that

$$P_{2n+2}(x_k) = l_k^2(\xi)$$

and

$$P'_{2n+2}(x_k) = 2(1-x_k^2)l_k(\xi)\psi_n(x_k, \xi) - \frac{4x_k}{1-x_k^2}l^2(\xi).$$

Hence from (3. 5) we have

$$[(1-x^2)\Phi_{n-1}(x, \xi)]^2 =$$

$$=\sum_{k=1}^{n}\left(\frac{1-x^2}{1-x_k^2}\right)^2\left[\left\{1-\frac{3x_k(x-x_k)}{1-x_k^2}\right\}l_k^2(x)l_k^2(\xi)+2(x-x_k)(1-x_k^2)l_k^2(x)\psi_n(x_k, \xi)\right],$$

and putting $\xi = x$ we have the lemma.

LEMMA 3. 2. *For* $-1 \leqq x \leqq 1$ *we have*

$$\left| 1 - \sum_{k=1}^{n} \lambda_k(x) \right| \leqq 3,$$

$$\sqrt{1-x^2} \left| 1 - \sum_{k=1}^{n} \lambda_k(x) \right| < \frac{3}{n}.$$

PROOF. From Lemma 3. 1 we have

$$\sum_{k=1}^{n} \lambda_k(x) = [(1-x^2)\Phi_{n-1}(x)]^2.$$

But from (3. 1)

$$(1-x^2)\Phi_{n-1}(x,x) = \frac{2(1-x^2)}{n+1}\left[1+\sum_{r=1}^{n-1} U_r^2(x)\right] =$$

$$= \frac{1}{n+1}\left[2\sin^2\theta + \sum_{r=1}^{n-1} 2\sin^2(r+1)\theta\right] = \frac{1}{n+1}\left[2\sin^2\theta + \sum_{k=1}^{n-1}(1-\cos(2r+2)\theta)\right] =$$

$$= \frac{1}{n+1}\left[n - \sum_{r=1}^{n}\cos 2r\theta\right] = 1 - \frac{1}{2n+2}\left[1+\frac{\sin(2n+1)\theta}{\sin\theta}\right].$$

Hence

$$1-[(1-x^2)\Phi_{n-1}(x,x)]^2 = \frac{\sin(n+1)\theta\cos n\theta}{(n+1)\sin\theta}\left[2-\frac{\sin(n+1)\theta\cos n\theta}{(n+1)\sin\theta}\right].$$

From this and Lemma 3. 1 we have

$$1-\sum_{k=1}^{n}\lambda_k(x) = \frac{\sin(n+1)\cos n\theta}{(n+1)\sin\theta}\left[2-\frac{\sin(n+1)\theta\cos n\theta}{(n+1)\sin\theta}\right].$$

Now

$$|\sin(n+1)\theta| \leqq (n+1)\sin\theta$$

$$|\cos n\theta| \leqq 1,$$

therefore

$$\left|1-\sum_{k=1}^{n}\lambda_k(x)\right| \leqq \frac{3}{n+1}\cdot\frac{|\sin(n+1)\theta|}{\sin\theta} \leqq 3$$

and

$$\sqrt{1-x^2}\left|1-\sum_{k=1}^{n}\lambda_k(x)\right| \leqq \frac{3}{n+1} < \frac{3}{n}$$

which proves the lemma.

4. In this article we shall estimate the sum

$$\sum_{k=1}^{n}\left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x)|x-x_k|.$$

Let us denote by E_k the expression

$$E_k \overset{\text{def}}{=} \left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x)|x-x_k|, \qquad 1 \leqq k \leqq n. \tag{4.1}$$

We write

$$\theta = \arccos x \qquad (-1 \leqq x \leqq 1)$$

$$\theta_i = \frac{i\pi}{n+1} \qquad (i = 0, 1, 2, \ldots, n+1)$$

then owing to (2. 1)

$$E_k = \frac{(1-x^2)^2 U_n^4(x)(1-x_k^2)^2}{(n+1)^4|x-x_k|^3} = \frac{\sin^4(n+1)\theta \sin^4\theta_k}{(n+1)^4|\cos\theta-\cos\theta_k|^3}. \tag{4.2}$$

We prove the following lemmas:

LEMMA 4. 1. *For* $1 \leqq k < i \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$E_k \leqq \frac{1}{n+1} \cdot \frac{1}{(i-k)^3}.$$

PROOF. Since

$$\sin\theta_k \leqq \sin\theta_k + \sin\theta = 2\sin\frac{\theta+\theta_k}{2}\cos\frac{\theta-\theta_k}{2} \leqq 2\sin\frac{\theta+\theta_k}{2}$$

$$|\cos\theta-\cos\theta_k| = 2\sin\frac{\theta+\theta_k}{2}\left|\sin\frac{\theta-\theta_k}{2}\right|, \tag{4.3}$$

$$\frac{\sin\theta_k}{|\cos\theta-\cos\theta_k|} \leqq \frac{1}{\left|\sin\frac{\theta-\theta_k}{2}\right|},$$

$$|\sin nt| \leqq 1,$$

therefore from (4. 2) we have

$$E_k \leqq \frac{1}{(n+1)^4} \cdot \frac{1}{\left|\sin\frac{\theta-\theta_k}{2}\right|^3}. \tag{4.4}$$

If

$$x_{i+1} \leqq x \leqq x_i$$

then

$$\theta_i \leqq \theta \leqq \theta_{i+1}.$$

In the case

$$1 \leqq k < i \leqq n-1,$$

$$\theta_i - \theta_k \leqq \theta - \theta_k \leqq \pi,$$

therefore,

$$\sin\frac{\theta-\theta_k}{2} \geqq \sin\frac{\theta_i-\theta_k}{2} = \sin\frac{i-k}{2(n+1)}\pi > \frac{i-k}{n+1}.$$

From this and (4. 4) we have the lemma.

LEMMA 4. 2. *If* $0 \leqq i \leqq n-2$, $i+2 \leqq k \leqq n$, *then in* $[x_{i+1}, x_i]$

$$E_k \leqq \frac{1}{(n+1)} \cdot \frac{1}{(k-i-1)^3}$$

PROOF. Here

$$\theta_k - \theta_{i+1} \leqq \theta_k - \theta \leqq \pi,$$

$$\sin \frac{\theta_k - \theta}{2} \geqq \sin \frac{\theta_k - \theta_{i+1}}{2} > \frac{k-i-1}{n+1}.$$

From this and (4. 4) follows the lemma.

LEMMA 4.3. *For* $1 \leqq i \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$E_i \leqq \frac{8}{n+1}.$$

PROOF. From (4. 2)

$$E_i = \frac{\sin^4 (n+1)\theta \sin^4 \theta_i}{(n+1)^4 |\cos \theta - \cos \theta_i|^3} \leqq \frac{|\sin^3 (n+1)\theta| \sin^3 \theta_i}{(n+1)^4 |\cos \theta - \cos \theta_i|^3} \tag{4. 5}$$

since

$$|\sin (n+1)\theta| = |\sin (n+1)\theta - \sin (n+1)\theta_i| = \left| 2 \sin (n+1) \frac{\theta - \theta_i}{2} \cos (n+1) \frac{\theta + \theta_i}{2} \right| \leqq$$

$$\leqq 2 \left| \sin (n+1) \frac{\theta - \theta_i}{2} \right| \leqq 2(n+1) \left| \sin \frac{\theta - \theta_i}{2} \right|, \tag{4. 6}$$

then from (4. 3) follows the inequality

$$\frac{|\sin (n+1)\theta| \sin \theta_i}{|\cos \theta - \cos \theta_i|} \leqq 2(n+1). \tag{4. 7}$$

Thus from this and (4. 5) follows the lemma.

LEMMA 4. 5. *If* $0 \leqq i \leqq n-1$, *then on* $[x_{i+1}, x_i]$

$$E_{i+1} \leqq \frac{8}{n+1}.$$

The PROOF follows analogous to Lemma 4. 3. Using the above lemmas we shall now prove the main lemma of this article.

LEMMA 4. 5. *For* $-1 \leqq x \leqq 1$ *we have*

$$\sum_{k=1}^{n} \left(\frac{1-x^2}{1-x_k^2} \right)^2 l_k^4(x) |x - x_k| \leqq \frac{20}{n+1}.$$

PROOF. We break the sum

$$\sum_{k=1}^{n} \left(\frac{1-x^2}{1-x_k^2} \right)^2 l_k^4(x) |x - x_k| \equiv \sum_{k=1}^{n} E_k$$

as

$$\sum_{k=1}^{n} E_k = \sum_{k=1}^{i-1} E_k + E_i + E_{i+1} + \sum_{k=i+2}^{n} E_k.$$

Now using Lemmas 4. 1, 4. 2, 4. 3 and Lemma 4. 4, we have

$$\sum_{k=1}^{n} E_k < \frac{1}{(n+1)} \sum_{k=1}^{i-1} \frac{1}{(i-k)^3} + \frac{16}{n+1} + \frac{1}{(n+1)} \sum_{k=i+2}^{n} \frac{1}{(k-i-1)^3} <$$

$$< \frac{2}{n+1} \sum_{j=1}^{\infty} \frac{1}{j^3} + \frac{16}{n+1} < \frac{20}{n+1}.$$

5. In this article we shall estimate the sum

$$\sum_{k=1}^{n} \left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x) \frac{(x-x_k)^2}{(1-x_k^2)}.$$

Let us denote

$$E_k^* \overset{\text{def}}{=} \left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x) \frac{(x-x_k)^2}{1-x_k^2}, \qquad 1 \leqq k \leqq n \tag{5.1}$$

then owing to (2. 1) and (4. 3),

$$E_k^* = \frac{\sin^4 (n+1)\theta \sin^2 \theta_k}{(n+1)^4 (\cos\theta - \cos\theta_k)^2} \leqq \frac{1}{(n+1)^4 \sin^2 \frac{\theta-\theta_k}{2}}. \tag{5.2}$$

Without going into the details of the calculation, which will be obvious from the previous article, we have the following

LEMMA 5. 1. *For* $-1 \leqq x \leqq 1$

$$\sum_{k=1}^{n} \left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x) \frac{(x-x_k)^2}{1-x_k^2} \leqq \frac{12}{(1+n)^2}.$$

6. In this article we shall estimate the quantity

$$\sum_{k=1}^{n} \left(\frac{1-x^2}{1-x_k^2}\right)^2 |(1-x_k^2) l_k^3(x) \psi_n(x_k, x)| (x-x_k)^2.$$

We shall first prove

LEMMA 6. 1. *We have*

$$\sqrt{1-x^2}\, |\psi_n(x_k, x)| \leqq \frac{3}{|\sin^3 \theta_k|} + \frac{2}{\sin^2 \theta_k} \left[\frac{1}{|\sin \frac{1}{2}(\theta-\theta_k)|} + \frac{1}{\sin \frac{1}{2}(\theta+\theta_k)}\right].$$

PROOF. From (2. 3)

$$\psi_n(x_k, x) = \frac{2}{n+1} \sum_{r=1}^{n-1} U_r'(x_k) U_r(x) =$$

$$= \frac{2}{n+1} \sum_{r=1}^{n-1} \frac{\sin (r+1)\theta}{\sin \theta} \left\{\frac{x_k}{1-x_k^2} U_r(x_k) - \frac{r+1}{1-x_k^2} \cos (r+1)\theta_k\right\}.$$

Therefore

$$\sqrt{1-x^2}\,\psi_n(x_k, x) = \frac{2}{n+1}\sum_{r=1}^{n-1}\sin(r+1)\theta\left[\frac{\cos\theta_k \sin(r+1)\theta_k}{\sin^3\theta_k} - \right.$$

$$\left. -(r+1)\frac{\cos(r+1)\theta_k}{\sin^2\theta_k}\right] = \frac{2\cos\theta_k}{(n+1)\sin^3\theta_k}\sum_{r=1}^{n-1}\sin(r+1)\theta\sin(r+1)\theta_k -$$

$$-\frac{2}{(n+1)\sin^2\theta_k}\sum_{r=1}^{n-1}(r+1)\sin(r+1)\theta\cos(r+1)\theta_k =$$

$$= \frac{\cos\theta_k}{(n+1)\sin^3\theta_k}\sum_{r=1}^{n-1}[\cos(r+1)(\theta-\theta_k) - \cos(r+1)(\theta+\theta_k)] -$$

$$-\frac{1}{(n+1)\sin^2\theta_k}\sum_{r=1}^{n-1}(r+1)[\sin(r+1)(\theta-\theta_k) + \sin(r+1)(\theta+\theta_k)] =$$

$$= \frac{\cos\theta_k}{(n+1)\sin^3\theta_k}\left[\cos(\theta+\theta_k) - \cos(\theta-\theta_k) + \frac{1}{2}\left\{\frac{\sin\frac{1}{2}(2n+1)(\theta-\theta_k)}{\sin\frac{1}{2}(\theta-\theta_k)} - \right.\right.$$

$$\left.\left. -\frac{\sin\frac{1}{2}(2n+1)(\theta+\theta_k)}{\sin\frac{1}{2}(\theta+\theta_k)}\right\}\right] - \frac{1}{(n+1)\sin^2\theta_k}\left[\sin(\theta-\theta_k) + \sin(\theta+\theta_k) + \right.$$

$$+\frac{1}{4}\left\{\frac{\cos\frac{1}{2}(\theta-\theta_k)\sin\frac{1}{2}(2n+1)(\theta-\theta_k)}{\sin^2\frac{1}{2}(\theta-\theta_k)} + \frac{\cos\frac{1}{2}(\theta+\theta_k)\sin\frac{1}{2}(2n+1)(\theta+\theta_k)}{\sin^2\frac{1}{2}(\theta+\theta_k)}\right\} -$$

$$\left. -\frac{1}{4}(2n+1)\left\{\frac{\cos\frac{1}{2}(2n+1)(\theta-\theta_k)}{\sin\frac{1}{2}(\theta-\theta_k)} + \frac{\cos\frac{1}{2}(2n+1)(\theta+\theta_k)}{\sin\frac{1}{2}(\theta+\theta_k)}\right\}\right].$$

$$|\cos\theta| \leqq 1$$

Now

$$|\sin n\theta| \leqq n\sin\theta$$

(6. 1)
$$|\cos n\theta| \leqq 1$$

$$|\sin n\theta| \leqq 1.$$

We have

$$\sqrt{1-x^2}\,|\psi_n(x_k, x)| \leqq \frac{1}{(n+1)\sin^3\theta_k}[2+(2n+1)] +$$

$$+\frac{1}{(n+1)\sin^2\theta_k}\left[2+(2n+1)\left\{\frac{1}{\left|\sin\frac{\theta-\theta_k}{2}\right|} + \frac{1}{\sin\frac{\theta+\theta_k}{2}}\right\}\right] \leqq$$

$$\leqq \frac{3}{\sin^3\theta_k} + \frac{2}{\sin^2\theta_k}\left[\frac{1}{\left|\sin\frac{\theta-\theta_k}{2}\right|} + \frac{1}{\sin\frac{\theta+\theta_k}{2}}\right]$$

which proves our Lemma 6. 1.

Let us now denote

$$E_k^{**} \overset{\text{def}}{=} \left(\frac{1-x^2}{1-x_k^2}\right)^2 |(1-x_k^2) l_k^3(x) \psi_n(x_k, x)| (x-x_k)^2, \qquad 1 \leqq k \leqq n, \tag{6.2}$$

then owing to (2. 1)

$$E_k^{**} = \left| \frac{(-1)^{k+1}}{(n+1)^3} \cdot \frac{\sin^4 \theta_k \sin^3 (n+1)\theta}{\cos \theta - \cos \theta_k} \sqrt{1-x^2}\, \psi_n(x_k, x) \right|$$

and using Lemma 6. 1 we have

$$E_k^{**} \leqq \frac{1}{(n+1)^3} \cdot \frac{\sin^4 \theta_k |\sin^3 (n+1)\theta|}{|\cos \theta - \cos \theta_k|} \cdot \left[\frac{3}{\sin^3 \theta_k} + \frac{2}{\sin^2 \theta_k} \left\{ \frac{1}{|\sin \frac{1}{2}(\theta - \theta_k)|} + \frac{1}{\sin \frac{1}{2}(\theta + \theta_k)} \right\} \right]. \tag{6.3}$$

From (6. 3) using the inequalities (6. 1) we have

$$E_k^{**} \leqq \frac{3}{(n+1)^3} \cdot \frac{\sin \theta_k}{|\cos \theta - \cos \theta_k|} + \frac{2}{(n+1)^3} \frac{\sin \theta_k}{|\cos \theta - \cos \theta_k|} \cdot \frac{1}{|\sin \frac{1}{2}(\theta - \theta_k)|} + + \frac{2}{(n+1)^3} \cdot \frac{\sin \theta_k}{|\cos \theta - \cos \theta_k|} \cdot \frac{\sin \theta_k}{\sin \frac{1}{2}(\theta + \theta_k)}.$$

On making use of the set of the inequalities (4. 3) we get

$$E_k^{**} \leqq \frac{3}{(n+1)^3} \cdot \frac{1}{|\sin \frac{1}{2}(\theta - \theta_k)|} + \frac{2}{(n+1)^3} \cdot \frac{1}{\sin^2 \frac{1}{2}(\theta - \theta_k)} + + \frac{4}{(n+1)^3 |\sin \frac{1}{2}(\theta - \theta_k)|} \leqq \frac{9}{(n+1)^3} \cdot \frac{1}{\sin^2 \frac{1}{2}(\theta - \theta_k)}. \tag{6.4}$$

Following the same reasonings as in Lemma 4. 1 and Lemma 4. 2 we have the following two lemmas.

LEMMA 6. 2. *For* $1 \leqq k < i \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$E_k^{**} \leqq \frac{9}{n+1} \cdot \frac{1}{(i-k)^2}.$$

LEMMA 6. 3. *For* $0 \leqq i \leqq n-2$, $i+2 \leqq k \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$E_k^{**} \leqq \frac{9}{n+1} \cdot \frac{1}{(k-i-1)^2}.$$

We shall now prove

LEMMA 6. 4. *For* $1 \leqq i \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$E_i^{**} < \frac{15}{n+1}.$$

PROOF. From (6.3) on using the inequalities (6.1) we have

$$E_i^{**} \leqq \frac{3}{(n+1)^3}\frac{\sin\theta_i|\sin(n+1)\theta|}{|\cos\theta-\cos\theta_i|}+\frac{2}{(n+1)^3}\frac{\sin\theta_i\sin^2(n+1)\theta}{|\cos\theta-\cos\theta_i|\,|\sin\frac{1}{2}(\theta-\theta_i)|}+ \tag{6.5}$$
$$+\frac{2}{(n+1)^3}\frac{\sin^2\theta_i\sin(n+1)\theta}{|\cos\theta-\cos\theta_i|\sin\frac{1}{2}(\theta+\theta_i)}.$$

From (4. 7) and (4. 3) we have

$$\frac{\sin\theta_i}{\sin\frac{1}{2}(\theta+\theta_i)} \leqq 2, \quad \frac{\sin\theta_i}{|\cos\theta-\cos\theta_i|} \leqq \frac{1}{|\sin\frac{1}{2}(\theta-\theta_i)|}, \tag{6.6}$$
$$\frac{\sin\theta_i|\sin(n+1)\theta|}{|\cos\theta-\cos\theta_i|} \leqq 2(n+1).$$

Thus from (6. 5)

$$E_i^{**} \leqq \frac{6}{(n+1)^2}+\frac{8}{n+1}+\frac{8}{(n+1)^2} < \frac{15}{n+1}.$$

Similarly we prove the following

LEMMA 6. 5. *For* $0 \leqq i \leqq n-1$, *we have in* $[x_{i+1}, x_i]$

$$E_{i+1}^{**} \leqq \frac{15}{n+1}$$

Using the Lemmas 6. 2, 6. 3, 6. 4 and Lemma 6. 5 we at once have the

LEMMA 6. 6. *For* $-1 \leqq x \leqq 1$, *we have*

$$\sum_{k=1}^{n}\left(\frac{1-x^2}{1-x_k^2}\right)^2|(1-x_k^2)l_k^3(x)\psi_n(x_k,x)|(x-x_k)^2 \leqq \frac{60}{n+1}.$$

7. In this article we estimate the quantity

$$\sum_{k=1}^{n}|\lambda_k(x)|\,|x-x_k|.$$

LEMMA 7. 1. *For* $-1 \leqq x \leqq 1$

$$\sum_{k=1}^{n}|\lambda_k(x)|\,|x-x_k| < \frac{176}{n}.$$

PROOF. From (2. 2) and (2. 4) we have

$$\sum_{k=1}^{n}|\lambda_k(x)|\,|x-x_k| \leqq \sum_{k=1}^{n}\left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x)\,|x-x_k|+3\sum_{k=1}^{n}\left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x)\frac{(x-x_k)^2}{1-x_k^2}+ \tag{7.1}$$
$$+2\sum_{k=1}^{n}\left(\frac{1-x^2}{1-x_k^2}\right)^2|(1-x_k^2)l_k^3(x)\psi_n(x_k,x)|(x-x_k)^2.$$

Owing to the Lemma 4. 5, Lemma 5. 1, Lemma 6. 6 and (7. 1) we have

$$\sum_{k=1}^{n} |\lambda_k(x)|\,|x-x_k| \leqq \frac{20}{n+1}+\frac{36}{n+1}+\frac{120}{n+1} = \frac{176}{n+1} < \frac{176}{n}.$$

8. In this article we shall estimate the sum

$$\sum_{k=1}^{n}\left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x).$$

If we denote

$$S_k = \left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x)$$

then using (2. 1) we have

$$S_k = \frac{\sin^4(n+1)\theta \sin^4\theta_k}{(n+1)^4(\cos\theta-\cos\theta_k)^4}. \tag{8.1}$$

LEMMA 8. 1. *For* $1 \leqq k < i \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$S_k \leqq \frac{1}{(i-k)^4}.$$

PROOF. From the inequalities (4. 3) and the expression for S_k in (8. 1) we get as in Lemma 4. 1

$$S_k \leqq \frac{1}{(n+1)^4 \sin^4\dfrac{\theta-\theta_k}{2}} \leqq \frac{1}{(i-k)^4}.$$

Similarly we have

LEMMA 8. 2. *For* $0 \leqq i \leqq n-2$, $i+2 \leqq k \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$S_k \leqq \frac{1}{(k-i-1)^4}.$$

We further prove

LEMMA 8. 3. *For* $1 \leqq i \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$S_i \leqq 16.$$

PROOF. From (8. 1)

$$S_i = \frac{\sin^4(n+1)\theta \sin^4\theta_i}{(n+1)^4(\cos\theta-\cos\theta_i)^4},$$

and on using (4. 6) we have

$$S_i \leqq 16.$$

Similarly we have

LEMMA 8. 4. *For* $0 \leqq i \leqq n-1$, *we have in* $[x_{i+1}, x_i]$

$$S_{i+1} \leqq 16.$$

Making use of these lemmas we finally have the

LEMMA 8. 5. *For* $-1 \leqq x \leqq 1$, *we have*

$$\sum_{k=1}^{n}{}' \left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x) \leqq 40.$$

9. In this article we shall estimate the sum

$$\sum_{k=1}^{n}{}' \left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x) \frac{|x-x_k|}{(1-x_k^2)}.$$

Denoting

$$S_k^* = \left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x) \frac{|x-x_k|}{(1-x_k^2)},$$

we have owing to (2. 1)

$$S_k^* = \frac{\sin^4(n+1)\theta \sin^2\theta_k}{(n+1)^4 |\cos\theta - \cos\theta_k|^3}. \tag{9. 1}$$

We have

LEMMA 9. 1. *For* $1 \leqq k < i \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$S_k^* \leqq \frac{1}{(i-k)^3}.$$

PROOF. From (9. 1) on using (4. 3) we have

$$S_k^* \leqq \frac{|\sin(n+1)\theta|}{|\cos\theta - \cos\theta_k|} \cdot \frac{1}{(n+1)^4 \sin^2 \frac{\theta-\theta_k}{2}}. \tag{9. 2}$$

Now

$$|\sin(n+1)\theta| = |\sin(n+1)\theta + \sin(n+1)\theta_k| = 2\left|\sin(n+1)\frac{\theta+\theta_k}{2}\cos(n+1)\frac{\theta-\theta_k}{2}\right| \leqq$$

$$\leqq 2(n+1)\sin\frac{\theta+\theta_k}{2}$$

and

$$|\cos\theta - \cos\theta_k| = 2\sin\frac{\theta+\theta_k}{2}\left|\sin\frac{\theta-\theta_k}{2}\right|$$

therefore

$$\frac{|\sin(n+1)\theta|}{|\cos\theta - \cos\theta_k|} \leqq \frac{n+1}{\left|\sin\frac{\theta-\theta_k}{2}\right|}. \tag{9. 3}$$

From (9. 2) and (9. 3) we have

$$S_k^* \leqq \frac{1}{(n+1)^3} \cdot \frac{1}{\left|\sin\frac{\theta-\theta_k}{2}\right|^3}. \tag{9. 4}$$

Owing to the conditions of the lemma we have, similar to Lemma 4. 1, the inequality

$$S_k^* \leqq \frac{1}{(i-k)^3}$$

which proves Lemma 9. 1. Similarly we have

LEMMA 9. 2. *For* $0 \leqq i \leqq n-2, \quad i+2 \leqq k \leqq n$ *we have in* $[x_{i+1}, x_i]$

$$S_k^* \leqq \frac{1}{(k-i-1)^3}.$$

We now prove

LEMMA 9. 3. *For* $1 \leqq i \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$S_i^* \leqq 16$$

PROOF. From (9. 1)

$$S_i^* = \frac{\sin^4(n+1)\theta \sin^2\theta_i}{(n+1)^4 |\cos\theta - \cos\theta_i|^3}.$$

Using (4. 7) we have

$$S_i^* \leqq \frac{\sin^2(n+1)\theta}{|\cos\theta - \cos\theta_i|} \cdot \frac{4}{(n+1)^2}. \tag{9. 5}$$

Since

$$|\sin(n+1)\theta| = |\sin(n+1)\theta - \sin(n+1)\theta_i| =$$

$$= 2\left|\sin(n+1)\frac{\theta-\theta_i}{2}\cos(n+1)\frac{\theta+\theta_i}{2}\right| \leqq 2(n+1)\left|\sin\frac{\theta-\theta_i}{2}\right|,$$

$$|\sin(n+1)\theta| = |\sin(n+1)\theta + \sin(n+1)\theta_i| =$$

$$= 2\left|\sin(n+1)\frac{\theta+\theta_i}{2}\cos(n+1)\frac{\theta-\theta_i}{2}\right| \leqq 2(n+1)\sin\frac{\theta+\theta_i}{2},$$

$$|\cos\theta - \cos\theta_i| = 2\sin\frac{\theta+\theta_i}{2}\left|\sin\frac{\theta-\theta_i}{2}\right|,$$

therefore

$$\frac{\sin^2(n+1)\theta}{|\cos\theta - \cos\theta_i|} \leqq 4(n+1)^2. \tag{9. 6}$$

From (9. 5) and (9. 6) we have

$$S_i^* \leqq 16.$$

Similarly we have

LEMMA 9. 4. *For* $0 \leqq i \leqq n-1$, *we have in* $[x_{i+1}, x_i]$

$$S_{i+1}^* \leqq 16.$$

Making use of the above lemmas we have

LEMMA 9. 5. *For* $-1 \leqq x \leqq 1$, *we have*

$$\sum_{k=1}^{n} \left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x) \frac{|x-x_k|}{(1-x_k^2)} \leqq 36.$$

10. In this article we shall estimate the quantity

$$\sum_{k=1}^{n} \left(\frac{1-x^2}{1-x_k^2}\right)^2 |(1-x_k^2) l_k^3(x) \psi_n(x_k, x)(x-x_k)|.$$

If we denote

$$S_k^{**} = \left(\frac{1-x^2}{1-x_k^2}\right)^2 |l_k^3(x)(1-x_k^2)\psi_n(x_k, x)(x-x_k)|$$

then owing to (2. 1) we have

$$S_k^{**} = \left| \frac{(-1)^{k+1}}{(n+1)^3} \frac{\sin^4\theta_k \sin^3(n+1)\theta}{(\cos-\cos\theta_k)^2} \sqrt{1-x^2}\, \psi_n(x_k, x) \right|.$$

Making use of Lemma 6. 1 we have

$$S_k^{**} \leqq \frac{1}{(n+1)^3} \frac{\sin^4\theta_k |\sin^3(n+1)\theta|}{(\cos\theta-\cos\theta_k)^2} \left[\frac{3}{\sin^3\theta_k} + \right.$$

$$\left. + \frac{2}{\sin^2\theta_k} \left\{ \frac{1}{|\sin\frac{1}{2}(\theta-\theta_k)|} + \frac{1}{\sin\frac{1}{2}(\theta+\theta_k)} \right\} \right]. \tag{10. 1}$$

From this on using the inequalities (6. 1) we have

$$S_k^{**} \leqq \frac{3}{(n+1)^3} \cdot \frac{\sin\theta_k}{|\cos\theta-\cos\theta_k|} \cdot \frac{|\sin(n+1)\theta|}{|\cos\theta-\cos\theta_k|} +$$

$$+ \frac{2}{(n+1)^3} \frac{\sin^2\theta_k}{(\cos\theta-\cos\theta_k)^2} \cdot \frac{1}{|\sin\frac{1}{2}(\theta-\theta_k)|} + \frac{2}{(n+1)^3} \frac{\sin^2\theta_k}{(\cos\theta-\cos\theta_k)^2} \cdot$$

$$\cdot \frac{|\sin(n+1)\theta|}{|\sin\frac{1}{2}(\theta+\theta_k)|}$$

which owing to (4. 3), (9. 3)

$$S_k^{**} \leqq \frac{3}{(n+1)^2} \cdot \frac{1}{\sin^2\frac{\theta-\theta_k}{2}} + \frac{4}{(n+1)^2} \cdot \frac{1}{\sin^2\frac{\theta-\theta_k}{2}} +$$

$$+ \frac{4}{(n+1)^2 \sin^2\frac{\theta-\theta_k}{2}} = \frac{11}{(n+1)^2} \frac{1}{\sin^2\frac{\theta-\theta_k}{2}}. \tag{10. 2}$$

Following the same reasoning as in Lemmas 4. 1 and 4. 2 we have the following two lemmas.

LEMMA 10. 1. *For* $1 \leqq k < i \leqq n$ *we have* $[x_{i+1}, x_i]$

$$S_k^{**} \leqq \frac{11}{(i-k)^2}.$$

LEMMA 10. 2. *For* $0 \leqq i \leqq n-2, \quad i+2 \leqq k \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$S_k^{**} \leqq \frac{11}{(k-i-1)^2}.$$

We shall now prove

LEMMA 10. 3. *For* $1 \leqq i \leqq n$, *we have in* $[x_{i+1}, x_i]$

$$S_i^{**} < 56.$$

PROOF. From (10. 1) on using the inequalities (6. 1), (6. 6), (9. 6) we have

$$S_i^{**} \leqq \frac{3}{(n+1)^3} \cdot \frac{\sin\theta_i |\sin(n+1)\theta|}{|\cos\theta - \cos\theta_i|} \cdot \frac{\sin^2(n+1)\theta}{|\cos\theta - \cos\theta_i|} +$$

$$+ \frac{2}{(n+1)^3} \frac{\sin^2\theta_i \sin^2(n+1)\theta}{(\cos\theta - \cos\theta_i)^2} \cdot \frac{|\sin(n+1)\theta|}{|\sin\frac{1}{2}(\theta-\theta_i)|} + \frac{2}{(n+1)^3} \frac{\sin^2\theta_i \sin^2(n+1)\theta}{(\cos\theta - \cos\theta_i)^2} \cdot$$

$$\cdot \frac{|\sin(n+1)\theta|}{\sin\frac{1}{2}(\theta+\theta_i)} < 24 + 16 + 16 = 56,$$

which proves the Lemma 10. 3. Similarly we have

LEMMA 10. 4. *For* $0 \leqq i \leqq n-1$, *we have in* $[x_{i+1}, x_i]$

$$S_{i+1}^{**} < 56.$$

Using the Lemmas 10. 1, 10. 2, 10. 3 and Lemma 10. 4 we at once have

LEMMA 10. 5. *For* $-1 \leqq x \leqq 1$ *we have*

$$\sum_{k=1}^{n} \left(\frac{1-x^2}{1-x_k^2}\right)^2 |(1-x_k^2) l_k^3(x) \psi_n(x_k, x)(x-x_k)| < 150.$$

11. In this article we shall estimate the quantity

$$\sum_{k=1}^{n} |\lambda_k(x)|.$$

LEMMA 11. 1. *For* $-1 \leqq x \leqq 1$

$$\sum_{k=1}^{n} |\lambda_k(x)| < 226.$$

PROOF. From (2. 2) and (2. 4) we have

$$\sum_{k=1}^{n}|\lambda_k(x)| \leqq \sum_{k=1}^{n}\left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x)+3\sum_{k=1}^{n}\left(\frac{1-x^2}{1-x_k^2}\right)^2 l_k^4(x)\frac{|x-x_k|}{1-x_k^2}+$$

$$+2\sum_{k=1}^{n}\left(\frac{1-x^2}{1-x_k^2}\right)^2|(1-x_k^2)l_k^3(x)\psi_n(x_k,x)(x-x_k)|.$$

Now making use of Lemma 8. 5, Lemma 9. 5 and Lemma 10. 5 we have

$$\sum_{k=1}^{n}|\lambda_k(x)| < 40+36+150 = 226.$$

12. PROOF of the theorem.

Let $\omega(f, \delta)$ denote the modulus of continuity of the function $f(x)$ then

$$\omega(f, \mu\delta) \leqq (\mu+1)\omega(f, \delta), \quad \mu > 0$$

and

$$\omega(f, |x-x_k|) \leqq \omega\left(f, \frac{1}{n}\right)+n|x-x_k|\,\omega\left(f, \frac{1}{n}\right).$$

Now

$$f(x)-\Lambda_n(f, x) = f(x)-\sum_{k=1}^{n}f(x)\lambda_k(x)+\sum_{k=1}^{n}f(x)\lambda_k(x)-$$

$$-\left\{\frac{1+x}{2}f(1)+\frac{1-x}{2}f(-1)\right\}-\sum_{k=1}^{n}\left[f(x_k)-\left\{\frac{1+x}{2}f(1)+\frac{1-x}{2}f(-1)\right\}\right]\lambda_k(x)$$

$$= f(x)\left[1-\sum_{k=1}^{n}\lambda_k(x)\right]+\sum_{k=1}^{n}[f(x)-f(x_k)]\lambda_k(x)-\left\{\frac{1+x}{2}f(1)+\frac{1-x}{2}f(-1)\right\}\cdot$$

$$\cdot\left[1-\sum_{k=1}^{n}\lambda_k(x)\right] = \frac{1}{2}[(1+x)\{f(x)-f(1)\}+(1-x)\cdot$$

$$\cdot\{f(x)-f(-1)\}\left[1-\sum_{k=1}^{n}\lambda_k(x)\right]+\sum_{k=1}^{n}[f(x)-f(x_k)]\lambda_k(x)$$

$$= \frac{1+x}{2}[f(x)-f(1)]\left[1-\sum_{k=1}^{n}\lambda_k(x)\right]+\frac{1-x}{2}[f(x)-f(-1)]\left[1-\sum_{k=1}^{n}\lambda_k(x)\right]+$$

$$+\sum_{k=1}^{n}[f(x)-f(x_k)]\lambda_k(x).$$

Now from Lemma 3.2

$$\left|\frac{1+x}{2}[f(x)-f(1)]\left[1-\sum_{k=1}^{n}\lambda_k(x)\right]\right| \leqq \frac{|1+x|}{2}\omega(f,|x-1|)\left|1-\sum_{k=1}^{n}\lambda_k(x)\right| \leqq$$

$$\leqq \frac{|1+x|}{2}[1+n|x-1|]\left|1-\sum_{k=1}^{n}\lambda_k(x)\right|\omega\left(f,\frac{1}{n}\right) < \left|1-\sum_{k=1}^{n}\lambda_k(x)\right|\omega\left(f,\frac{1}{n}\right)+$$

$$+n\frac{|1+x|\cdot|1-x|}{2}\left|1-\sum_{k=1}^{n}\lambda_k(x)\right|\omega\left(f,\frac{1}{n}\right) < \left|1-\sum_{k=1}^{n}\lambda_k(x)\right|\omega\left(f,\frac{1}{n}\right)+$$

$$+n\sqrt{1-x^2}\left|1-\sum_{k=1}^{n}\lambda_k(x)\right|\omega\left(f,\frac{1}{n}\right) < (3+3)\,\omega\left(f,\frac{1}{n}\right) = 6\omega\left(f,\frac{1}{n}\right), -1 \leqq x \leqq 1.$$

Similarly we prove that

$$\left|\frac{1-x}{2}[f(x)-f(-1)]\left[1-\sum_{k=1}^{n}\lambda_k(x)\right]\right| < 6\omega\left(f,\frac{1}{n}\right).$$

Further on account of the Lemma 7. 1 and Lemma 11. 1 we have for $-1 \leqq x \leqq 1$

$$\sum_{k=1}^{n}|f(x)-f(x_k)|\,|\lambda_k(x)| \leqq \omega\left(f,\frac{1}{n}\right)\sum_{k=1}^{n}(1+n|x-x_k|)\,|\lambda_k(x)| \leqq$$

$$\leqq \omega\left(f,\frac{1}{n}\right)\left[\sum_{k=1}^{n}|\lambda_k(x)|+n\sum_{k=1}^{n}|\lambda_k(x)|\,|x-x_k|\right] <$$

$$< (226+176)\omega\left(f,\frac{1}{n}\right) = 402\omega\left(f,\frac{1}{n}\right).$$

Thus

$$|f(x)-\Lambda_n(f,x)| < 414\omega\left(f,\frac{1}{n}\right), \qquad -1 \leqq x \leqq 1.$$

and the proof of our theorem is completed.

REFERENCES

[1] FREUD, G.: Egy Jackson-féle interpolációs eljárásról, *Mat. Lapok* **4** (1964) 330—336.

[2] SALLAY, M.: Über ein Interpolationsverfahren, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **9** (1964) 607—615.

[3] FANTA, K., KISS, O.: О сходимости интерполяционных методов решения граничных задач для обыкновенных дифференциальных уравнений, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **9** (1964) 89—112.

ALBERTA UNIVERSITY, EDMONTON, CANADA
and
LUCKNOW UNIVERSITY, LUCKNOW, INDIA

(Received June 28, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 185—191.

ОБ ОЦЕНКЕ ПОГРЕШНОСТИ ПРИ НАХОЖДЕНИИ СОБСТВЕННЫХ ФУНКЦИЙ МЕТОДОМ КОНЕЧНЫХ РАЗНОСТЕЙ

L. VEIDINGER

1. Рассмотрим в открытой двумерной области R с границей C задачу на собственные значения:

$$Lu + \lambda u = 0 \text{ в } R, \tag{1}$$

$$u = 0 \text{ на } C,$$

где

$$Lu = \frac{\partial}{\partial x}\left[a(x,y)\frac{\partial u}{\partial x}\right] + \frac{\partial}{\partial y}\left[b(x,y)\frac{\partial u}{\partial y}\right] - f(x,y)u,$$

$$a(x,y) \geqq \alpha > 0, \quad b(x,y) \geqq \alpha > 0, \quad f(x,y) \geqq 0, \quad \alpha = \text{const}$$

Пусть $\lambda^1 \leqq \lambda^2 \leqq \ldots$ — собственные значения задачи (1). Собственные функции $u^1, u^2, \ldots$, соответствующие собственным значениям $\lambda^1, \lambda^2, \ldots$, можно выбрать ортонормированными в том смысле, что

$$\iint_R u^k u^l \, dx\, dy = \delta_{kl} = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases}$$

В работе [1] была получена равномерная оценка погрешности собственных функций задачи (1), вычисляемых методом конечных разностей. Однако, при выводе оценки в работе [1] предполагается, что область R является объединением ячеек сетки и

$$u^k(x,y) \in C^{(8)}(\bar{R})^*, \quad \text{где} \quad \bar{R} = R \cup C.$$

Эти два условия обычно одновременно не выполняются (см. [2] или [3]).

В работе [4] дается оценка погрешности собственных функций в норме L_2. При выводе оценки предполагается, что $u^k(x,y) \in C^{(4)}(\bar{R})$.

В настоящей работе получается равномерная оценка погрешности собственных функций при предположении, что $a(x,y) \in C^{(3)}(\bar{R})$, $b(x,y) \in C^{(3)}(\bar{R})$ и $u^k(x,y) \in C^{(4)}(\bar{R})$. При выводе оценки погрешности мы используем некоторые идеи из работы [5].

2. Пусть бесконечная плоскость, в которой находится область R, разделена двумя семействами параллельных прямых, образующими квадратную сетку. Пусть этими прямыми будут $x = mh$ и $y = nh$ ($m, n = 0, \pm 1, \ldots$). Точки

* Функция $\varphi(x,y)$, определенная в ограниченной замкнутой области T, принадлежит классу $C^{(m)}(T)$, если она имеет в области T непрерывные частные производные m-го порядка.

(mh, nh) называются узлами сетки. Через R_h обозначим множество узлов сетки, которые вместе с четырьмя отрезками прямых, соединяющими их с соседними узлами, принадлежат R, а через C_h^* — множество остальных узлов, принадлежащих R. Пусть $S_h = R_h \cup C_h^*$. Обозначим через C_h множество точек пересечения C с прямыми сетки.

Если $P \in S_h$, то ближайшие к P точки множества $S_h \cup C_h$, лежащие на прямых сетки, проходящих через P, называются соседними точками узла P.

Введем следующие обозначения

$$V_{\bar{x}}(P) = \frac{V(P)-V(W)}{h_W}, \quad V_{\hat{x}}(P) = \frac{V(E)-V(P)}{0{,}5(h_E+h_W)},$$

$$V_{\bar{y}}(P) = \frac{V(P)-V(S)}{h_S}, \quad V_{\hat{y}}(P) = \frac{V(N)-V(P)}{0{,}5(h_N+h_S)},$$

где $V = V(P)$ — любая функция, определенная в точках $S_h \cup C_h$; $E = (x_P + h_E, y_P)$, $S = (x_P, y_P - h_S)$, $W = (x_P - h_W, y_P)$, $N = (x_P, y_P + h_N)$ — соседние точки узла $P = (x_P, y_P)$. (Если $P \in R_h$, то $h_E = h_S = h_W = h_N = h$.)

Задача (1) аппроксимируется разностной задачей

$$\begin{aligned} L_h U + \lambda_h U &= 0 \quad \text{в} \quad S_h, \\ U &= 0 \quad \text{на} \quad C_h, \end{aligned} \tag{2}$$

где L_h — пятиточечный разностный оператор Микеладзе, определенный формулой

$$L_h U = (a_1 U_{\bar{x}})_{\hat{x}} + (b_1 U_{\bar{y}})_{\hat{y}} - fU.$$

Здесь $a_1(P) = a(W')$, $W' = (x_P - 0{,}5h_W, y_P)$, $b_1(P) = b(S')$, $S' = (x_P, y_P - 0{,}5h_S)$.

Следуя [6], введем скалярное произведение сеточных функций. Пусть V и W — произвольные функции, заданные на S_h. Обозначим

$$(V, W) = \sum_{P \in S_h} V(P)\, W(P) H_P,$$

где

$$H_P = \frac{(h_E+h_W)(h_N+h_S)}{4}.$$

Пусть $\lambda_h^1 \leqq \lambda_h^2 \leqq \ldots$ — собственные значения задачи (2). Собственные функции $U^1, U^2, \ldots$, соответствующие собственным значениям $\lambda_h^1, \lambda_h^2, \ldots$, можно выбрать ортонормированными в том смысле, что

$$(U^k, U^l) = \delta_{kl}.$$

Лемма 1. *При $k = 1, 2, \ldots$ имеет место неравенство*

$$c_1 k < \lambda_h^k < c_2 k, \tag{3}$$

где c_1 и c_2 — положительные постоянные, зависящие только от области R и от коэффициентов оператора L.

Лемма 2. *При* $k=1, 2, \ldots$ *имеет место неравенство*

$$\max_{P\in S_h} |U^k(P)| < c_3(\lambda_h^k)^{1/2} < c_4 k^{1/2}, \tag{4}$$

где c_3 *и* c_4 *— положительные постоянные, зависящие только от области* R *и от коэффициентов оператора* L.

Мы не останавливаемся на доказательстве этих лемм, так как оно проводится подобно [1], с некоторым изменением для нашего случая.

Пусть p — число узлов множества S_h. Из (3) и (4) следует, что

$$\sum_{j=1}^{p} \frac{[U^j(P)]^2}{(\lambda_h^j)^2} < c_5 \log h^{-1}, \quad P\in S_h, \tag{5}$$

где c_5 — положительная постоянная, зависящая только от области R и от коэффициентов оператора L.

Введем разностную функцию Грина $G_h(P, Q)$, полагая

$$L_{h,P}G_h(P, Q) = -\frac{\delta(P, Q)}{H_P}, \quad P\in S_h, Q\in S_h\cup C_h,$$

$$G_h(P, Q) = 0, \ P\in C_h, Q\in S_h\cup C_h,$$

где $\delta(P, Q)=0$ при $P\neq Q$, $\delta(P, P)=1$. Индекс P в символе $L_{h,P}$ означает, что оператор должен быть применен относительно переменной P.

Легко показать, что для функции $G_h(P, Q)$ справедлива формула

$$G_h(P, Q) = \sum_{j=1}^{p} \frac{U^j(P)U^j(Q)}{\lambda_h^j}.$$

3. Теорема. *Предположим, что* $a(x, y)\in C^{(3)}(\bar{R})$, $b(x, y)\in C^{(3)}(\bar{R})$ и $u^k(x, y)\in C^{(4)}(\bar{R})$. *Тогда, если* λ^k *— простое собственное значение задачи* (1), *то имеет место неравенство*

$$|u^k(P)-U^k(P)| < c_6 h^2 \log h^{-1}, \quad P\in S_h,$$

где c_6 *— положительная постоянная, зависящая только от* k, *от области* R *и от коэффициентов оператора* L. *Если* $\lambda^k=\lambda^{k+1}=\ldots=\lambda^{k+n-1}$ *— n-кратное собственное значение задачи* (1), *то имеет место оценка*

$$\left| u^k(P)-\sum_{i=0}^{n-1} \beta_h^{k+i} U^{k+i}(P) \right| < c_7 h^2 \log h^{-1}, \quad P\in S_h,$$

где $\beta_h^k, \ldots, \beta_h^{k+n-1}$ *— действительные коэффициенты,* c_7 *— положительная постоянная, зависящая только от* k, *от области* R *и от коэффициентов оператора* L.

Доказательство. Рассмотрим сначала случай, когда λ^k — простое собственное значение.

Пусть

$$\Phi_h^k = L_h u^k - Lu^k + (\lambda_h^k-\lambda^k)u^k. \tag{6}$$

Из предположений теоремы следует, что (см. [7])

$$L_h u^k(P) - L u^k(P) = \begin{cases} O(h^2), & P \in R_h, \\ O(h), & P \in C_h^*. \end{cases} \tag{7}$$

В работе [4] доказана оценка

$$\lambda_h^k - \lambda^k = O(h^2). \tag{8}$$

Подставляя (7) и (8) в (6), получим

$$\Phi_h^k(P) = \begin{cases} O(h^2), & P \in R_h, \\ O(h), & P \in C_h^*, \end{cases} \tag{9}$$

Пусть $c_k = (U^k, u^k)$ и $W^k = u^k - c_k U^k$. Функция W^k удовлетворяет условию

$$(U^k, W^k) = 0. \tag{10}$$

Далее, W^k является решением задачи

$$\begin{aligned} L_h W^k + \lambda_h^k W^k &= \Phi_h^k \quad \text{в} \quad S_h, \\ W^k &= 0 \quad \text{на} \quad C_h. \end{aligned} \tag{11}$$

Из (10) и (11) следует, что (см. например, [8], стр. 43)

$$W^k = \sum_{\substack{j=1 \\ (j \neq k)}}^{p} \frac{(\Phi_h^k, U^j)}{\lambda_h^k - \lambda_h^j} U^j, \tag{12}$$

где p—число узлов множества S_h.

Мы можем записать (12) в виде

$$W^k(P) = \sum_{Q \in S_h} R_h^k(P, Q) \Phi_h^k(Q) H_Q, \tag{13}$$

где

$$R_h^k(P, Q) = \sum_{\substack{j=1 \\ (j \neq k)}}^{p} \frac{U^j(P) U^j(Q)}{\lambda_h^k - \lambda_h^j} =$$

$$= -\sum_{j=1}^{p} \frac{U^j(P) U^j(Q)}{\lambda_h^j} + \frac{U^k(P) U^k(Q)}{\lambda_h^k} + \sum_{\substack{j=1 \\ (j \neq k)}}^{p} \frac{U^j(P) U^j(Q)}{(\lambda_h^j)^2} \frac{\lambda_h^k}{\frac{\lambda_h^k}{\lambda_h^j} - 1} =$$

$$= -G_h(P, Q) + \frac{U^k(P) U^k(Q)}{\lambda_h^k} + \sum_{\substack{j=1 \\ (j \neq k)}}^{p} \frac{U^j(P) U^j(Q)}{(\lambda_h^j)^2} \frac{\lambda_h^k}{\frac{\lambda_h^k}{\lambda_h^j} - 1}.$$

Пусть

$$S_1(P) = -\sum_{Q \in S_h} G_h(P, Q) \Phi_h^k(Q) H_Q.$$

Аналогично [9] можно показать, что

$$S_1(P) = O(h^2). \tag{14}$$

Пусть

$$S_2(P) = \sum_{Q \in S_h} \frac{U^k(P)\,U^k(Q)}{\lambda_h^k}\,\Phi_h^k(Q)\,H_Q.$$

По лемме 2 собственные функции U^k равномерно по h ограничены в S_h. Отсюда и из (9) следует, что

$$S_2(P) = O(h^2). \tag{15}$$

Пусть

$$S_3(P) = \sum_{Q \in S_h} \sum_{\substack{j=1 \\ (j \neq k)}}^{p} \frac{U^j(P)\,U^j(Q)}{(\lambda_h^j)^2}\,\frac{\lambda_h^k}{\frac{\lambda_h^k}{\lambda_h^j} - 1}\,\Phi_h^k(Q)\,H_Q.$$

Применяя неравенство Коши-Буняковского, получим

$$\sum_{\substack{j=1 \\ (j \neq k)}}^{p} \frac{U^j(P)\,U^j(Q)}{(\lambda_h^j)^2}\,\frac{\lambda_h^k}{\frac{\lambda_h^k}{\lambda_h^j} - 1} = O\left(\sqrt{\sum_{j=1}^{p} \frac{[U^j(P)]^2}{(\lambda_h^j)^2}}\,\sqrt{\sum_{j=1}^{p} \frac{[U^j(Q)]^2}{(\lambda_h^j)^2}}\right). \tag{16}$$

Подставляя (5) в (16), находим

$$\sum_{\substack{j=1 \\ (j \neq k)}}^{p} \frac{U^j(P)\,U^j(Q)}{(\lambda_h^j)^2}\,\frac{\lambda_h^k}{\frac{\lambda_h^k}{\lambda_h^j} - 1} = O(\log h^{-1}).$$

Отсюда и из (9) следует, что

$$S_3(P) = O(h^2 \log h^{-1}). \tag{17}$$

Подставляя (14), (15) и (17) в (13), получим

$$W^k(P) = O(h^2 \log h^{-1}). \tag{18}$$

Пусть $Z^k = u^k - U^k$. Тогда

$$Z^k = \frac{W^k}{c_k} + \frac{c_k - 1}{c_k}\,u^k. \tag{19}$$

Из определения c_k и W^k следует, что

$$(W^k, u^k) = (u^k, u^k) - c_k(U^k, u^k) = (u^k, u^k) - c_k^2,$$

откуда

$$c_k^2 = (u^k, u^k) - (W^k, u^k). \tag{20}$$

Но, как легко проверить,

$$(u^k, u^k) = 1 + O(h^2). \tag{21}$$

Подставляя (18) и (21) в (20), получим

$$c_k^2 = 1 + O(h^2 \log h^{-1}). \tag{22}$$

Выберем знаки функций u^k и U^k так, чтобы $c_k > 0$. Тогда из (18), (19) и (22) следует, что

$$Z^k(P) = \frac{W^k(P)}{c_k} + \frac{c_k^2 - 1}{c_k(c_k+1)} u^k(P) = O(h^2 \log h^{-1}).$$

Пусть теперь $\lambda^k = \lambda^{k+1} = \ldots = \lambda^{k+n-1}$ — n-кратное собственное значение. Рассмотрим функцию

$$V^k = u^k - \sum_{i=0}^{n-1} \beta_h^{k+i} U^{k+i},$$

где

$$\beta_h^{k+i} = (U^{k+i}, u^k).$$

Функция V^k при $i = 0, 1, \ldots, n-1$ удовлетворяет условию

$$(V^k, U^{k+i}) = 0. \tag{23}$$

Далее, легко видеть, что V^k является решением задачи

$$\begin{aligned} L_h V^k + \lambda_h^k V^k &= \Psi_h^k \quad \text{в} \quad S_h, \\ V^k &= 0 \quad \text{на} \quad C_h, \end{aligned} \tag{24}$$

где

$$\Psi_h^k = \Phi_h^k + \sum_{i=0}^{n-1} (\lambda_h^k - \lambda_h^{k+i}) \beta_h^{k+i} U^{k+i}.$$

Из (23) и (24) следует, что (см. [8], стр. 43)

$$V^k = \sum_{\substack{j=1 \\ (j \neq k, k+1, \ldots, k+n-1)}}^{p} \frac{(\Psi_h^k, U^j)}{\lambda_h^k - \lambda_h^j}.$$

Отсюда, рассуждая, как в случае простого собственного значения, получим

$$V^k(P) = O(h^2 \log h^{-1}).$$

Этим и заканчивается доказательство теоремы.

БИБЛИОГРАФИЯ

[1] Саульев, В. К.: *Об оценке погрешности при нахождении собственных функций методом конечных разностей,* В сб. „Вычислительная математика" № 1, Изд-во АН СССР, Москва, 1957, стр. 87—115.

[2] Кондратьев, В. А.: Краевые задачи для эллиптических уравнений в конических областях, *Докл. Акад. Наук* СССР **153** (1963) 27—29.

[3] Кондратьев, В. А.: Краевые задачи для эллиптических уравнений высших порядков при наличии особенностей границы, *Материалы к Совместному советско-американскому симпозиуму по уравнениям с частными производными,* Изд-во Сибирского отделения АН СССР, Новосибирск, 1963.

[4] Приказчиков, В. Г.: Разностная задача на собственные значения для эллиптического оператора, *Ж. Вычисл. Мат. и Мат. Физ.* **5** (1965) 648—657.

[5] Тихонов, А. Н. и Самарский, А. А.: Разностная задача Штурма-Лиувилля, *Ж. Вычисл. Мат. и Мат. Физ.* **1** (1961) 784—805.

[6] Самарский, А. А.: Локально-одномерные схемы на неравномерных сетках, *Ж. Вычисл. Мат. и Мат. Физ.* **3** (1963) 431—466.

[7] Тихонов, А. Н. и Самарский, А. А.: Однородные разностные схемы на неравномерных сетках, *Ж. Вычисл. Мат. и Мат. Физ.* **2** (1962) 812—832.

[8] Gould, S. H.: *Variational methods for eigenvalue problems,* Univ. Toronto Press, Toronto, 1957.

[9] Самарский, А. А.: О точности метода сеток для задачи Дирихле в произвольной области, *Apl. Mat.* **10** (1965) 293 – 296

МАТЕМАТИЧЕСКИЙ ИНСТИТУТ ВЕНГЕРСКОЙ АКАДЕМИИ НАУК, БУДАПЕШТ

(Поступило 12-ого августа 1966 г.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 193—199.

О РАЗНОСТНОМ МЕТОДЕ PÓLYA

L. VEIDINGER

В работе [1] PÓLYA предложил разностный метод получения верхних границ для собственных значений оператора Лапласа. В настоящей работе метод PÓLYA обобщается на более общие эллиптические операторы и устанавливается порядок погрешности собственных значений, вычисляемых разностным методом PÓLYA.

1. Рассмотрим сначала задачу

$$\begin{aligned} \Delta u + \lambda u &= 0 \quad \text{в} \quad R, \\ u &= 0 \quad \text{на} \quad C, \end{aligned} \tag{1}$$

где R — открытая двумерная область, граница которой C состоит из конечного числа кусочно-аналитических простых замкнутых кривых. Обозначим через $\lambda^1 \leqq \lambda^2 \leqq \ldots$ собственные значения задачи (1).

Пусть бесконечная плоскость, в которой находится область R, разделена двумя семействами параллельных прямых, образующими квадратную сетку. Пусть этими прямыми будут $x = mh$ и $y = nh$ ($m, n = 0, \pm 1, \pm 2, \ldots$). Точки (mh, nh) называются узлами сетки. Наименьший квадрат, ограниченный четырьмя прямыми сетки, мы называем ячейкой сетки. Пусть R^* — объединение всех ячеек квадратной сетки, лежащих внутри области R, а C^* — граница области R^*. Пусть R_h состоит из всех внутренних узлов R^* и пусть C_h состоит из всех узлов на C^*.

В работе [1] задача (1) аппроксимируется разностной задачей

$$\begin{aligned} \Delta_h U(P) + \mu_h \frac{6U(P) + U(E) + U(SE) + U(S) + U(W) + U(NW) + U(N)}{12} &= 0, \\ U(P) = 0, \quad P \in C_h, & \qquad P \in R_h, \end{aligned} \tag{2}$$

где $E = (x_P + h, y_P)$, $SE = (x_P + h, y_P - h)$, $S = (x_P, y_P - h)$, $W = (x_P - h, y_P)$, $NW = (x_P - h, y_P + h)$, $N = (x_P, y_P + h)$ — соседние точки узла $P = (x_P, y_P)$, а Δ_h — пятиточечный разностный оператора Лапласа, определенный формулой

$$\Delta_h U(P) = h^{-2}[U(E) + U(S) + U(W) + U(N) - 4U(P)].$$

Обозначим через $\mu_h^1 \leqq \mu_h^2 \leqq \ldots$ собственные значения задачи (2), а через $\lambda_h^1 \leqq \lambda_h^2 \leqq \ldots$ собственные значения задачи

$$\Delta_h U(P) + \lambda_h U(P) = 0, \quad P \in R_h, \qquad U(P) = 0, \quad P \in C_h.$$

В работе [2] доказано, что если $h^2\lambda_h^k<4$, то имеет место неравенство

$$\lambda^k \leqq \mu_h^k \leqq \frac{\lambda_h^k}{1-\frac{h^2}{4}\lambda_h^k}. \tag{3}$$

Теорема 1. *Пусть R состоит из конечного числа конгруэнтных квадратов со стороной h_0, ограниченных прямыми сетки. Если $h=\frac{h_0}{p}$, где p — положительное целое число, то при $k=1, 2, \ldots$ имеем*

$$-c_1 h^{4/3} < \lambda^k - \mu_h^k \leqq 0, \tag{4}$$

где c_1 — положительная постоянная, зависящая только от k и от области R.

Доказательство. При предположениях Теоремы 1 имеет место оценка (см. [3])

$$\lambda_h^k \leqq \lambda^k + c_2 h^{4/3}, \tag{5}$$

где c_2 — положительная постоянная, зависящая только от k и от области R. Сопоставляя неравенства (3) и (5) получаем желаемую оценку (4).

Теорема 2. *Пусть R—произвольная открытая двумерная область, граница которой C состоит из конечного числа кусочно-аналитических простых замкнутых кривых. Тогда при $k=1, 2, \ldots$ имеем*

$$-c_3 h < \lambda^k - \mu_h^k \leqq 0, \tag{6}$$

где c_3 — положительная постоянная, зависящая только от k и от области R.

Доказательство. При предположениях Теоремы 2 имеет место оценка (см. [3])

$$\lambda_h^k \leqq \lambda^k + c_4 h, \tag{7}$$

где c_4 — положительная постоянная, зависящая только от k и от области R. Сопоставляя неравенства (3) и (7), получаем желаемую оценку (6).

2. Рассмотрим теперь более общую задачу

$$\begin{aligned} Lu+\lambda u &= 0 \quad \text{в} \quad R, \\ u &= 0 \quad \text{на} \quad C, \end{aligned} \tag{8}$$

где L — линейный самосопряженный дифференциальный оператор

$$Lu = \frac{\partial}{\partial x}\left[a(x,y)\frac{\partial u}{\partial x}\right]+\frac{\partial}{\partial x}\left[b(x,y)\frac{\partial u}{\partial y}\right]+$$

$$+\frac{\partial}{\partial y}\left[b(x,y)\frac{\partial u}{\partial x}\right]+\frac{\partial}{\partial y}\left[c(x,y)\frac{\partial u}{\partial y}\right]-f(x,y)u$$

эллиптического типа, т. е. такой, что для всех $(x, y)\in R$

$$a\xi^2+2b\xi\eta+c\eta^2 \geqq \alpha(\xi^2+\eta^2) \qquad (\alpha = \text{const} > 0)$$

при любых действительных ξ, η. Предположим, что $f(x, y)\geqq 0$.

Введем обозначения

$$V_x(P) = h^{-1}[V(E)-V(P)], \quad V_{\bar{x}}(P) = h^{-1}[V(P)-V(W)],$$
$$V_y(P) = h^{-1}[V(N)-V(P)], \quad V_{\bar{y}}(P) = h^{-1}[V(P)-V(S)],$$

где $V = V(P)$ — любая функция, определенная в узлах сетки.

Задача (8) аппроксимируется разностной задачей

$$L'_h U(P) + \mu_h \frac{6U(P)+U(E)+U(SE)+U(S)+U(W)+U(NW)+U(N)}{12} = 0, \quad P \in R_h,$$

$$U(P) = 0, \quad P \in C_h, \tag{9}$$

где

$$\begin{aligned} L'_h U(P) = {} & \{[A'(P)+D'(P)]\,U_x(P)\}_{\bar{x}} + \{[B'_1(P)+E'_1(P)]\,U_y(P)\}_{\bar{x}} + \\ & + \{[B'_2(P)+E'_2(P)]\,U_{\bar{y}}(P)\}_x + \{[B'_1(P)+E'_1(P)]\,U_x(P)\}_{\bar{y}} + \\ & + \{[B'_2(P)+E'_2(P)]\,U_{\bar{x}}(P)\}_y + \{[C'(P)+G'(P)]\,U_y(P)\}_{\bar{y}} - \\ & - H'_1(P)\,U_x(P) + [H'_1(P)\,U(P)]_{\bar{x}} - H'_2(P)\,U_{\bar{x}}(P) + [H'_2(P)\,U(P)]_x - \\ & - K'_1(P)\,U_y(P) + [K'_1(P)\,U(P)]_{\bar{y}} - K'_2(P)\,U_{\bar{y}}(P) + [K'_2(P)\,U(P)]_y - \\ & - F'(P)\,U(P). \end{aligned}$$

Коэффициенты оператора L'_h определяются соотношениями

$$A'(P) = h^{-2}\iint\limits_{D_h} a\,dx\,dy, \quad B'_1(P) = h^{-2}\iint\limits_{T_h^1} b\,dx\,dy,$$

$$B'_2(P) = h^{-2}\iint\limits_{T_h^2} b\,dx\,dy, \quad C'(P) = h^{-2}\iint\limits_{E_h} c\,dx\,dy,$$

$$D'(P) = h^{-2}\left[\iint\limits_{T_h^1} f(x-x_P)^2\,dx\,dy + \iint\limits_{T_h^3} f(x-x_E)^2\,dx\,dy\right],$$

(10)

$$E'_1(P) = h^{-2}\iint\limits_{T_h^1} f(x-x_P)(y-y_P)\,dx\,dy, \quad E'_2(P) = h^{-2}\iint\limits_{T_h^2} f(x-x_P)(y-y_P)\,dx\,dy,$$

$$G'(P) = h^{-2}\left[\iint\limits_{T_h^1} f(y-y_P)^2\,dx\,dy + \iint\limits_{T_h^4} f(y-y_N)^2\,dx\,dy\right],$$

$$H'_1(P) = h^{-2}\iint\limits_{T_h^1} f(x-x_P)\,dx\,dy, \quad H'_2(P) = h^{-2}\iint\limits_{T_h^2} f(x-x_P)\,dx\,dy,$$

$$K'_1(P) = h^{-2}\iint\limits_{T_h^1} f(y-y_P)\,dx\,dy, \quad K'_2(P) = h^{-2}\iint\limits_{T_h^2} f(y-y_P)\,dx\,dy,$$

$$F'(P) = h^{-2}\left[\iint\limits_{T_h^1} f\,dx\,dy + \iint\limits_{T_h^2} f\,dx\,dy\right].$$

Здесь D_h — четырехугольник, ограниченный прямыми $x = x_P$, $x+y = x_p + y_P + h$, $x = x_P + h$, $x+y = x_P + y_P$; T_h^1 — треугольник, ограниченный прямыми $x = x_P$, $y = y_P$, $x+y = x_P + y_P + h$; T_h^2 — треугольник, ограниченный прямыми $x = x_P$, $y = y_P$, $x+y = x_P + y_P - h$; E_h — четырехугольник, ограниченный прямыми $y = y_P$, $x+y = x_P + y_P + h$, $y = y_P + h$, $x+y = x_P + y_P$; T_h^3 — треугольник, ограниченный прямыми $x = x_P + h$, $y = y_P$, $x+y = x_P + y_P$; T_h^4 — треугольник, ограниченный прямыми $x = x_P$, $y = y_P + h$, $x+y = x_P + y_P$.

Мы обозначим через $\lambda^1 \leqq \lambda^2 \leqq \ldots$ собственные значения задачи (8), а через $\mu_h^1 \leqq \mu_h^2 \leqq \ldots$ собственные значения задачи (9).

Теорема 3. *Пусть R — произвольная открытая двумерная область, граница которой C состоит из конечного числа кусочно-аналитических простых замкнутых кривых. Предположим, что коэффициенты оператора L аналитичны в некоторой открытой области, содержащей R с границей. Тогда при $k = 1, 2, \ldots$ справедлива двусторонняя оценка*

$$-c_5 h < \lambda^k - \mu_h^k \leqq 0, \tag{11}$$

где c_5 — положительная постоянная, зависящая только от k, от области R и от коэффициентов оператора L.

Доказательство. Собственные значения λ^k являются стационарными значениями частного Рэлея (см., например, [4], стр. 186).

$$\varrho(v) = \frac{\iint\limits_R (av_x^2 + 2bv_x v_y + cv_y^2 + fv^2)\,dx\,dy}{\iint\limits_R v^2\,dx\,dy},$$

где $v = v(x, y)$ — любая функция, удовлетворяющая следующим условиям:

1. v непрерывна на множестве $\overline{R} = R \cup C$;

2. вектор-функция $\nabla v = (v_x, v_y)$ кусочно-непрерывна в R, а $|\nabla v|^2$ интегрируем по Лебегу на $\overline{R}$.

3. $v = 0$ на G.

Пусть

$$V(P) = \sum_{i=1}^{k} t_i V^i(P),$$

где $t_1, \ldots, t_k$ — действительные параметры, $V^1, \ldots, V^k$ — любые линейно-независимые сеточные функции, обращающиеся в нуль в узлах C_h.

Разобьем каждую ячейку сетки диагональю, параллельной прямой $x+y = 0$, на два треугольника. Обозначим через $v(x, y)$ функцию, которая в каждом треугольнике линейна, совпадает с $V(P)$ в узлах R_h и обращается в нуль вне R^*.

Если вершинами треугольника T_h^1 являются точки $P = (x_P, y_P)$, $E = (x_P + h, y_P)$, $N = (x_P\ y_P + h)$, то для $(x, y) \in T_h^1$ функция $v(x, y)$ определяется равенством $v(x, y) = h^{-1}\{V(P)h + [V(E) - V(P)]\ (x - x_P) + [V(N) - V(P)]\ (y - y_P)\}$. Нетрудно проверить, что

$$\iint\limits_{T_h^1} v^2\,dx\,dy = \frac{h^2}{12}\{[V(P)]^2 + [V(E)]^2 + [V(N)]^2 + V(P)\,V(E) + V(P)\,V(N) + V(E)\,V(N)\} \tag{12}$$

и

$$\iint\limits_{T_h^1} (av_x^2+2bv_xv_y+cv_y^2+fv^2)\,dx\,dy = [A(P)+D(P)][V(E)-V(P)]^2+$$

$$2[B(P)+E(P)][V(E)-V(P)][V(N)-V(P)]+[C(P)+G(P)][V(N)-V(P)]^2+$$

(13)

$$+2hH(P)V(P)[V(E)-V(P)]+2hK(P)V(P)[V(N)-V(P)]+h^2F(P)[V(P)]^2,$$

где

$$A(P)=h^{-2}\iint\limits_{T_h^1} a\,dx\,dy,\quad B(P)=h^{-2}\iint\limits_{T_h^1} b\,dx\,dy,$$

$$C(P)=h^{-2}\iint\limits_{T_h^1} c\,dx\,dy,\quad D(P)=h^{-2}\iint\limits_{T_h^1} f(x-x_P)^2\,dx\,dy,$$

$$E(P)=h^{-2}\iint\limits_{T_h^1} f(x-x_P)(y-y_P)\,dx\,dy,\quad G(P)=h^{-2}\iint\limits_{T_h^1} f(y-y_P)^2\,dx\,dy,$$

$$H(P)=h^{-2}\iint\limits_{T_h^1} f(x-x_P)\,dx\,dy,\quad K(P)=h^{-2}\iint\limits_{T_h^1} f(y-y_P)\,dx\,dy,$$

$$F(P)=h^{-2}\iint\limits_{T_h^1} f\,dx\,dy.$$

Суммирование выражений (12) по всем треугольникам сетки приводит к соотношению

$$\iint\limits_{R^*} v^2\,dx\,dy = J_h'(V)=\frac{h^2}{12}\sum_{P\in R_h} V(P)[6V(P)+V(E)+V(SE)+V(S)+V(NW)+$$

$$+V(N)] = h^2\sum_{P\in R_h}\Big\{[V(P)]^2-\frac{1}{12}[V(E)-V(P)]^2-\frac{1}{12}[V(S)-V(P)]^2-$$

(14)

$$-\frac{1}{12}[V(SE)-V(P)]^2\Big\}.$$

Суммируя (13) по сетке, мы найдем, что

$$\iint\limits_{R^*} (av_x^2+2bv_xv_y+cv_y^2+fv^2)\,dx\,dy = I_h'(V)=$$

$$=\sum_{P\in R_h}\{[A'(P)+D'(P)][V(E)-V(P)]^2+2[B_1'(P)+E_1'(P)][V(E)-V(P)]\times$$

$$\times[V(N)-V(P)]+2[B_2'(P)+E_2'(P)][V(W)-V(P)][V(S)-V(P)]+$$

$$+[C'(P)+G'(P)][V(N)-V(P)]^2+2hH_1'(P)V(P)[V(E)-V(P)]+$$

(15)

$$+2hH_2'(P)V(P)[V(P)-V(W)]+2hK_1'(P)V(P)[V(N)-V(P)]+$$

$$+2hK_2'(P)V(P)[V(P)-V(S)]+h^2F'(P)[V(P)]^2\}.$$

Здесь коэффициенты $A'(P)$, $B'(P)$ и т. д. определяются соотношениями (10).

Легко показать, что собственные значения задачи (9) являются стационарными значениями частного Рэлея

$$\varrho_h'(V) = \frac{I_h'(V)}{J_h'(V)}. \tag{16}$$

Пусть V^i $(i=1, 2, \ldots, k)$ — сеточная функция, для которой частное Рэлея (16) принимает стационарное значение μ_h^i. Тогда, применяя неравенство Пуанкаре (см. [5] и [6], стр. 63), получим

$$\lambda^k \leqq \mu_h^k. \tag{17}$$

Пусть $\lambda_h^1 \leqq \lambda_h^2 \leqq \ldots$ — собственные значения задачи

$$L_h U(P) + \lambda_h U(P) = 0, \quad P \in R_h, \qquad U(P) = 0, \quad P \in C_h, \tag{18}$$

где

$$L_h U(P) = 0{,}5\{[a(P)\,U_x(P)]_{\bar{x}} + [a(P)\,U_{\bar{x}}(P)]_x + [b(P)\,U_y(P)]_{\bar{x}} + [b(P)\,U_{\bar{y}}(P)]_x +$$

$$+ [b(P)\,U_x(P)]_{\bar{y}} + [b(P)\,U_{\bar{x}}(P)]_y + [c(P)\,U_{\bar{y}}(P)]_y + [c(P)\,U_y(P)]_{\bar{y}}\} - f(P)\,U(P).$$

При предположениях теоремы 3 имеет место оценка (см. [7])

$$\lambda^k - \lambda_h^k = O(h). \tag{19}$$

Собственные значения задачи (18) являются стационарными значениями частного Рэлея

$$\varrho_h(V) = \frac{I_h(V)}{J_h(V)}.$$

где

$$J_h(V) = h^2 \sum_{P \in R_h} [V(P)]^2$$

и

$$I_h(V) = 0{,}5h^2 \sum_{P \in R_h} \{a(P)\big[(V_x(P))^2 + (V_{\bar{x}}(P))^2\big] + 2b(P)[V_x(P)\,V_y(P) +$$

$$+ V_{\bar{x}}(P)\,V_{\bar{y}}(P)] + c(P)\big[(V_y(P))^2 + (V_{\bar{y}}(P))\big]^2 +$$

$$+ 2f(P)[V(P)]^2\} = \sum_{P \in R_h} \{a'(P)[V(E) - V(P)]^2 + b(P)[V(E) - V(P)] \times$$

$$\times [V(N) - V(P)] + b(P)[V(W) - V(P)][V(S) - V(P)] + c'(P)[V(N) - V(P)]^2 +$$

$$+ h^2 f(P)[V(P)]^2\}.$$

Здесь $$a'(P) = 0{,}5[a(E) + a(P)], \quad c'(P) = 0{,}5[c(N) + c(P)].$$

С помощью теоремы Тэйлора легко показать, что

$$A'(P) = a'(P) + O(h), \quad 2B_1'(P) = b(P) + O(h),$$
$$2B_1'(P) = b(P) + O(h), \quad C'(P) = c'(P) + O(h). \tag{20}$$

Далее

$$\sum_{P\in R_h}\{D'(P)[V(E)-V(P)]^2+2E_1'(P)[V(E)-V(P)][V(N)-V(P)]+$$

$$+2E_2'(P)[V(W)-V(P)][V(S)-V(P)]+G'(P)[V(N)-V(P)]^2+$$

$$+2hH_1'(P)V(P)[V(E)-V(P)]+2hH_2'(P)V(P)[V(P)-V(W)]+ \tag{21}$$

$$+2hK_1'(P)V(P)[V(N)-V(P)]+2hK_2'(P)V(P)[V(P)-V(S)]+h^2F'(P)[V(P)]^2\}=$$

$$=h^2\sum_{P\in R_h}f(P)[V(P)]^2+O\big(hI_h(V)\big)+O\big(hJ_h(V)\big).$$

Подставляя (20) и (21) в (15), находим, что

$$I_h'(V)=I_h(V)+O\big(hI_h(V)\big)+O\big(hJ_h(V)\big). \tag{22}$$

С другой стороны, из (14) следует, что

$$J_h'(V)\geqq h^2\sum_{P\in R_h}[V(P)]^2-\frac{1}{4}h^2\sum_{P\in R_h}\{[V(E)-V(P)]^2+[V(N)-V(P)]^2\}= \tag{23}$$

$$=J_h(V)+O\big(h^2I_h(V)\big).$$

Из (22) и (23) мы видим, что

$$\mu_h^k\leqq\lambda_h^k+O(h). \tag{24}$$

Сопоставляя неравенства (17), (19) и (24), получаем желаемую оценку (11).

БИБЛИОГРАФИЯ

[1] PÓLYA, G.: Sur une interprétation de la méthode des différences finies qui peut fournir des bornes supérieures et inférieures, *C. R. Acad. Sci. Paris* **235** (1952) 995—997.

[2] WEINBERGER, H. F.: Lower bounds for higher eigenvalues by finite-difference methods, *Pacific J. Math.* **8** (1958) 339—368.

[3] VEIDINGER, L.: О вычислении собственных значений мембраны методом конечных разностей, *Ж. Вычисл. Мат. и Мат. Физ.* **4** (1964) 1037—1044.

[4] Вазов, В. и Форсайт, Дж.: *Разностные методы решения дифференциальных уравнений в частных производных,* Издательство Ин. Лит., Москва, 1963.

[5] POINCARÉ, H.: Sur les équations aux derivées partielles de la physique mathématique, *Amer. J. Math.* **12** (1890) 211—294.

[6] GOULD, S. H.: *Variational methods for eigenvalue problems,* Univ. Toronto Press, Toronto, 1957.

[7] VEIDINGER, L.: Об оценке погрешности при нахождении собственных значений методом конечных разностей, *Ж. Вычисл. Мат. и Мат. Физ.* **5** (1965) 806—815.

МАТЕМАТИЧЕСКИЙ ИНСТИТУТ ВЕНГЕРСКОЙ АКАДЕМИИ НАУК, БУДАПЕШТ

(Поступило 12-ого августа 1966 г.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 201—213.

ÜBER EINE ERWEITERUNG DES ZYGMUNDSCHEN APPROXIMATIONSPROZESSES IN ZWEI DIMENSIONEN

von

M. SALLAY

1.

Es sei $f(x)$ stetige, nach 2π periodische Funktion, weiter $\{T_n\}$ eine Folge von linearen beschränkten Transformationen, die den Raum $C_{2\pi}$ in den Raum der trigonometrischen Polynome höchstens n-ter Ordnung abbilden. Die Folge $\{T_n\}$ nennen wir ZYGMUNDsche Approximationsfolge, falls für $f \in C_{2\pi}$

$$|f(x) - T_n\{f; x\}| \leqq c\,\omega_2(f; n^{-1}), \qquad (c = \text{Konst.})$$

besteht, wo

$$\omega_2(f; \delta) = \max_{\substack{|h| \leqq \delta \\ x \in [0,\, 2\pi]}} |f(x+h) + f(x-h) - 2f(x)|$$

der Stetigkeitsmodul zweiter Ordnung von $f(x)$ ist. In der Arbeit [2] hat G. FREUD gezeigt, daß eine Folge $\{T_n\}$ dann und nur dann eine ZYGMUNDsche Approximationsfolge darstellt, falls die folgenden Bedingungen erfüllt sind:

a) $|T_n\{f; x\}| \leqq c_1 \max_x |f(x)|$,

für $f \in C_{2\pi}$,

b) $|T_n\{g; x\} - g(x)| \leqq c_2 n^{-2} \max_x |g''(x)|$,

für jedes zweimal stetig differenzierbare, nach 2π periodische $g(x)$; wobei die Konstanten c_1 und c_2 von x, n und von der Wahl von $f(x)$ und $g(x)$ unabhängig sind.

G. FREUD hat die folgende Frage gestellt (vgl. [4] S. 182): „Wäre es möglich, diese Aussage auf mehrere Dimensionen zu erweitern? In der Bedingung b) müßte eine Summe über die Normen der zweiteren partiellen Ableitungen auftreten; die Schwierigkeit aber dürfte in einer geeigneten Definition von ω_2 liegen."

In unserer Arbeit versuchen wir eine Definition von ω_2 in zwei Dimensionen zu geben, mit deren Hilfe wir das Problem von G. FREUD lösen können.

2.

Es sei D ein beschränktes Gebiet der Ebene, es seien $P(x, y) = P$ die Punkte von D und $f(P)$ in dem Gebiet D stetige Funktion. Wir betrachten einen beliebigen Punkt P von D und schreiben um den Punkt P einen Kreis R_ϱ mit dem Radius ϱ. Wir bezeichnen die Eckpunkte des Durchmessers von R_ϱ in der Richtung α mit $P^{\alpha}_{-\varrho}$, resp. $P^{\alpha}_{+\varrho}$.[1] Wir definieren den Stetigkeitsmodul zweiter Ordnung von $f(P)$

[1] Es wird $R_\varrho \subset D$ nicht verlangt, es werden nur die Richtungen α beachtet für welche $P^{\alpha}_{\varrho}, P^{-\alpha}_{\varrho} \in D$ besteht.

in dem Gebiet D folgenderweise:

$$(1) \qquad \underset{D}{\omega_2}(f;\delta) \overset{\text{def}}{=} \sup_{\substack{0\leqq\alpha\leqq 2\pi \\ P,P_{+\varrho},P_{-\varrho}\in D \\ \varrho\leqq\delta}} |f(P^{\alpha}_{+\varrho}+f(P^{\alpha}_{-\varrho})-2f(P)|.$$

In dem Falle, wenn $f(P)$ die Funktion $f(x, y)$ der rechtwinkeligen Koordinaten x resp. y ist, ergibt sich

$$(2) \qquad \underset{D}{\omega_2}(f;\delta) \overset{\text{def}}{=} \sup_{\substack{0\leqq\alpha\leqq 2\pi \\ \varrho\leqq\delta \\ (x,y)\in D}} |f(x+\varrho\cos\alpha, y+\varrho\sin\alpha)+f(x-\varrho\cos\alpha, y-\varrho\sin\alpha)-2f(x,y)|.$$

BEMERKUNGEN:

1) Der Stetigkeitsmodul $\underset{D}{\omega_2}(f;\delta)$ ist eine nicht abnehmende Funktion von δ resp. D, d. h.

$$\underset{D}{\omega_2}(f;\delta_1) \leqq \underset{D}{\omega_2}(f;\delta_2),$$

für $\delta_1 < \delta_2$;

$$\underset{D_1}{\omega_2}(f;\delta) \leqq \underset{D_2}{\omega_2}(f;\delta)$$

für $D_1 \subset D_2$.

2) Wir können (2) wie die Erweiterung des wohlbekannten Stetigkeitsmoduls zweiter Ordnung betrachten; es seien z.B. $f(x, y) \equiv f(x)$ und D das Interval $[a, b]$, dann ist

$$\omega_2(f;\delta) = \sup_{\substack{\varrho\leqq\delta \\ x\in[a,b]}} |f(x+\varrho)+f(x-\varrho)-2f(x,y)|.$$

3) Es sei α eine festgesetzte Richtung, E_α eine Gerade in der Richtung α. Betrachten wir die zu dem Gebiet D gehörenden Strecken von E_α. Aus den Bemerkungen 1. und 2. folgt

$$(3) \qquad \omega_2(f;\delta) \leqq \sup_{0\leqq\alpha\leqq 2\pi} \underset{E_\alpha}{\omega_2}(f;\delta) = \underset{D}{\omega_2}(f;\delta).$$

4) Es sei vorausgesetzt, daß die partiellen Ableitungen zweiter Ordnung von $f(x, y)$ in D existieren und stetig sind. Gemäß der Definition der Ableitung in Richtung α ist

$$|f(x+\varrho\cos\alpha, y+\varrho\sin\alpha)+f(x-\varrho\cos\alpha, y-\varrho\sin\alpha)-2f(x,y)| =$$

$$= \left|\varrho^2\left(\cos^2\alpha\,\frac{\partial^2 f}{\partial x^2}+2\cos\alpha\sin\alpha\,\frac{\partial^2 f}{\partial x\,\partial y}+\sin^2\alpha\,\frac{\partial^2 f(x,y)}{\partial y^2}\right)\right|+o(\varrho^2),$$

und hieraus folgt, daß

$$\underset{D}{\omega_2}(f;\delta) \leqq \delta^2\left(\max_{x,y\in D}\left|\frac{\partial^2 f(x,y)}{\partial x^2}\right|+2\max_{x,y\in D}\left|\frac{\partial^2 f(x,y)}{\partial x\,\partial y}\right|+\right.$$

$$\left.+\max_{x,y\in D}\left|\frac{\partial^2 f(x,y)}{\partial y^2}\right|\right)$$

mit $\delta_1 = \delta\cos\alpha$, $\delta_2 = \delta\sin\alpha$.

5) Es ist aus (3) und aus dem eindimensionalen Falle leicht ersichtlich, daß für $\vartheta > 1$

$$\underset{D}{\omega_2}(f; \vartheta\delta) \leqq (2\vartheta)^2 \underset{D}{\omega_2}(f; \delta)$$

ist.

6) Es sei D ein n-dimensionales Gebiet, und $f(P)$ eine Funktion in D; wir können die Definition (1) ohne Schwierigkeit in n Dimension erweitern. Wir müssen nur statt R_ϱ eine n-dimensionale Kugel mit dem Radius ϱ betrachten.

3.

Es sei D_ϱ ein Parallelogramm — mit dem Mittelpunkt $(x_0, y_0) = P_0$ und mit der Kantenlänge $2\varrho_1$, resp. $2\varrho_2$, $(\varrho = \sqrt{\varrho_1^2 + \varrho_2^2})$ — deren Seiten mit den Achsen x resp. y parallel sind. Es seien die Koordinaten der Eckpunkte P_i $(i = 1, 2, 3, 4)$ wie folgt:

$$P_1 : (x_0 + \varrho_1,\ y_0 + \varrho_2); \quad P_2 (x_0 + \varrho_1,\ y_0 - \varrho_2); \quad P_3 : (x_0 - \varrho_1,\ y_0 + \varrho_2);$$

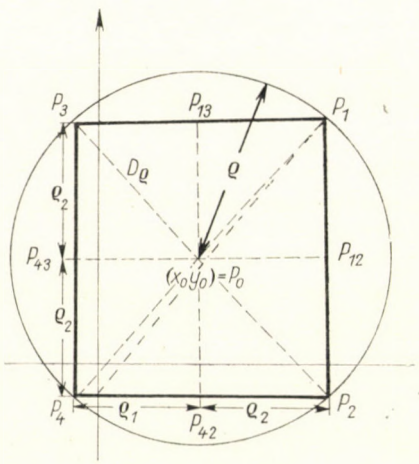
$$P_4 : (x_0 - \varrho_1,\ y_0 - \varrho_2); \quad \varrho = \sqrt{\varrho_1^2 + \varrho_2^2}; \quad \varrho_1 = \varrho \cos\alpha, \quad \varrho_2 = \varrho \sin\alpha.$$

LEMMA 1.[2] Es sei $\varphi(x, y)$ eine in dem Gebiet D_ϱ in beiden Veränderlichen stetige Funktion, die in den Eckpunkten von D_ϱ verschwindet, d.h.:

$$\varphi(P_i) = 0, \qquad (i = 1, 2, 3, 4). \tag{4}$$

Dann gilt

$$\max_{D_\varrho} |\varphi(x, y)| \leqq 2\underset{D_\varrho}{\omega_2}(\varphi; \varrho).$$



Figur 1

BEWEIS. Es nehme die Funktion $|\varphi(x, y)|$ ihr Maximum in dem Punkt (ξ, η) an.

Es sei ferner

$$\Omega(\varphi, \varrho) \overset{\text{def}}{=} |\varphi(x + \varrho_1,\ y + \varrho_2) +$$

$$+ \varphi(x - \varrho_1,\ y - \varrho_2) + \varphi(x - \varrho_1,\ y + \varrho_2) + \tag{5}$$

$$+ \varphi(x + \varrho_1,\ y - \varrho_2) - 4\varphi(x, y)|.$$

Wir setzen in (5) für (x, y), ϱ_1, ϱ_2 nacheinander die folgenden Werte ein:

$$x = \xi \quad \varrho_1 = \xi - (x_0 - \varrho_1) \quad \text{für} \quad \xi \leqq x_0$$

$$x = \xi \quad \varrho_1 = (x_0 + \varrho_1) - \xi \quad \text{für} \quad \xi > x_0$$

$$y = \eta \quad \varrho_2 = \eta - (y_0 - \varrho_2) \quad \text{für} \quad \eta \leqq y_0$$

$$y = \eta \quad \varrho_2 = (y_0 + \varrho_2) - \eta \quad \text{für} \quad \eta > y_0.$$

[2] Dieser Hilfssatz ist die Verallgemeinerung auf zwei Dimensionen des Lemmas 5. 2 von H. BURKILL in [1]. Unser Beweis geht nach der Idee des Burkillschen Beweises. Es ist auch ohne Schwierigkeit möglich, den Hilfssatz auf n-Dimensionen zu erweitern.

Nach der Einsetzung erhalten wir z.B. in dem Fall $\xi \leqq x_0$, $\eta > y_0$

$$\Omega(\varphi, \varrho) = |\varphi(2\xi - (x_0 - \varrho_1), y_0 + \varrho_2) + \varphi(x_0 - \varrho_1, 2\eta - (y_0 + \varrho_2)) +$$
$$+ \varphi(x_0 - \varrho_1, y_0 + \varrho_2) + \varphi(2\xi - (x_0 - \varrho_1), \eta - (y_0 + \varrho_2)) - 4\varphi(\xi, \eta)|.$$

Nach der Bedingung (4) ergibt sich

$$\Omega(\varphi, \varrho) = |\varphi(\xi, \eta) + [3\varphi(\xi, \eta) - \varphi(2\xi - (x_0 - \varrho_1), y_0 + \varrho_2) -$$
$$- \varphi(2\xi - (x_0 - \varrho_1), 2\eta - (y_0 + \varrho_2)) - \varphi(x_0 - \varrho_1, 2\eta - (y_0 + \varrho_2))]| \geqq |\varphi(\xi, \eta)|,$$

woraus wegen $\Omega(\varphi, \varrho) \leqq 2\omega_2(\varphi, \varrho)$ unsere Behauptung folgt. In den anderen Fällen geht der Beweis ähnlicherweise.

LEMMA 2. Man kann zu jeder Funktion $f(x, y)$ in D_ϱ ein bilineares Polynom der Gestalt

$$p(x, y) = (a_1 x + b_1)(a_2 y + b_2) \tag{6}$$

konstruieren, für welches

$$\max_{D_\varrho} |f(x, y) - p(x, y)| \leqq 4\omega_2(f; \varrho), \quad \varrho = \sqrt{\varrho_1^2 + \varrho_2^2} \tag{7}$$

gültig ist.

BEWEIS: Es nehme die Funktion $p(x, y)$ in den Eckpunkten des Parallelogramms D_ϱ die gleichen Werte wie $f(x, y)$ an. Da die Funktion $\varphi(x, y) = f(x, y) - p(x, y)$ in den Eckpunkten verschwindet, ergibt sich aus dem Lemma 1

$$\max_{D_\varrho} |f(x, y) - p(x, y)| \leqq 2\omega_2(f - p; \varrho) \leqq 2(\omega_2(f; \varrho) + \omega_2(p; \varrho)). \tag{8}$$

Bezeichnen wir mit f_i den Wert von $f(x, y)$ in den Eckpunkten P_i und schreiben wir $p(x, y)$ in der Form

$$p(x, y) = \frac{1}{4\varrho_1 \varrho_2}[f_1 + f_4 - f_2 - f_3](x - x_0)(y - y_0) +$$
$$+ \frac{1}{4\varrho_1}[f_2 - f_1 - f_4 - f_3](x - x_0) + \frac{1}{4\varrho_2}[f_3 - f_1 - f_4 - f_2](y - y_0) + \frac{1}{4}\sum_{i=1}^{4} f_i. \tag{9}$$

Setzen wir (9) in das zweite Glied der rechten Seite von (8) ein. Aus der Definition (2) folgt

$$\omega_2(p; \varrho) \leqq \tfrac{1}{2} \max_{D_\varrho} |f_1 + f_4 - f_2 - f_3| \leqq$$
$$\leqq \tfrac{1}{2}\{\max_{D_\varrho} |f_1 + f_4 - 2f(x_0, y_0)| + \max_{D_\varrho} |f_2 + f_3 - 2f(x_0, y_0)|\} \leqq \omega_2(f; \varrho)$$

und hieraus erhalten wir (7).

LEMMA 3. Man kann zu jeder Funktion $f(x, y)$ in D_ϱ ein Polynom der Gestalt

$$r(x, y) = (a_1 x^2 + b_1 x + c_1)(a_2 x^2 + b_2 y + c_2) \tag{10}$$

konstruieren, für welche die Ungleichungen

$$\max_{D_\varrho} |r_{xx}(x,y)| \leqq \frac{2}{\varrho_1^2}\,\omega_{2\,D_\varrho}(f;\varrho)$$

$$\max_{D_\varrho} |r_{xy}(x,y)| \leqq \frac{4}{\varrho_1\varrho_2}\,\omega_{2\,D_\varrho}(f;\varrho) \tag{11}$$

$$\max_{D_\varrho} |r_{yy}(x,y)| \leqq \frac{2}{\varrho_2^2}\,\omega_{2\,D_\varrho}(f;\varrho)$$

$$\max_{D_\varrho} |r(x,y)-p(x,y)| \leqq 2\omega_{2\,D_\varrho}(f;\varrho) \tag{12}$$

bestehen.

BEWEIS. Bezeichnen wir mit f_{ij} $(i=1,4;\ j=2,3)$ die Werte von $f(P)$ in den Mittelpunkten P_{ij} der Seiten von D_ϱ, ferner mit f_0 den Wert von $f(x_0,y_0)=f(P_0)$. Das Polynom $r(x,y)$ sei folgenderweise beschaffen:

$$r(P_i)=f_i, \qquad (i=1,2,3,4)$$

$$r(P_{ij})=\frac{1}{4}(f_i+f_j+2f_{ij}); \qquad (i=1,4;\ j=2,3)$$

$$r(P_0)=\frac{1}{8}\left(\sum_{i=1}^{4} f_i+4f_0\right)$$

Durch Ausrechnen erhalten wir:

$$\begin{aligned} r(x,y)=\Bigg\{ & \frac{1}{8\varrho_1^2\varrho_2^2}\big((f_1+f_2-2f_{12})+(f_3+f_4-2f_{43})- \\ & -2(f_{13}+f_{42}-f_0)\big)(x-x_0)^2(y-y_0)^2 \\ & +\frac{1}{8\varrho_1^2\varrho_2}\big((f_1+f_2-2f_{12})-(f_2+f_4-2f_{42})\big)(x-x_0)^2(y-y_0) \\ & +\frac{1}{8\varrho_1\varrho_2^2}\big((f_1+f_3-2f_{13})-(f_2+f_4-2f_{42})\big)(x-x_0)(y-y_0)^2 \\ & +\frac{1}{4\varrho_1\varrho_2}\big((f_1+f_4-2f_0)-(f_2+f_3-2f_0)\big)(x-x_0)(y-y_0) \\ & +\frac{1}{4\varrho_1^2}(f_{12}+f_{43}-2f_0)(x-x_0)^2+\frac{1}{4\varrho_2^2}(f_{13}+f_{42}-2f_0)(y-y_0)^2 \\ & +\frac{1}{8\varrho_1}(f_1+f_2-f_3-f_4+2f_{12}-2f_{43})(x-x_0)+ \\ & +\frac{1}{8\varrho_2}(f_1+f_3-f_2-f_4+2f_{13}-2f_{42})(y-y_0)+\frac{1}{8}(f_1+f_2+f_3+f_4+4f_0), \end{aligned}$$

woraus nach den Relationen $|x-x_0| \leqq \varrho_1$, $|y-y_0| \leqq \varrho_2$ und wegen der Bemerkung (3) die Abschätzungen (11) folgen.

Man kann auch die Ungleichung (12) mit Rücksicht auf die Darstellung (9) von $p(x, y)$ leicht beweisen.

4.

Es sei im weiteren das Gebiet D_ϱ das Quadrat $D = \{-1 \leqq x \leqq 1,\ -1 \leqq y \leqq 1\}$. Es bezeichne $\mathscr{C}_D$ die Klasse der in D stetigen Funktionen. Es sei $\{A_{nm}\}$ eine von den Parametern n, m (n und m sind ganze Zahlen) abhängige lineare Transformationsfolge, welche den Raum $\mathscr{C}_D$ in sich selbst transformiert. Es sei die Norm des Operators

$$\|A_{nm}\| = \max_{\substack{x,y \in D \\ \|f(x,y)\| \leqq 1}} |A_{nm}\{f; x, y\}|. \tag{13}$$

Setzen wir weiter voraus, daß

$$\|A_{nm}\| \leqq \varkappa_1 \tag{14}$$

besteht, wo $\varkappa_1$ eine von n und m unabhängige Konstante ist.

Wir nennen die Approximationsfolge $\{A_{nm}\}$ — dem eindimensionalen Fall entsprechend — Zygmundsche Approximationsfolge, falls für $f(x, y) \in \mathscr{C}_D$ die Relation

$$\|f(x, y) - A_{nm}\{f; x, y\}\| \leqq K\omega_2\left(f; \sqrt{\frac{1}{n^2}+\frac{1}{m^2}}\right) \tag{15}$$

befriedigt ist, wo K eine von f, n, m und A_{nm} unabhängige Konstante ist.

Es bezeichne $\mathscr{C}_D^2$ die Klasse der zu dem Raume $\mathscr{C}_D$ gehörenden Funktionen, deren partielle Ableitungen zweiter Ordnung existieren und stetig sind. Wir setzen voraus, daß für $f \in \mathscr{C}_D^2$

$$\|f - A_{nm}f\| \leqq \varkappa_2\left(\frac{1}{n^2}+\frac{1}{m^2}\right)\left\{\|f_{xx}\| + 2\|f_{xy}\| + \|f_{yy}\|\right\} \tag{16}$$

gültig ist, wo $\varkappa_2$ eine von f, n, m und A_{nm} unabhängige Konstante ist.

SATZ 1. *Eine Folge $\{A_{nm}\}$ linearer Transformationen stellt dann und nur dann eine Zygmundsche Approximationsfolge dar, falls die beiden Bedingungen* (14) *und* (16) *erfüllt sind.*

Es sei nun $f(x, y)$ in beiden Veränderlichen stetige und nach 2π periodische Funktion, und $\{T_{nm}\}$ eine Folge linearer Transformationen, die den Raum $C_{2\pi}$ der stetigen Funktionen in den Raum der trigonometrischen Polynome höchstens n resp. m-ter Ordnung transformiert. Wir setzen voraus, daß für $\{A_{nm}\} = \{T_{nm}\}$ die Bedingungen (14) und (16) erfüllt sind.

SATZ 2. *Die Bedingungen* (14) *und* (16) *sind hinreichend und notwendig dafür, daß*

$$\|f(x, y) - T_{nm}\{f; x, y\}\| \leqq 2\big(\varkappa_2 + 2(\varkappa_1+1)\big)\pi^2\omega_2\left(f; \sqrt{\frac{1}{n^2}+\frac{1}{m^2}}\right) \tag{17}$$

erfüllt sei.

Wir zeigen zuerst, daß die Klasse der Operatoren, die die Bedingungen (14) und (16) befriedigen, nicht leer ist.

Es sei $f(x, y)$ eine in beiden Veränderlichen nach 2π periodische Funktion, deren partielle Ableitungen zweiter Ordnung in dem Gebiet $D=\{-\pi\leqq x\leqq\pi, -\pi\leqq y\leqq\pi\}$ existieren und stetig sind.

Es sei

$$\tau_{nm}\{f; x, y\} = k\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(u, v)\left(\frac{\sin\frac{n}{2}(x-u)}{\sin\frac{1}{2}(x-u)}\right)^4\left(\frac{\sin\frac{m}{2}(y-v)}{\sin\frac{1}{2}(y-v)}\right)^4 du\,dv$$

mit

$$k = \left(\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left(\frac{\sin\frac{n}{2}(x-u)}{\sin\frac{1}{2}(x-u)}\right)^4\left(\frac{\sin\frac{m}{2}(y-v)}{\sin\frac{1}{2}(y-v)}\right)^4 du\,dv\right)^{-1}.$$

Aus (13) ist es offenbar, daß für die Norm des Operators τ_{nm}

$$\|\tau_{nm}\|\leqq 1$$

besteht, wir müssen noch zeigen, daß für $f(x, y)\in C^2_{2\pi}$ die Bedingung (16) erfüllt ist.

Wohlbekannterweise kann man zeigen, daß

$$|f(x,y)-\tau_{nm}\{f; x, y\}| =$$

$$= \left|\frac{k}{4}\int_0^{\frac{\pi}{2}}\int_0^{\frac{\pi}{2}}(f(x+u,y+v)+f(x+u,y-v)+f(x-u,y+v)+\right.$$

$$\left.+f(x-u,y-v)-4f(x,y))\left(\frac{\sin nu}{\sin u}\frac{\sin mv}{\sin v}\right)^4 du\,dv\right|$$

ist. Nach $f(x, y)\in C^2_{2\pi}$ besteht

$$|f(x+u,y+v)+f(x-u,y+v)+f(x+u,y-v)+f(x-u,y-v)-4f(x,y)|\leqq \tag{18}$$

$$\leqq 2\operatorname*{Max}_{x,y\in D}\left|u^2\frac{\partial^2 f(x,y)}{\partial x^2}+2uv\frac{\partial^2 f(x,y)}{\partial x\,\partial y}+v^2\frac{\partial^2 f(x,y)}{\partial y^2}\right|.$$

Aus den bekannten trigonometrischen Relationen (vgl. z.B. [6] Seite 87—89)

$$\int_0^{\frac{\pi}{2}}\left(\frac{\sin\mu t}{\sin t}\right)^4 dt = \frac{\pi\mu(2\mu^2+1)}{6},$$

$$\int_0^{\frac{\pi}{2}} t\left(\frac{\sin\mu t}{\sin t}\right)^4 dt \leqq \frac{\pi^2\mu^2}{4},$$

$$\int_0^{\frac{\pi}{2}} t^2\left(\frac{\sin\mu t}{\sin t}\right)^4 dt \leqq \frac{\pi^3\mu}{8}$$

und nach (18) bekommen wir die Relation (16).

BEWEIS DES SATZES 1.

Es sei $\{A_{nm}\}$ eine Zygmundsche Approximationsfolge, dann konvergiert $A_{nm}\{f; x, y\}$ in $\mathscr{C}$ für jede festgestellte Funktion $f(x, y)$ gleichmäßig. Nach dem Banach—Steinhausschen Satz ist

$$\|A_{nm}\| \leqq \varkappa_1, \qquad (\varkappa = \text{konst.}).$$

Hieraus und aus der Bemerkung 4 folgt, daß die Bedingungen (14) und (16) notwendig sind, man muß noch zeigen, daß (14) und (16) hinreichend sind.

Teilen wir das Quadrat D mit den Geraden

$$x = \frac{2k-1}{n}, \qquad \left(k = 0, \pm 1, \pm 2, \ldots, \pm\left[\frac{n-1}{2}\right]\right);$$

$$y = \frac{2l-1}{m}, \qquad \left(l = 0, \pm 1, \pm 2, \ldots, \pm\left[\frac{m-1}{2}\right]\right),$$

so ist das Quadrat in nm Parallelogramme zerlegt.

Betrachten wir die Parallelogramme mit dem Mittelpunkt $\left(\frac{2k}{n}, \frac{2l}{m}\right)$, $\left(k = 0, \pm 1, \ldots, \pm\left[\frac{n-1}{1}\right],\ l = 0, \pm 1, \ldots, \pm\left[\frac{m-1}{2}\right]\right)$ und mit den Kantenlängen $\frac{2k}{n}, \frac{2l}{m}$, und konstruieren für jedes Parallelogramm die Funktionen $p(x, y)$ resp. $r(x, y)$, mit $\varrho_1 = \frac{1}{n}$, $\varrho_2 = \frac{1}{m}$, $\varrho = \sqrt{\frac{1}{n^2} + \frac{1}{m^2}}$.

Auf diese Art haben wir Funktionen $p_{nm}(x, y)$ resp. $r_{nm}(x, y)$ definiert, die in D stetig sind und deren partielle Ableitungen zweiter Ordnung, die Punkte der Geraden $x = \frac{2k-1}{n}$, $y = \frac{2l-1}{m}$ ausgenommen, existieren. Wir zeigen, daß man zu jedem $r_{nm}(x, y)$ eine Funktion $\gamma_{nm}(x, y)$ konstruieren kann, deren partielle Ableitungen zweiter Ordnung in D existieren und stetig sind, so daß

$$\|r_{nm}(x, y) - \gamma_{nm}(x, y)\| \leqq \varepsilon \tag{19}$$

$$\left\|\frac{\partial^2}{\partial x^2}\gamma_{nm}(x, y)\right\| \leqq \left\|\frac{\partial}{\partial x^2} r_{nm}(x, y)\right\| + \varepsilon$$

$$\left\|\frac{\partial^2}{\partial x\,\partial y}\gamma_{nm}(x, y)\right\| \leqq \left\|\frac{\partial^2}{\partial x\,\partial y} r_{nm}(x, y)\right\| + \varepsilon \tag{20}$$

$$\left\|\frac{\partial^2}{\partial y^2}\gamma_{nm}(x, y)\right\| \leqq \left\|\frac{\partial^2}{\partial y^2} r_{nm}(x, y)\right\| + \varepsilon$$

befriedigt sind.

Aus der Konstruktion von $r_{nm}(x, y)$ folgt, daß die Schnitte $x = x_0$ resp. $y = y_0$, $(x_0, y_0 \in D,\ x_0 = \text{konst}, y_0 = \text{konst.})$ aus Parabelbögen zweiter Ordnung bestehen, die in den Punkten

$$x = \frac{2k}{n}, \qquad \left(k = 0, \pm 1, \ldots, \pm\left[\frac{n-1}{2}\right]\right),$$

$$y = \frac{2l}{m}, \qquad \left(l = 0, \pm 1, \ldots, \pm\left[\frac{m-1}{2}\right]\right)$$

dieselben Werte wie $f(x, y)$ annehmen. Es ist leicht zu zeigen (vgl. z.B. [2]), daß man für eine Funktion $r_v(t)$, die in einem Intervall $[a, b]$ abschnittweise aus Parabelbögen zweiter Ordnung besteht, eine stetige und zweimal stetig differenzierbare Funktion $\gamma_v(t)$ konstruieren kann mit den Eigenschaften

$$\|r_v(t) - \gamma_v(t)\| \leqq \varepsilon_1, \quad \|\gamma_v'(t)\| \leqq \|r_v'(t)\| + \varepsilon_1, \quad \|\gamma_v''\| \leqq \|r_v''\| + \varepsilon_1. \tag{21}$$

Betrachten wir die Darstellung (10) von $r(x, y)$ und schreiben wir die Funktion $r_{nm}(x, y)$ in der Form $r_{nm}(x, y) = r_{nl}(x) r_{km}(y)$ $\left(l = 0, \pm 1, \ldots, \pm\left[\frac{n-1}{2}\right]\right.$, $\left.k = 0 \pm 1, \ldots, \pm\left[\frac{m-1}{2}\right]\right.$, wo die Funktionen $r_{nl}(x)$ resp. $r_{km}(y)$ in den Intervallen $[-1 \leqq x \leqq 1]$ resp. $[-1 \leqq y \leqq 1]$ abschnittweise aus Parabelbögen bestehen. Konstruieren wir zu $r_{km}(y)$ resp. $r_{nl}(x)$ die entsprechenden Funktionen $\gamma_{nl}(x)$ resp. $\gamma_{km}(y)$ und bilden wir die Funktion $\gamma_{nm}(x, y) = \gamma_{nl}(x)\gamma_{km}(y)$, deren partielle Ableitungen zweiter Ordnung existieren. Nach (21) ist es leicht zu beweisen, daß die Funktion $\gamma_{nm}(x, y)$ die Bedingungen (19) und (20) befriedigt.

Es sei jetzt $f(x, y) \in \mathscr{C}$, dann ist

$$\begin{aligned}\|f(x, y) - A_{nm}\{f; x, y\}\| = \|&(f(x, y) - p_{nm}(x, y)) + \\ &+ (p_{nm}(x, y) - r_{nm}(x, y)) + (r_{nm}(x, y) - \gamma_{nm}(x, y)) + \\ &+ (\gamma_{nm}(x, y) - A_{nm}\{\gamma_{nm}; x, y\}) + (A_{nm}\{\gamma_{nm}; x, y\} - A_{nm}\{r_{nm}; x, y\}) + \\ &+ (A_{nm}\{r_{nm}; x, y\} - A_{nm}\{p_{nm}; x, y\}) + \\ &+ (A_{nm}\{p_{nm}; x, y\} - A_{nm}\{f; x, y\})\|.\end{aligned}$$

Nach der Anwendung der Relationen (7), (12), (13), (16), (20) resp. (14) ist

$$\|f(x, y) - A_{nm}\{f; x, y\}\| \leqq (4(1 + \varkappa_1) + 2\varkappa_2 + \varepsilon)\omega_2\left(f; \sqrt{\frac{1}{n^2} + \frac{1}{m^2}}\right),$$

woraus für $\varepsilon \to 0$ die Behauptung des Satzes folgt.

Es sei $f(x, y) \in C_{2\pi}$; $0 \leqq x \leqq 2\pi$, $0 \leqq y \leqq 2\pi$. Wir erweitern die Funktion $f(x, y)$ in dem Gebiet $D_{2\pi}$: $\{-2\pi \leqq x \leqq 2\pi, -2\pi \leqq y \leqq 2\pi\}$ so, daß $f(x, y)$ eine gerade Funktion der Veränderlichen sei. Wir transformieren $D_{2\pi}$ in D, wo die Behauptung des Satzes 1 gilt. So haben wir auch den Satz 2 bewiesen.

5.

P. P. KOROVKIN hat in [5] den folgenden Satz gezeigt: eine folge linearer Transformationen $(A_n\{f; x\})$ befriedingt dann und nur dann die Relation

$$|A_n\{f; x\}-f(x)| \leqq \text{Konst. } \omega(f; \delta) \quad \text{mit} \quad \omega(f; \delta) = \sup_{h \leqq \delta} |f(x+h)-f(x)|,$$

wenn es die konstanten Funktionen identisch darstellt und die Funktion $g_x(t) = \sin^2 \frac{(t-x)}{2}$ in x gleichmäßig in der Größenordnung $O(n^{-2})$ approximiert.

Den Satz von P. P. KOROVKIN hat G. FREUD für den Stetigkeitsmodul zweiter Ordnung erweitert [3].

Mit Hilfe des Satzes 1 geben wir eine Erweiterung in zwei Dimensionen der Sätze von P. P. KOROVKIN [5] und G. FREUD [3].

SATZ 3. *Es sei* $f(x, y) \in \mathscr{C}$, $(x, y \in D)$, *es sei ferner vorausgesetzt, daß für die Funktionen* $f(x, y) \in \mathscr{C}_D^2$ *einen Punkt* $(\xi, \eta) \in D$ *gibt mit* $\frac{\partial}{\partial x} f(\xi, \eta) = \frac{\partial}{\partial y} f(\xi, \eta) = 0$. *Die Folge linearer Transformationen* $\{A_{nm}\}$ *stellt dann und nur dann eine Zygmundsche Approximationsfolge dar, falls die Bedingungen*

$$\begin{aligned} A_{nm}\{1\} &\equiv 1 \\ A\mu_1\mu_2\{t_i\} &= t_i + O\left(\frac{1}{\mu_i^2}\right), \qquad (i = 1, 2) \\ A\mu_1\mu_2\{t_i t_j\} &= t_i t_j + O\left(\frac{1}{\mu_i \mu_j}\right), \qquad (i, j = 1, 2) \end{aligned} \tag{22}$$

mit $t_1 = x$, $t_2 = y$, $\mu_1 = n$, $\mu_2 = m$ *erfüllt sind.*

BEWEIS. Nach dem Satz 1 ist es hinreichend zu zeigen, daß für $f(x, y) \in \mathscr{C}_D^2$ aus (21) die Relation (16) folgt. Schreiben wir die Taylorsche Reihe von $f(x, y)$ in dem Punkt $(\bar{x}, \bar{y})$ folgenderweise:

$$\begin{aligned} f(x, y) = f(\bar{x}, \bar{y}) &+ \left((x-\bar{x})\frac{\partial}{\partial x} + (y-\bar{y})\frac{\partial}{\partial y}\right) f(\bar{x}, \bar{y}) + \\ &+ \frac{1}{2}\left((x-\bar{x})\frac{\partial}{\partial x} + (y-\bar{y})\frac{\partial}{\partial y}\right)^2 f(\bar{x} + \vartheta_1(x-\bar{x}), \bar{y} + \vartheta_2(y-\bar{y})) \end{aligned} \tag{23}$$

mit $0 < \vartheta_1 < 1$, $0 < \vartheta_2 < 1$.

Wenden wir die lineare Transformation $\{A_{nm}\}$ für die beiden Seiten von (22) an. Wegen (21) ist

$$\begin{aligned} A_{nm}\{f; x, y\} = f(\bar{x}, \bar{y}) &+ \left((x-\bar{x})\frac{\partial}{\partial x} + (y-\bar{y})\frac{\partial}{\partial y}\right) f(\bar{x}, \bar{y}) + \\ &+ \frac{1}{2}\left((x-\bar{x})\frac{\partial}{\partial x} + (y-\bar{y})\frac{\partial}{\partial y}\right)^2 f(\bar{x} + \vartheta_1(x-\bar{x}), \bar{y} + \vartheta_2(y-\bar{y})) + O\left(\frac{1}{n^2}\right)\frac{\partial}{\partial x} f(\bar{x}, \bar{y}) + \\ &+ O\left(\frac{1}{m^2}\right)\frac{\partial}{\partial y} f(\bar{x}, \bar{y}) + \frac{1}{2}\left(O\left(\frac{1}{n}\right)\frac{\partial}{\partial x} + O\left(\frac{1}{m}\right)\frac{\partial}{\partial y}\right)^2 f(\bar{x} + \vartheta_1(x-\bar{x}), \bar{y} + \vartheta_2(y-\bar{y})). \end{aligned}$$

Hieraus ist für jedes $(\bar{x}, \bar{y}) \in D$

$$|A_{nm}\{f(\bar{x}, \bar{y})\} - f(\bar{x}, \bar{y})| \leqq O\left(\frac{1}{n^2}\right)\left\|\frac{\partial}{\partial x} f(x, y)\right\| + O\left(\frac{1}{m^2}\right)\left\|\frac{\partial}{\partial y} f(x, y)\right\| +$$

(24)

$$+ O\left(\frac{1}{n^2}\right)\left\|\frac{\partial^2}{\partial x^2} f(x, y)\right\| + O\left(\frac{1}{nm}\right)\left\|\frac{\partial^2}{\partial x\, \partial y} f(x, y)\right\| + O\left(\frac{1}{n^2}\right)\left\|\frac{\partial^2}{\partial y^2} f(x, y)\right\|.$$

Da die Funktionen $f_x(x, y)$ und $f_y(x, y)$ in D nach x und y differenzierbare Funktionen sind und in dem Punkt (ξ, η) verschwinden, gelten die Relationen

$$f_x(x, y) = \int_{\xi}^{x} f_{xx}(t, \eta)\, dt + \int_{\eta}^{y} f_{xy}(x, u)\, du$$

$$f_y(x, y) = \int_{\eta}^{y} (\xi, t)\, dt + \int_{\xi}^{y} f_{xy}(u, y)\, du.$$

Es ergibt sich

$$\| f_x(x, y)\| \leqq 2\left(\|f_{xx}(x, y)\| + \|f_{xy}(x, y)\|\right) \tag{25}$$

resp.

$$\| f_y(x, y)\| \leqq 2\left(\|f_{yy}(x, y)\| + \|f_{xy}(x, y)\|\right). \tag{26}$$

Nach der Einsetzung von (26) und (25) in (24) folgt, daß die Bedingung (16) befriedigt ist.

Es seien $f(x, y) \in C_{2\pi}$, $D = \{-\pi \leqq x \leqq \pi,\ -\pi \leqq y \leqq \pi\}$ und $\{T_{nm}\}$ eine Folge linearer Transformationen, die den Raum $C_{2\pi}$ in den Raum der trigonometrischen Polynome höchstens n resp. m-ter Ordnung transformiert. Aus (17) können wir die folgende Erweiterung des Satzes von G. FREUD [3] beweisen.

SATZ 4. *Die positive lineare Transformationsfolge $\{T_{nm}\}$ stellt eine Zygmundsche Approximationsfolge dann und nur dann dar, falls die Bedingungen*

$$T_{nm}\{1\} = 1$$

$$T_{nm}\{\sin(x-u)\} = O\left(\frac{1}{n^2}\right)$$

$$T_{nm}\{\sin(y-v)\} = O\left(\frac{1}{m^2}\right)$$

(27)

$$T_{nm}\left\{\sin^2\frac{x-u}{2}\right\} = O\left(\frac{1}{n^2}\right)$$

$$T_{nm}\left\{\sin^2\frac{y-v}{2}\right\} = O\left(\frac{1}{m^2}\right)$$

$$T_{nm}\left\{\sin\frac{x-u}{2}\sin\frac{y-v}{2}\right\} = O\left(\frac{1}{nm}\right)$$

erfüllt sind.

BEWEIS. Es ist hinreichend zu zeigen, daß für $f(x, y) \subset C^2_{2\pi}$ aus (27) die Relation (16) folgt. Wir betrachten die Funktion $g(x, y) = f(x, y) - f(u, v) - [f_x(u, v) \sin(x-u) + f_y(u, v) \sin(y-v)]$. Da $g(u, v) = g_x(u, v) = g_y(u, v) = 0$ ist, können wir $g(x, y)$ aus den partiellen Ableitungen zweiter Ordnung folgenderweise ausdrücken:

$$g(x, y) = \int_u^x \int_u^x g_{xx}(\xi, y)\, d\xi\, d\xi + 2 \int_u^x \int_v^x g_{xy}(\xi, \eta)\, d\xi\, d\eta + \int_v^y \int_v^y g_{yy}(x, \eta)\, d\eta\, d\eta$$

Aus der Definition von $g(x, y)$ ist es offenbar, daß die Abschätzungen

$$|g_{xx}| \leqq \|f_{xx}\| + \|f_x\|, \quad |g_{yy}| \leqq \|f_{yy}\| + \|f_y\|, \quad |g_{xy}| \leqq \|f_{xy}\|$$

gelten, woraus

$$-\left\{(\|f_{xx}\| + \|f_x\|)\frac{(x-u)^2}{2} + (\|f_{yy}\| + \|f_y\|)\frac{(y-v)^2}{2} + \|f_{xy}\|\big((x-u)+(y-v)\big)^2\right\} \leqq$$

$$\leqq g(x, y) \leqq (\|f_{xx}\| + \|f_x\|)\frac{(x-u)^2}{2} + (\|f_{yy}\| + \|f_y\|)\frac{(y-v)^2}{2} + \tag{28}$$

$$+ \|f_{xy}\|\big((x-u)+(y-v)\big)^2$$

folgt. Setzen wird die Relation $t^2 = \pi^2 \sin^2 \frac{t}{2}$ in (28) ein, und wenden wir die lineare Transformation $\{T_{nm}\}$ für (28) an. Wir bekommen

$$-\pi_2\{(\|f_{xx}\| + \|f_x\|)\}\, T_{nm}\left\{\sin^2\frac{x-u}{2}; x\right\} + (\|f_{yy}\| + \|f_y\|)\, T_{nm}\left\{\sin^2\frac{y-v}{2}; y\right\} +$$

$$+ \|f_{xy}\|\, T_{nm}\left\{\sin\frac{x-u}{2} + \sin\frac{y-v}{2}; x, y\right\} \leqq T_{nm}\{f; x, y\} - f(u, v)\, T_{nm}\{1; x, y\} -$$

$$-[f_x\, T_{nm}\{\sin(x-u); x\} + f_y\, T_{nm}\{\sin y; y-v\}] \leqq \pi^2\left[(\|f_{xx}\| + \|f_x\|)\, T_{nm}\left\{\sin^2\frac{x-u}{2}; x\right\} +\right.$$

$$\left. + (\|f_{yy}\| + \|f_y\|)\, T_{nm}\left\{\sin^2\frac{y-v}{2}; y\right\} + \|f_{xy}\|\, T_{nm}\left\{\left[\sin\frac{x-u}{2} + \sin\frac{y-v}{2}\right]^2; x, y\right\}\right].$$

Nach den Bedingungen (27) und den Ungleichungen (25) und (26) folgt die Abschätzung (16).

6.

Die Funktion $f(x, y)$ gehört in dem Gebiet D zu der Zygmundschen Klasse, d.h. $f(x, y) \in Z$, falls für jedes $(x, y) \in D$ und für beliebige Richtung α $0 \leqq \alpha \leqq 2\pi$

$$\sup_{\substack{\alpha \\ (x,y)\in D}} |f(x + \varrho\cos\alpha, y + \varrho\sin\alpha) + f(x - \varrho\cos\alpha, y - \varrho\sin\alpha) -$$

$$-2f(x, y)| \leqq M\varrho \qquad (\varrho = \text{Konst.}) \tag{29}$$

ist. Offenbar ist die Definition (29) mit der Relation

$$\omega_2(f;\delta) \leqq M\delta \tag{30}$$

identisch.

BEMERKUNGEN.

1) Aus (2) ist es klar, daß eine Funktion $f(x, y)$ dann und nur dann $f(x, y) \in Z$ ist, falls in jeder Geraden E_α in der Richtung α in gewöhnlichem Sinne zu der Zygmundsche Klasse gehört.

2) Es ist leicht ersichtlich, daß für die Funktionen $f(x, y) \in \mathrm{Lip}\,\alpha$, $f(x, y) \in Z$.

3) Es sei $f(x, y) \in C_{2\pi}$ und bezeichnen wir die beste Approximation von $f(x, y)$ durch trigonometrische Polynomen mit der Ordnung n resp. m mit $E_{nm}(f)$. Aus dem Satz 1. können wir ohne Schwierigkeit beweisen, daß die Funktion $f(x, y)$ dann und nur dann zu der Zygmundschen Klasse gehört, falls

$$E_{nm}(f) \leqq \text{Konst.} \sqrt{\frac{1}{n^2}+\frac{1}{m^2}}$$

ist. Dieser Satz ist eine Erweiterung in zwei Dimensionen des Satzes von A. ZYGMUND [7].

LITERATURVERZEICHNIS

[1] BURKILL, H.: Cesaro-Perron almost periodic functions, *Proc. London Math. Soc.* **2** (1952) 157.
[2] FREUD, G.: Sui procedimenti lineari d'approssimazione, *Atti Accad. Naz. Lincei. Rend. Cl. Sci. Fis. Mat. Nat.* **26** (1959) 641—643.
[3] FREUD, G.: Über eine positive Zygmundsche Approximationsfolge, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961) 71—75.
[4] *On Approximation Theory*, JSNM vol. 5 Birkhäuser Verlag, Basel—Stuttgart, 1964, S. 183—183.
[5] КОРОВКИН, П. П.: *Линейные операторы и теория приближений*, Физматгиз, Москва, 1959.
[6] NATANSZON, I. P.: *Konstruktív függvénytan*, Akadémiai Kiadó, Budapest, 1952, p. 87—89.
[7] ZYGMUND, A.: Smooth Functions, *Duke Math. J.* **12** (1945) 47—76.

MATHEMATISCHES INSTITUT DER UNGARISCHEN AKADEMIE
DER WISSENSCHAFTEN, BUDAPEST

(Eingegangen: 5. September, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 215—219.

ON RATIONAL APPROXIMATION

by

G. FREUD and J. SZABADOS

1. In his paper [1] one of us proved some theorems concerning the rational approximation. Two of these theorems were generalized in [2]. In this paper we will give further generalizations of the theorems in question towards an other direction.

Let us denote by Θ_n the class of piecewise constant functions defined in [0, 1] having less than n jumps (we need not care about the values of a $\vartheta_n(x) \in \Theta_n$ at the points of jump). Let us further denote by the best approximation of a bounded function $f(x)$ $(0 \leqq x \leqq 1)$ the quantity

$$E(f; n) = \inf_{\vartheta_n \in \Theta_n} \sup_{x \in [0,1]} |f(x) - \vartheta_n(x)|$$

THEOREM. *Let $r \geqq 1$ and assume that the $(r-1)^{th}$ derivative of a function $f(x)$ exists in* [0, 1] *and $f^{(r-1)}(x)$ is a primitive function of an integrable function $f^{[r]}(x)$. Then there exist rational functions $r_N(x)$ of degree N for which*

$$|f(x) - r_N(x)| = O\left(\frac{E\left(f^{[r]}; \left[\frac{N}{\log^2 N^{9[r+1]}}\right]\right)}{N^r} + \frac{1}{N^{r+1}} \right) \qquad (0 \leqq x \leqq 1)$$

further

$$|r_N'(x)| = O(N^{28r+29} \log^3 N) \qquad (0 \leqq x \leqq 1).$$

PROOF. We make use of the same method as in [1]. Consider

$$\varphi(x) = f(x) - \sum_{k=0}^{r-1} \frac{f^{[k]}(0)}{k!} x^k - \frac{f^{[r]}(+0)}{r!} x^r \tag{1.1}$$

then $\varphi^{[r]}(x) = f^{[r]}(x) - f^{[r]}(+0)$ and $E(f^{[r]}; n) = E(\varphi^{[r]}; n)$. Let

$$s_N = \left[\frac{N}{\log^2 N^{9(r+1)}}\right]. \tag{1.2}$$

Consider the function $\Phi_{s_N}^{[r]}(x) \in \Theta_{s_N}$ approximating $\varphi^{[r]}(x)$ with an error smaller than $2E(\varphi^{[r]}; s_N)$. Let $0 = \xi_0 < \xi_1 < \ldots < \xi_{t_N} = 1$ $(t_N \leqq 2s_N)$ a sequence containing all the jumps of $\Phi_{s_N}^{[r]}(x)$ and the points k/s_N $(k = 0, 1, \ldots, s_N)$. In this way we have

$$\xi_{i+1} - \xi_i \leqq \frac{1}{s_N} \qquad (i = 0, 1, \ldots, t_N - 1)$$

and

$$|\varphi^{[r]}(x)-\varphi^{[r]}(y)|\leqq 4E(f^{[r]};\, s_N) \quad \text{for} \quad x,\, y\in J_i=(\xi_i,\, \xi_{i+1}) \qquad (i=0,\, 1,\, \ldots,\, t_N-1).$$

Now we divide the J_i's into two classes. Let

$$\gamma_N=\frac{1}{N^{r+1}} \tag{1.3}$$

and denote by J_i' that intervals for which

$$\xi_{i+1}-\xi_i\geqq\gamma_N$$

and by J_i'' for which

$$\xi_{i+1}-\xi_i<\gamma_N.$$

For the J_i''s we get polynomials $p_{v_N}^{\{i\}}(x)$ of degree

$$v_N=2\left[\frac{8}{65}\log^2 N^{9(r+1)}\right] \tag{1.4}$$

for which (see [1])

$$|\varphi(x)-p_{v_N}^{\{i\}}(x)|=O(1)\frac{E(f^{[r]};\, s_N)}{v_N^r s_N^r}=O(1)\frac{E(f^{[r]};\, s_N)}{N^r} \qquad (x\in\bar{J}_i')$$

and

$$p_{v_N}^{\{i\}}(\xi_i)=\varphi(\xi_i), \qquad p_{v_N}^{\{i\}}(\xi_{i+1})=\varphi(\xi_{i+1}).$$

Further let

$$X^{\{i\}}=\frac{2}{\xi_{i+1}-\xi_i}\left(x-\frac{\xi_{i+1}+\xi_i}{2}\right)$$

and T_{v_N} the Čebyšev-polynomial of degree v_N; finally

$$r_{v_N}^{\{i\}}(x)=\frac{(1+\gamma_N)p_{v_N}^{\{i\}}(x)}{1+\gamma_N\cdot T_{v_N}(X^{\{i\}})}$$

a rational function of degree v_N. We obtain in the same way as in [1]

$$r_{v_N}^{\{i\}}(\xi_i)=\varphi(\xi_i), \qquad r_{v_N}^{\{i\}}(\xi_{i+1})=\varphi(\xi_{i+1}), \tag{1.5}$$

$$|\varphi(x)-r_{v_N}^{\{i\}}(x)|=O(1)\frac{E(f^{[r]};\, s_N)}{N^r} \qquad (x\in\bar{J}_i'), \tag{1.6}$$

$$|r_{v_N}^{\{i\}'}(x)|=O(1)\,v_N^2\gamma_N^{-3} \qquad (x\in[0,\, 1]). \tag{1.7}$$

For the J_i'''s let $r_{v_N}^{\{i\}}(x)$ be that linear function for which (1.5) holds. Then we have (see [1])

$$|\varphi(x)-r_{v_N}^{\{i\}}(x)|=O(1)\cdot\gamma_N \qquad (x\in\bar{J}_i'') \tag{1.8}$$

$$|r_{v_N}^{\{i\}'}(x)|=O(1) \qquad (x\in[0,\, 1]). \tag{1.9}$$

Now let

$$\varphi_N(x)=\frac{r_{v_N}^{\{1\}}(x)+r_{v_N}^{\{t_N\}}(x)}{2}+\frac{1}{2}\sum_{i=2}^{t_N}|x-\xi_i|\,\frac{r_{v_N}^{\{i\}}(x)-r_{v_N}^{\{i-1\}}(x)}{x-\xi_i}. \tag{1.10}$$

Evidently

$$\varphi_N(x) = r_{v_N}^{\{i\}}(x) \quad \text{for} \quad x \in \bar{J}_i$$

and hence we obtain by (1. 6) and (1. 8)

$$|\varphi(x) - \varphi_N(x)| = O(1) \max\left[\frac{E(f^{[r]}; s_N)}{N^r}, \gamma_N\right] = \tag{1.11}$$

$$= O(1)\left[\frac{E(f^{[r]}; s_N)}{N^r} + \gamma_N\right] \qquad (0 \leqq x \leqq 1).$$

Let $R_{v_N}(x-\xi_i)$ denote Newman's rational function of degree v_N approximating $|x-\xi_i|$, then the rational function

$$r_N(x) = \sum_{k=0}^{r-1} \frac{f^{(k)}(0)}{k!} x^k + \frac{f^{[r]}(+0)}{r!} x^r + \frac{r_{v_N}^{\{1\}}(x) + r_{v_N}^{\{t_N\}}(x)}{2} + \tag{1.12}$$

$$+ \frac{1}{2} \sum_{i=2}^{t_N} R_{v_N}(x-\xi_i) \frac{r_{v_N}^{\{i\}}(x) - r_{v_N}^{\{i-1\}}(x)}{x-\xi_i}$$

is of degree (see (1. 2) and (1. 4))

$$r + 2v_N + (t_N - 1)2v_N = r + 2t_N v_N \leqq r + 4s_N v_N \leqq r + \frac{64}{65} N \leqq N$$

at most, provided that $N \geqq 65r$. Further, taking into account (1. 1), (1. 12), (1. 11), (1. 10), (1. 7), (1. 9) and (1. 3) we have

$$|f(x) - r_N(x)| = \left|\varphi(x) - \frac{r_{v_N}^{\{1\}}(x) + r_{v_N}^{\{t_N\}}(x)}{2} - \frac{1}{2} \sum_{i=2}^{t_N} R_{v_N}(x-\xi_i) \frac{r_{v_N}^{\{i\}}(x) - r_{v_N}^{\{i-1\}}(x)}{x-\xi_i}\right| \leqq$$

$$\leqq O(1)\left[\frac{E(f^{[r]}; s_N)}{N^r} + \gamma_N\right] + \frac{3e^{-\sqrt{v_N}}}{2} \sum_{i=2}^{t_N} \left[\left|\frac{r_{v_N}^{\{i\}}(x) - r_{v_N}^{\{i\}}(\xi_i)}{x-\xi_i}\right| + \right.$$

$$\left. + \left|\frac{r_{v_N}^{\{i-1\}}(x) - r_{v_N}^{\{i-1\}}(\xi_i)}{x-\xi_i}\right|\right] \leqq O(1)\left[\frac{E(f^{[r]}; s_N)}{N^r} + \gamma_N + e^{-\sqrt{v_N}} s_N v_N^2 \gamma_N^{-3}\right] =$$

$$= O(1)\left[\frac{E(f^{[r]}; s_N)}{N^r} + \gamma_N + \gamma_N^{\frac{36}{\sqrt{65}}} v_N^2 \gamma_N^{-3}\right] = O(1)\left[\frac{E(f^{[r]}; s_N)}{N^r} + \gamma_N\right] \qquad (0 \leqq x \leqq 1)$$

which proves the first part of the theorem. As regards the second statement of the theorem, we have from (1. 2)—(1. 4), (1. 7) and (1. 9)

$$|r_{v_N}^{\{i\}}(x)| = O(\gamma_N^{-1}) = O(N^{r+1}) \qquad (x \text{ arbitrary})$$

$$|r_{v_N}^{\{i\}'}(x)| = O(v_N^2 \gamma_N^{-3}) = O(N^{3(r+1)} \log^4 N) \qquad (x \text{ arbitrary}).$$

Apply the Lemma from [2] according to our notations:

$$\left|\frac{R_{v_N}(x-\xi_i)}{x-\xi_i}\right| = O(1), \quad \left|\left(\frac{R_{v_N}(x-\xi_i)}{x-\xi_i}\right)'\right| = O\left(v_N^{5/2} e^{6\sqrt{v_N}}\right) \qquad \left(|x| \leqq 1 + \frac{1}{v_N}\right).$$

From these relations we obtain

$$|r'_N(x)| \leqq O\left[\frac{N}{\log^2 N}(N^{27(r+1)}\log^5 N \cdot N^{r+1} + N^{3(r+1)}\log^2 N)\right] =$$
$$= O(N^{28r+29}\log^3 N)$$

thus our theorem is proved. The last estimation for $r'_N(x)$ gives a possibility for using Theorem 1A from [2].

2. A corollary and an example. *Let especially $f^{[r]}(x)$ be of bounded variation with total variation $V(f^{[r]})$. Then there exist rational functions $r_N(x)$ of degree N such that*

$$|f(x)-r_N(x)| = O(1)V(f^{[r]})\frac{\log^2 N}{N^{r+1}} \qquad (x\in[0,1]).$$

This corollary is Satz 2 from [1]. We can obtain it from our Theorem by observing that if $f^{[r]}(x)$ is of bounded variation then

$$E(f^{[r]};n) = O\left(\frac{1}{n}\right). \tag{2.1}$$

To prove (2. 1) let $f^{[r]}(x)=g_1(x)-g_2(x)$, where $g_1(x)$ and $g_2(x)$ are both monotonically increasing. Let $m=\left[\frac{n}{2}\right]$ and

$$\zeta_m^{(\mu)} = \inf\left\{x: g_1(x) \geqq \mu\frac{g_1(1)-g_1(0)}{m}+g_1(0)\right\} \qquad (\mu = 0, 1, \ldots, m).$$

Consider the jump function

$$\varphi_m(x) = \mu\frac{g_1(1)-g_1(0)}{m}+g_1(0) \quad \text{for} \quad x\in(\zeta_m^{(\mu)}, \zeta_m^{(\mu+1)}) \qquad (\mu = 0, 1, \ldots, m-1)$$

with m jump. Clearly

$$|g_1(x)-\varphi_m(x)| \leqq \frac{g_1(1)-g_1(0)}{m} \qquad (x\in[0,1]).$$

Analogously, one can find a function $\psi_m(x)\in\Theta_m$ such that

$$|g_2(x)-\psi_m(x)| \leqq \frac{g_2(1)-g_2(0)}{m} \qquad (x\in[0,1]).$$

But in this case for the function $\Phi_n(x)=\varphi_m(x)-\psi_m(x)\in\Theta_n$ we have

$$|f^{[r]}(x)-\Phi_n(x)| = O\left(\frac{1}{m}\right) = O\left(\frac{1}{n}\right)$$

i.e. (2. 1) holds.

AN EXAMPLE. *There exist rational functions $r_N(x)$ of degree N such that*

$$\left|\int_0^x t^\alpha\left|\sin\frac{1}{t^\alpha}\right|dt - r_N(x)\right| = O\left(\frac{\log^3 N}{N^2}\right) \qquad (x\in[0,1], 0<\alpha<1). \tag{2.2}$$

Namely, we may apply our Theorem with

$$f(x) = \int_0^x t^\alpha \left| \sin \frac{1}{t^\alpha} \right| dt, \qquad r = 1.$$

In this case an easy calculation shows that

$$E(f'; n) = E\left(x^\alpha \left| \sin \frac{1}{x^\alpha} \right|; n\right) = O\left(\frac{\log n}{n}\right). \tag{2. 3}$$

To prove (2. 3) let $x_k = \left(\frac{2}{k\pi}\right)^{\frac{1}{\alpha}}$ then for even k's $f'(x_k) = 0$, for odd k's $f'(x_k) = \frac{2}{k\pi}$ is a local maximum. We obtain a function $\vartheta_n(x) \in \Theta_n$ approximating $f'(x)$ with an error ε as follows. Let

$$\vartheta_n(x) = \varepsilon \quad \text{if} \quad x \leqq x_{k_0}, \quad \text{where} \quad k_0 = 2\left[\frac{1}{2\pi\varepsilon}\right] + 2.$$

As regards the intervals (x_{k+1}, x_k) $(k \leqq k_0)$, we can approximate $f'(x)$ by a jump function which has at most

$$\left[\frac{1}{k\pi\varepsilon}\right] + 1$$

jump points. Thus the total number of jump points is at most

$$1 + \sum_{k=1}^{k_0}{}' \left(\left[\frac{1}{k\pi\varepsilon}\right] + 1\right) \leqq 3 + \frac{1}{\pi\varepsilon} + \frac{\log k_0}{\pi\varepsilon} \leqq \frac{|\log \varepsilon|}{\varepsilon}.$$

Correspondingly, if we put $\varepsilon = \frac{\log n}{n}$ then we can see that (2. 3) holds.

Thus the Theorem gives (2. 2). Clearly $f'(x) \notin \operatorname{Lip} \beta$ if $\beta > \alpha$ so the magnitude of the polynomial approximation is not better than $O(N^{-1-\alpha})$ which is worse than (2. 2). The Corollary or Satz 2 from [1] cannot be applied because $f'(x)$ is not of bounded variation.

REFERENCES

[1] FREUD, G.: Über die Approximation reeller Funktionen durch rationale gebrochene Funktionen, *Acta Math. Acad. Sci. Hungar.* **17** (1966) 313—324.

[2] SZABADOS, J.: Generalization of two theorems of G. Freud concerning the rational approximation, *Studia Sci. Math. Hung.* **2** (1967) 73—80.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST
and
EÖTVÖS L. UNIVERSITY, BUDAPEST

(Received September 19, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 221—226.

ÜBER STARKE APPROXIMATION DURCH DIFFERENZIERTE FOLGEN VON APPROXIMIERENDEN POLYNOMEN

von

G. FREUD

Es sei $f(x)$ eine im endlichen Intervall $[c, d]$ stetig differenzierbare Funktion, $\pi_{\nu n}(x)$ $(\nu = 1, 2, \ldots, n;\ n = 1, 2, \ldots)$ eine zeilenfinite Matrix von Polynomen, welche $f(x)$ in $[c, d]$ im folgenden Sinne stark mit Exponenten p approximieren:

$$\left\{\frac{1}{n}\sum_{\nu=1}^{n}|f(x)-\pi_{\nu n}(x)|^p\right\}^{1/p} \leqq \varepsilon_n, \qquad \varepsilon_n \to 0 \tag{1}$$

$$\text{für } x \in [c, d], \qquad (n = 1, 2, \ldots).$$

Es sei weiter vorausgesetzt, daß die Polynome $\pi_{\nu n}(x)$ $(\nu = 1, 2, \ldots, n)$ höchstens vom Grade $2n$ sind. Gefragt wird nach einer Abschätzung des Ausdruckes

$$h_n^{(1)}(p; f; \pi_{\nu n}; x) = \left\{\frac{1}{n}\sum_{\nu=1}^{n}|f'(x)-\pi'_{\nu n}(x)|^p\right\}^{1/p}. \tag{2}$$

Wir zeigen, daß — oberflächlich ausgedrückt — falls die Polynome $\pi_{\nu n}$ selbst die Funktion $f(x)$ im starken Sinne hinreichend gut approximieren, dann approximieren auch die differenzierten Ausdrücke $\pi'_{\nu n}$ im starken Sinne, mit dem gleichen Exponenten p und mit einer passenden Genauigkeit die Funktion $f'(x)$. Als Anwendung seien Abschätzungen der starken de la Vallée Poussinschen Summen und der Starken $(C, 1)$ Summen von Orthogonalpolynomreihen bewiesen.

Wir bezeichnen — wie üblich — mit

$$E_n(f; c, d) = \operatorname*{Min}_{\pi_n} \operatorname*{Max}_{x \in [c,d]} |f(x)-\pi_n(x)| \tag{3}$$

falls $\pi_n(x)$ alle Polynome höchstens n-ten Grades durchläuft.

SATZ 1. *Aus* (1) *folgt für jedes* $0 < \delta < \dfrac{d-c}{2}$

$$|h_n^{(1)}(p; f; \pi_{\nu n}; x)| \leqq K[n\varepsilon_n + E_{n-1}(f'; c, d)] \tag{4}$$

$$\text{für} \quad x \in [c+\delta, d-\delta]$$

wo K eine sowohl von n, wie von der Wahl der Polynom-matrix $|\pi_{\nu n}|$ und der Funktion f unabhängige positive Größe ist. [1]

[1] Das gleiche sei über die später auftretenden Größen K_1, K_2, vorausgesetzt.

BEWEIS: Im Falle $p>1$ sei $p^{-1}+q^{-1}=1$ und wir bezeichnen mit $R_n^{(p)}$ die Menge der n-gliedrigen reellen Folgen $\{\varrho_\nu\}$ mit

$$\sum_{\nu=1}^{n} |\varrho_\nu|^q \leqq 1.$$

Im Falle $p=1$ sei $R_n^{(1)}$ die Menge der n-gliedrigen Folgen $\{\varrho_\nu\}$ mit $|\varrho_\nu| \leqq 1$ $(\nu=1, 2, \ldots, n)$. Infolge der Hölderschen Ungleichung gilt für $p>1$

$$\sum_{\nu=1}^{n} |\varrho_\nu| \leqq n^{1/p} \quad \text{für} \quad \{\varrho_\nu\} \in R_n^{(p)} \tag{5}$$

und diese Formel behält ihre Gültigkeit für $p=1$.

Nach einem wohlbekannten Satze der Analysis ist

$$\left\{\sum_{\nu=1}^{n} |a_\nu|^p\right\}^{1/p} = \operatorname*{Max}_{\{\varrho_\nu\} \in R_n^{(p)}} \left|\sum_{\nu=1}^{n} \varrho_\nu a_\nu\right|, \tag{6}$$

diese Formel bildet die Grundlage des Beweises.

Aus (1) und (6) ergibt sich für jedes $\{\varrho_\nu\} \in R_n^{(p)}$

$$\left|\sum_{\nu=1}^{n} \varrho_\nu [f(x)-\pi_{\nu n}(x)]\right| \leqq n^{1/p} \varepsilon_n. \tag{7}$$

Es sei $\tau_n(f; x)$ das — nach einem bekannten Satze von E. BOREL vorhandene — Polynom höchstens n-ten Grades, für welches

$$|f(x)-\tau_n(f; x)| \leqq E_n(f; c, d) \quad \text{für} \quad x \in [c, d] \tag{8}$$

gültig ist. Wegen (5), (7) und (8) ist

$$\left|\sum_{\nu=1}^{n} \varrho_\nu [\tau_n(f; x)-\pi_{\nu n}(x)]\right| \leqq n^{1/p} [\varepsilon_n + E_n(f; c, d)] \quad \text{für} \quad x \in [c, d].$$

Unter dem Betragszeichen steht ein Polynom höchstens n-ten Grades; durch Anwendung der Bernsteinschen Ungleichung haben wir

$$\left|\sum_{\nu=1}^{n} \varrho_\nu [\tau_n'(f; x)-\pi_{\nu n}'(x)]\right| \leqq K_1 n^{1+1/p} [\varepsilon_n + E_n(f; c, d)] \tag{9}$$

$$\text{für} \quad x \in [c+\delta, d-\delta].$$

Laut eines früheren Ergebnisses des Verfassers (G. FREUD [5]; vgl. auch J. CZIPSZER und G. FREUD [4]) folgt aus (8)

$$|f'(x)-\tau_n'(f; x)| \leqq K_2 [n E_n(f; c, d) + E_{n-1}(f'; c, d)] \quad \text{für} \quad x \in [c+\delta, d-\delta]. \tag{10}$$

Aus (9) und (10) ergibt sich unter Beachtung der bekannten Ungleichung [2]

$$E_n(f; c, d) \leqq K_3 \frac{E_{n-1}(f'; c, d)}{n}$$

daß

$$\left| \sum_{\nu=1}^{n} \varrho_\nu [f'(x) - \pi'_{\nu n}(x)] \right| \leqq K_4 n^{1/p} [n\varepsilon_n + E_n(f'; c, d)] \tag{11}$$

$$\text{für } x \in [c+\delta, d-\delta]$$

gültig ist.

Da (11) für alle Folgen $\{\varrho_\nu\} \in R_n^{(p)}$ gültig ist, ergibt sich unter Beachtung der Beziehung (6)

$$\left\{ \sum_{\nu=1}^{n} |f'(x) - \pi'_{\nu n}(x)|^p \right\}^{1/p} \leqq K_4 n^{1/p} [n\varepsilon_n + E_n(f'; c, d)]$$

$$\text{für } x \in [c+\delta, d-\delta]$$

und das zeigt die Gültigkeit von (4) (vgl. (2)) w.z.b.w.

Durch wiederholte Anwendung des Satzes 1 ergibt sich

SATZ 2.: *Es sei $f(x)$ in $[c, d]$ r-mal stetig differenzierbar, dann folgt aus* (1) *für jedes* $0 < \delta < \frac{d-c}{2}$ *und* $n > r$

$$h_n^{(r)}(p; f; \pi_{\nu n}; x) = \left\{ \frac{1}{n} \sum_{\nu=1}^{n} |f^{(r)}(x) - \pi_{\nu n}^{(r)}(x)|^p \right\}^{1/p} \leqq$$

$$\leqq K^{(r)} [n(n-1)\ldots(n-r+1)\varepsilon_n + E_{n-r}(f^{(r)}; c, d)] \tag{12}$$

wo $K^{(r)}$ sowohl von n, wie auch von der Wahl der Polynommatrix $|\pi_{\nu n}|$ und der Funktion f unabhängig ist.

Einige wichtige Anwendungen dieser Sätze ergeben sich aus den Untersuchungen von G. ALEXITS und D. KRÁLIK [2], [3], L. LEINDLER [6] und endlich G. ALEXITS [1] bezüglich der starken $(C, 1)$-Summierung von Orthogonalpolynomreihen. Mit Hilfe der in den zitierten Arbeiten entwickelten Methode ergeben sich folgende Sätze:

SATZ A: Es seien $w(x)$ eine in einem endlichen Intervalle $[a, b]$ definierte nichtnegative L-integrierbare Gewichtsfunktion, $\{p_n(w; x)\}$ die Folge der normierten

[2] *Beweis:* Es sei $\psi_{n-1}(x)$ das Polynom höchstens $n-1$-ten Grades mit $|f'(x) - \psi_{n-1}(x)| \leqq \leqq E_{n-1}(f'; c, d)$ für $x \in [c, d]$, ferner sei $\Psi_n(x)$ eine primitive Funktion von $\psi_{n-1}(x)$. Infolge $|(f - \Psi_n)'| \leqq E_{n-1}(f')$ gilt für das Jacksonsche Approximationspolynom $J_n(x) = J_n(f; \Psi_n; x)$ die Ungleichung $|f(x) - \Psi_n(x) - J_n(x)| \leqq 10(d-c)\frac{E_{n-1}(f'; c, d)}{n}$, es ist also $E_n(f; c, d) \leqq$ $\leqq 10(d-c)\frac{E_{n-1}(f'; c, d)}{n}$ w. z. b. w.

Orthogonalpolynome bezüglich dieser Gewichtsfunktion. Es sei vorausgesetzt, daß $w(x)$ in einem echten Teilintervall $[c, d] \subset (a, b)$ beschränkt ist und

$$\sum_{\nu=0}^{n-1} p_\nu^2(w; x) = O(n) \tag{13}$$

gleichmäßig bezüglich $x \in [c, d]$ befriedigt ist. Es sei ferner $f(x)$ eine in $[a, b]$ stetige Funktion, und $s_\nu(w; f; x)$ sei die ν-te Teilsumme der Orthogonalentwicklung von $f(x)$ gemäß $\{p_\nu(w; x)\}$; dann gilt

$$\frac{1}{n} \sum_{\nu=n+1}^{2n} |f(x) - s_\nu(w; f; x)| \leqq c_1 E_n(f; a, b) \tag{14}$$

mit einer positiven Zahl C_1, welche weder von n, noch von f abhängt.[3]

SATZ B: Wir machen die gleichen Voraussetzungen wie bei Satz A, es sei aber an Stelle von (13) sogar

$$p_\nu(w; x) = O(1) \tag{15}$$

bezüglich $x \in [c, d]$ gleichmäßig erfüllt; dann gilt sogar für jedes $p \geqq 1$

$$\left\{\frac{1}{n} \sum_{\nu=n+1}^{2n} |f(x) - s_\nu(w; f; x)|^p\right\}^{1/p} \leqq C(p) E_n(f; a, b). \tag{16}$$

Unter Anwendung des Satzes 2 ergeben sich folgende weitere Sätze:

SATZ 3: *Es sei $f(x)$ in $[a, b]$ r-mal stetig differenzierbar, $r \geqq 1$. Dann gilt unter den Voraussetzungen des Satzes A*

$$\frac{1}{n} \sum_{\nu=n+1}^{2n} |f^{(r)}(x) - s_\nu^{(r)}(w; f; x)| \leqq c_2 E_{n-r}(f^{(r)}; a, b) \tag{17}$$

und unter den Voraussetzungen des Satzes B, für jedes $p \geqq 1$

$$\left\{\frac{1}{n} \sum_{\nu=n+1}^{2n} |f^{(r)}(x) - s_\nu^{(r)}(w; f; x)|^p\right\}^{1/p} \leqq c_3(p) E_{n-2}(f^{(r)}; a, b) \tag{18}$$

$$\text{für } x \in [c + \delta, d - \delta].$$

BEWEIS: Wir setzen $\pi_{\nu n}(x) = s_{n+\nu}(w; f; x)$, und schließen aus Satz 2 und Satz A bzw. Satz 2 und Satz B, daß

$$\left\{\frac{1}{n} \sum_{\nu=n+1}^{2n} |f^{(r)}(x) - s_\nu^{(r)}(w; f; x)|^p\right\}^{1/p} \leqq c_4(p)[n(n-1)\ldots(n-r+1) E_n(f; a, b) +$$

$$+ E_{n-r}(f; c, d)]$$

$$\text{für} \quad x \in [c + \delta, d - \delta]$$

[3] Das gleiche sei für die mit $C(p)$, c_2 u. s. w. bezeichneten Konstanten gültig.

im Falle $p=1$, bzw. im Falle $p\geqq 1$ gültig ist. Die Behauptungen des Satzes folgen nun aus den Ungleichungen

$$E_{n-r}(f^{(r)};\, c,\, d)\leqq E_{n-r}(f^{(r)};\, a,\, b)$$

und (für $n>r$)

$$n(n-1)\ldots(n-r+1)E_n(f;\, a,\, b)\leqq c_5 E_{n-r}(f^{(r)};\, a,\, b)$$

(vgl. Fußnote 2).

Zwecks Übergang zur Abschätzung von starken $(C, 1)$-Summen und verwandten Mittelbildungen bedienen wir uns folgenden (ziemlich trivialen) Hilfssatzes, welcher auch unter anderen ähnlichen Verhältnissen von Nutzen sein kann:

HILFSSATZ: *Es sei* $\{A_\nu\}$ *eine beliebige Folge positiver Zahlen, und* $\{\varepsilon_\nu\}$ *eine abnehmende Folge positiver Zahlen, s eine beliebige reelle Zahl,* $r>0$ *eine ganze Zahl; dann folgt aus*

$$\frac{1}{n}\sum_{\nu=n+1}^{2n} A_\nu \leqq \varepsilon_n \qquad (n=r, r+1, \ldots), \tag{19}$$

daß für jedes $n>r$

$$\sum_{\nu=r+1}^{n} \nu^{s-1}A_\nu \leqq \gamma_s r^s \varepsilon_r + 2\gamma_s^2 \sum_{\nu=r+1}^{n} \nu^{s-1}\varepsilon_\nu \tag{20}$$

mit

$$\gamma_s = \operatorname{Max}(1, 2^{s-1})$$

befriedigt ist.

BEWEIS: (Durch elementares Rechnen) Für $m\geqq 1$ ergibt sich unter Beachtung von (19) und der Monotonie von $\{\varepsilon_\nu\}$

$$\sum_{\nu=r2^m+1}^{r2^{m+1}} \nu^{s-1}A_\nu \leqq \gamma_s (r2^m)^{s-1} \sum_{r2^m+1}^{r2^{m+1}} A_\nu \leqq \gamma_s (r2^m)^s \varepsilon_{r2^m} \leqq$$

$$\leqq 2\gamma_s (r2^m)^{s-1} \sum_{\nu=r2^{m-1}+1}^{r2^m} \varepsilon_\nu \leqq 2\gamma_s^2 \sum_{\nu=r2^{m-1}+1}^{r2^m} \nu^{s-1}\varepsilon_\nu .$$

Wir addieren diese Ungleichungen für $m=1, 2, \ldots, N-1$ miteinander und mit der — sich aus (19) ergebenden — Ungleichung

$$\sum_{\nu=r+1}^{2r} \nu^{s-1}A_\nu \leqq \gamma_s r^{s-1} \sum_{\nu=r+1}^{2r} A_\nu \leqq \gamma_s r^s \varepsilon_r$$

und erhalten

$$\sum_{\nu=r+1}^{r2^N} \nu^{s-1}A_\nu \leqq \gamma_s r^s \varepsilon_r + 2\gamma_s^2 \sum_{\nu=r+1}^{r2^{N-1}} \nu^{s-1}\varepsilon_\nu \qquad (N=1, 2, \ldots)^4. \tag{21}$$

[4] Für $N=1$ setzen wir die letzte Summe laut Verabredung gleich Null.

Es sei nun $n>r$, $2^{N-1}\leqq n<2^N r$; dann folgt aus (21)

$$\sum_{\nu=r+1}^{n} \nu^{s-1}A_\nu \leqq \sum_{\nu=r+1}^{r2^N} \nu^{s-1}A_\nu \leqq \gamma_s r^s \varepsilon_r + 2\gamma_s^2 \sum_{\nu=r+1}^{r2^{N-1}} \nu^{s-1}\varepsilon_\nu \leqq \gamma_s r^s \varepsilon_r +$$

$$+ 2\gamma_s^2 \sum_{\nu=r+1}^{n} \nu^{s-1}\varepsilon_\nu$$

w.z.b.w.

Aus dem Hilfssatz, angewandt auf $A_\nu = f^{(r)}(x) - s_\nu^{(r)}(w; f; x)$ und $\varepsilon_n^p = c_3(p)E_{n-r}(f^{(r)}; a, b)$ erhalten wir

SATZ 4. *Aus* (18) *folgt*

$$\sum_{\nu=r+1}^{n} \nu^{s-1}|f^{(r)}(x) - s_\nu^{(r)}(w; f; x)|^p \leqq$$

(22)

$$\leqq c_6(p, s)\left\{r^s[E_0(f^{(r)}; a, b)]^p + \sum_{\nu=r+1}^{n} \nu^{s-1}[E_{\nu-r}(f^{(r)}; a, b)]^p\right\}$$

$$\text{für} \quad x\in[c+\delta, d-\delta] \qquad (n=r+1, r+2, \ldots).$$

Ist z.B. $f^{(r)}(x)$ m-mal stetig differenzierbar, $f^{(r+m)}(x)\in \mathrm{Lip}\,\alpha$ $(0<\alpha\leqq 1)$ und $(m+\alpha)p<s$, dann folgt aus (22)

$$\left\{\frac{1}{n^s}\sum_{\nu=r+1}^{n} \nu^{s-1}|f^{(r)}(x) - s_\nu^{(r)}(w; f; x)|^p\right\}^{1/p} = O(n^{-m-\beta})$$

gleichmäßig bezüglich $x\in[c+\delta, d-\delta]$.

LITERATURVERZEICHNIS

[1] ALEXITS, G.: Einige Beiträge zur Approximationstheorie, *Acta Sci. Math.* (Szeged) **26** (1965) 211—224.

[2] ALEXITS, G. und KRÁLIK, D.: Über die Approximation im starken Sinne, *Acta Sci. Math.* **26** (1965) 93—101.

[3] ALEXITS, G. und KRÁLIK, D.: Über die Approximation mit starken de la Vallée Poussinschen Mitteln, *Acta Math. Acad. Sci. Hungar.* **16** (1965) 43—50.

[4] CZIPSZER, J. et FREUD, G.: Sur l'approximation d'une fonction periodique et ses dérivées successives par un polynome trigonometrique et par ses derivées successives, *Acta Math.* **99** (1958) 33—51.

[5] FREUD, G.: Über die $(C, 1)$-Summen der Entwickelungen nach orthogonalen Polynomen, *Acta Math. Acad. Sci. Hungar.* **14** (1963) 197—208.

[6] LEINDLER, L.: Über die Approximation im starken Sinne, *Acta Math. Acad. Sci. Hungar.* **16** (1965) 255—262.

MATHEMATISCHES INSTITUT DER UNGARISCHEN AKADEMIE DER WISSENSCHAFTEN, BUDAPEST

(Eingegangen: 27. September, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 227—247.

ON THE GENERAL BRANCHING PROCESS WITH CONTINUOUS TIME-PARAMETER

by

D. SZÁSZ

§ 0. Introduction

There are various types of applications, in which the following model can be set up: in some space (for example on the real line or in the three-dimensional euclidian space) particles are distributed at random. The particles may have different life-times, depending on the chance, and then they divide. The new particles are spread out in the space at random. This phenomenon may be repeated. This process of random distributions of particles in the space is the spreading process, or clustering process, or with the expression used in the theory of branching processes — the general branching process with continuous time-parameter ([1] III. 17.). In fact our process is an age-dependent branching process, but the points are elements of some space, and we describe not only the entire number of points, but also their situation in the space. The necessity of such an investigation arises in the stochastic foundation of cosmology (J. NEYMAN [2]), in the theory of epidemics, in biology (for example in the description of propagation of plants), in physics (in the description of a nuclear reaction), etc.

This paper is the starting step of the systematic description of the general branching process with continuous time-parameter. For the case, when the life-times are not random variables, but they have unit length, the process was studied by A. PRÉKOPA ([3]), and some of our theorems generalizes his results. J. A. MOYAL ([4]) investigated the multiplicative population process, which corresponds to our homogeneous spreading process, but his approach and results differ from ours.

§ 1 contains the concept of random point distribution in an abstract space, some definitions, and a theorem of A. PRÉKOPA, which will be used later. In § 2 we study the single spreading, because the results obtained for it are very useful for the investigation of the spreading process. In § 3 we consider the spreading process and in § 4 we give the expectations of the random variables characterizing the spreading process. An interesting type of the spreading process is the homogeneous spreading process. This is considered in § 5. § 6 contains an interesting property of the single spreading and an application of this property for the spreading process.

Some of the results contained in this paper is the content of the thesis of the author ([5]).

At last I should like to express my sincere thanks to A. PRÉKOPA and A. RÉNYI for their valuable remarks.

§ 1. Random Point Distributions in Abstract Spaces

Let T be an abstract space, and σ_T such a σ-algebra of subsets of T, which contains the countable subsets of T.

Let Ω contain the non-negative, integer-valued (allowing the value $+\infty$ too) functions $\omega(t)$ defined on T and having values different from 0 only on a countable subset of T. In what follows Ω will be called the sample space of a random point distribution. Every ω will be called a realization of the random point distribution. If $\omega(t) \geqq 1$, we say that t is a point of the realization ω ($t \in \omega$), and if $\omega(t) \geqq 2$, then t is a multiple point of the realization ω. The realization ω is determined by the set of its points with their multiplicities.

In case of $A \in \sigma_T$ we define the function $\xi(\omega, A) = \xi(A)$ as follows

$$\xi(\omega, A) = \sum_{t \in A} \omega(t)$$

(the value of $\xi(\omega, A)$ may also be $+\infty$).

We form a σ-algebra containing subsets of Ω so that it should contain the sets

$$\{\omega : \omega \in \Omega,\ \xi(\omega, A_1) = k_1, \ldots, \xi(\omega, A_n) = k_n\} \tag{1. 1}$$

where $A_1, \ldots, A_n$ are sets from σ_T and $k_1, \ldots, k_n$ are non-negative integers. Let σ_0 be the system of sets of type (1. 1), and σ_Ω the smallest σ-algebra containing σ_0. This definition guarantees that in case of any $A_1, \ldots, A_n$ ($A_i \in \sigma_T$, $i = 1, \ldots, n$) $(\xi(A_1), \ldots, \xi(A_n))$ is a random vector variable. We suppose further that on σ_Ω a probability measure $\mathbf{P}$ is defined ($\mathbf{P}(\Omega) = 1$).

If Ω, σ_Ω and $\mathbf{P}$ are given in this manner corresponding to the space T, we say that in T a *random point distribution* is given. In this paper we suppose that for each $t \in T$

$$\mathbf{P}(\xi(\{t\}) < \infty) = 1.$$

Obviously $\mathbf{M}\xi(A)$ is a measure on σ_T. We call $\mathbf{M}\xi(A)$ the *expectation measure* of the random point distribution ξ.

The probabilities of the sets of σ_0 determine the probability distributions of the random vector variables $(\xi(A_1), \ldots, \xi(A_n))$ $(A_i \in \sigma_T)$, and these distributions determine the measure $\mathbf{P}$ on σ_Ω; thus $\mathbf{P}$ is uniquely determined by its values taken on elements of σ_0.

If the sets $A_1, \ldots, A_n$ in (1. 1) are not disjoint, then one can find disjoint sets $B_1, \ldots, B_r$ in σ_T such that for every A_k ($k = 1, \ldots, r$)

$$A_k = \sum_{l=1}^{l_k} B_{k_l}.$$

If we suppose that random variables corresponding to disjoint sets are independent, then the measure $\mathbf{P}$ is uniquely determined by the distributions of the random variables $\xi(A)$ ($A \in \sigma_T$), namely these distributions determine the probabilities of the sets (1. 1).

DEFINITION 1. 1. We say that a random point distribution is of *Poisson type*, if random variables corresponding to disjoint sets of σ_T are independent, and there is a σ-finite measure μ on σ_T so that if $A \in \sigma_T$ and $\mu(A) < \infty$, then

$$\mathbf{P}(\xi(A) = k) = \frac{\mu^k(A)}{k!} e^{-\mu(A)} \qquad (k = 0, 1, \ldots)$$

The measure μ mentioned in the above definition will be called the *parameter-measure* of the Poisson random point distribution. If $\mu(T) < \infty$, then the random point distribution will be called *finite Poisson type*.

DEFINITION 1. 2. We say that the random point distribution ξ is *atomless*, if for every $t \in T$

$$\mathbf{P}(\xi(\{t\}) = 0) = 1.$$

A Poisson random point distribution is clearly atomless if and only if for its parameter-measure $\mu(\{t\}) = 0$ $(t \in T)$.

In what follows we shall state the product space theorem proved by A. PRÉKOPA ([6]), which will be a useful tool in characterizing Poisson spreading processes. We suppose that in the space T a random point distribution is given. This point distribution will be called the *underlying* point distribution. Let us suppose that every point of the underlying point distribution is the starting point of a random happening taking place in the abstract space Y. These happenings are symbolized by elements of the space Y. If $\omega = (t_1, t_2, \ldots)$ $(t_k \in T)$ is a realization of the underlying random point distribution, then the whole phenomenon is described by the set of pairs of points $((t_1, y_1), (t_2, y_2), \ldots)$, where $y_k \in Y$ $(k = 1, 2, \ldots)$. So the sample space Ω_1 of the whole phenomenon consists of point distributions defined in the space $Z = T \times Y$. We suppose that a σ-algebra σ_Y of subsets of Y is given and denote by σ_Z the σ-algebra $\sigma_T \times \sigma_Y$ defined in Z.

As we said, the sample space Ω_1 of the whole phenomenon contains the point distributions $((t_1, y_1), (t_2, y_2), \ldots)$ for which $(t_1, t_2, \ldots) \in \Omega$, and $y_1, y_2 \ldots$ are arbitrary elements of Y. Naturally the elements of Ω_1 can be identified with non-negative, integer-valued functions $\omega_1(z)$ defined on Z and having values different from 0 only on a countable subset of Z. In case of $D \in \sigma_Z$ we define the function $\eta(\omega_1, D) = \eta(D)$ as follows

$$\eta(\omega_1, D) = \sum_{z \in D} \omega_1(z).$$

σ_{Ω_1} will be the smallest σ-algebra for which the functions $\eta(\omega_1, D)$ are measurable for every $D \in \sigma_Z$. It will not lead to misunderstanding if we denote the probability measure defined on σ_{Ω_1} also by $\mathbf{P}$ $(\mathbf{P}(\Omega_1) = \mathbf{1})$. We remark that the σ-algebra σ_Ω is isomorphic to a system of subsets of σ_{Ω_1} so we need not separate the underlying point distribution from the whole phenomenon. Thus we can denote the isomorphic σ-algebra also by σ_Ω, and we can keep the notations ω, $\xi(A)$ too.

Provisionally we suppose that

$$\mathbf{P}(\xi(T) < \infty) = \mathbf{1}$$

i.e. the realizations of the underlying distribution contain finite number of points with probability 1.

Let us consider the following conditional probabilities

$$\mathbf{P}\big(\eta(D_1)=k_1, \ldots, \eta(D_n)=k_n|\sigma_\Omega\big)$$

$$(D_i=A_i\times C_i,\ A_i\in\sigma_T,\ C_i\in\sigma_Y,\ i=1, \ldots, n),$$

where $k_1, \ldots, k_n$ are non-negative integers. Because of the isomorphism the conditional probability can be written in the following form

$$\mathbf{P}\big(\eta(D_1)=k_1, \ldots, \eta(D_n)=k_n|t_1, \ldots, t_r\big)$$

where $\omega=(t_1, \ldots, t_r)\in\Omega$.

We assume that the underlying points generate the secondary happenings independently of each other, and we formulate this assumption as follows

α) If $D_i=A_i\times C_i$, where $A_i\in\sigma_T$, $C_i\in\sigma_Y$, and $t_i\in A_i$ $(i=1, \ldots, n)$, $A_iA_k=0$ (if $i\neq k$), then

$$\mathbf{P}\big(\eta(D_1)=1, \ldots, \eta(D_n)=1|t_1, \ldots, t_n\big)=\varepsilon(C_1, t_1)\ldots\varepsilon(C_n, t_n) \tag{1.2}$$

where for every fixed t $\varepsilon(C, t)$ is a probability measure on the σ-algebra σ_Y.

This measure $\varepsilon(C, t)$ is the probability distribution of the secondary happening, if its starting point is t. But in case of a fixed $C\in\sigma_Y$ $\varepsilon(C, t)$ is also a measurable function of the variable t with respect to the σ-algebra σ_T. To prove this let us consider the subset $\Omega^{(1)}$ of the sample space Ω, the elements of which consist of a single point $t\in T$. From (1.2)

$$\mathbf{P}\big(\eta(D)=1|t\big)=\varepsilon(C, t)$$

where $D=T\times C$, $C\in\sigma_Y$, $t\in T$. From the definition of conditional probability it follows that a set $\{\omega\}=\{(t)\}\subset\Omega^{(1)}$ is measurable with respect to σ_Ω if and only if the set $\{t\}\subset T$ is measurable with respect to σ_Ω, so

$$\{\omega\colon \omega\in\Omega^{(1)},\ a\leqq\big(\eta(D)=1|\omega\big)\leqq b\}\in\sigma_\Omega$$

and thus

$$\{t\colon t\in T,\ a\leqq\varepsilon(C, t)\leqq b\}\in\sigma_T$$

what means the measurability of $\varepsilon(C, t)$ with respect to σ_T.*

If for the underlying distribution $\mathbf{P}\big(\xi(T)<\infty\big)<1$, then with positive probability it happens that ω has infinite number of points, so in (1.2) we have an infinite product. To exclude this we say that the phenomenon satisfies the condition α), if for every $B\in\sigma_T$, for which $\mathbf{P}\big(\xi(B)<\infty\big)=1$, the condition α) is satisfied for the set B instead of the set T.

THEOREM 1.1. *If in T an atomless Poisson random point distribution ξ is given with parameter measure μ and for the whole phenomenon in $T\times Y$ the condition α) holds, then the random point distribution η of the whole phenomenon is also of atomless Poisson type and if $D=A\times C$ $(A\in\sigma_T,\ C\in\sigma_Y)$, then*

$$\mathbf{M}\eta(D)=\int_A\varepsilon(C, t)\mu(dt). \tag{1.3}$$

* The proof above is taken over from [6].

If D' is an arbitrary element of σ_Z, then

$$\mathbf{M}\eta(D') = \nu^*(D')$$

where ν^* is the extension to σ_Z of the measure $\mathbf{M}\eta(D)$ defined for the rectangular sets $D = A \times C$ $(A \in \sigma_T,\ C \in \sigma_Y)$.

The proof of the theorem can be found in [6].

2. § Single Spreading

We suppose that in T a random point distribution ξ is given. In case of the spreading process the secondary happenings are again point distributions, so the space Y consists of the point distributions in T, that is it is identical with Ω defined in § 1. Similarly the σ-algebra σ_Y is identical with σ_Ω. In case of $A \in \sigma_T$, $y \in Y$ let $\psi(y, A) = \psi(A)$ be defined by

$$\psi(y, A) = \sum_{t \in A} y(t).$$

If $\omega = (t_1, t_2, \ldots)$ is a realization of the underlying random point distribution, then to every $t_i \in \omega$ we choose a random element y_i of Y, i. e. every t_i generates a random point distribution in T.

By the superposition of these point distributions we obtain a new point distribution ζ in T, that is

$$\zeta(\omega_1, A) = \sum_{l=1}^{\infty} \psi(y_l, A)$$

where $\omega_1 = ((t_1, y_1), (t_2, y_2), \ldots)$. This phenomenon will be called *single spreading*. We remark that here we supposed that the underlying points die, but if they do not, they can be taken into account in the y's, and so our model contains this case too.

DEFINITION 2. 1. We say that a single spreading is *independent*, if it satisfies Condition α) of § 1.

To determine the expectation of $\zeta(A)$, we use the equality

$$\mathbf{M}\zeta(A) = \mathbf{M}[\mathbf{M}(\zeta(A)|\omega)]$$

where $\mathbf{M}(\zeta(A)|\omega) = \mathbf{M}(\zeta(A)|\sigma_\Omega)$. Supposing that $\omega = (t_1, t_2, \ldots)$

$$\mathbf{M}(\zeta(A)|\omega) = \mathbf{M}\left[\sum_{l=1}^{\infty} \psi(y_l, A)|t_1, t_2, \ldots\right] = \sum_{l=1}^{\infty} \mathbf{M}(\psi(y_l, A)|t_1, t_2, \ldots).$$

If the spreading is independent, then

$$\mathbf{M}(\psi(y_l, A)|t_1, t_2, \ldots) = M(A, t_l) \tag{2. 1}$$

where for fixed $t \in T$ $M(A, t)$ is a measure on σ_T, and for fixed $A \in \sigma_T$, $M(A, t)$ is a measurable function of t with respect to σ_T. This latter statement follows from

the measurability of $\varepsilon(C, t)$ for a fixed C. Using the concept and the properties of the random integral defined in [1] (III. 4 and III. 9. 1) we have

$$\mathbf{M}(\zeta(A)|\omega) = \int_T M(A, t)\omega(dt)$$

and thus

$$\mathbf{M}\zeta(A) = \mathbf{M}\left(\int_T M(A, t)\omega(dt)\right) = \int_T M(A, t)\mu(dt) \tag{2. 2}$$

where μ is the expectation measure of the underlying distribution. Thus we have

THEOREM 2. 1. *If in case of an independent single spreading μ is the expectation measure of the underlying distribution, $M(A, t)$ is the expectation of the number of points spread by the underlying point t to the set A (the exact definition of $M(A, t)$ is* (2. 1)*), then the expectation of $\zeta(A)$ can be expressed in the form* (2. 2).

This theorem was known on more special conditions and is proved for example in [1].

DEFINITION 2. 1. We say that a probability distribution is of compound Poisson type, if its characteristic function is

$$\exp\left\{\sum_{k=1}^{\infty} c_k(e^{iuk} - 1)\right\}$$

where the c_k's are non-negative real numbers, and $\sum_{k=1}^{\infty} c_k$ is convergent.

It is clear that the distribution of the random variable ζ is of compound Poisson type if ζ is the sum of a series

$$\eta_1 + 2\eta_2 + 3\eta_3 + \ldots$$

where $\eta_1, \eta_2, \ldots$ are independent random variables having Poisson distributions and the series converges with probability 1.

Let us consider an independent single spreading where the underlying random point distribution is of atomless Poisson type with expectation measure μ and suppose that for the set A $\mathbf{P}(\zeta(A) < \infty) = 1$. Let $C_{A,k}$ be that subset of Y, for the points of which

$$\psi(y, A) = k \qquad (k = 0, 1, 2, \ldots)$$

and introduce the notation

$$D_{A,k} = T \times C_{A,k} .$$

From the definition of σ_Y it follows that $C_{A,k} \in \sigma_Y$, and so $D_{A,k} \in \sigma_Z = \sigma_T \times \sigma_Y$. Obviously

$$\zeta(A) = \sum_{k=1}^{\infty} k\eta(D_{A,k}) .$$

Using our suppositions

$$\mathbf{P}(\eta(D_{A,k}) < \infty) = 1 .$$

In case of $i \neq j$, $D_{A,i} D_{A,k} = \emptyset$, so in consequence of Theorem 1. 1 the random

variables $\eta(D_{A,1}), \eta(D_{A,2}), \ldots$ are independent with Poisson distributions, and the expectation of $\eta(D_{A,k})$ is

$$\mathbf{M}\eta(D_{A,k}) = \int_T \varepsilon(C_{A,k}, t)\mu(dt).$$

Let us denote the random variable $\psi(y, A)$ by $\psi_t(A)$, if y is selected from Y to the point t. In this case

$$\varepsilon(C_{A,k}, t) = \mathbf{P}(\psi_t(A) = k)$$

and so

$$\mathbf{M}\eta(D_{A,k}) = \int_T \mathbf{P}(\psi_t(A) = k)\mu(dt) \tag{2. 3}$$

But we supposed that $\mathbf{P}(\zeta(A) < \infty) = 1$ and this means that $\zeta(A)$ has a compound Poisson distribution with the characteristic function

$$f(u, A) = \mathbf{M}(e^{iu\zeta(A)}) = \exp\left\{\sum_{k=1}^{\infty} \int_T \mathbf{P}(\psi_t(A) = k)\mu(dt)\,(e^{iuk} - 1)\right\}$$

and the series

$$\sum_{k=1}^{\infty} \int_T \mathbf{P}(\psi_t(A) = k)\mu(dt)$$

converges. Thus

$$\sum_{k=1}^{\infty} \int_T |\mathbf{P}(\psi_t(A) = k)(e^{iuk} - 1)|\,\mu(dt) \leqq 2 \sum_{k=1}^{\infty} \int_T \mathbf{P}(\psi_t(A) = k)\mu(dt)$$

so we can invert the order of summation and integration. Accordingly

$$f(u, A) = \exp\left\{\int_T \mathbf{P}(\psi_t(A) = k)(e^{iuk} - 1)\mu(dt)\right\} =$$

$$= \exp\left\{\int_T \sum_{k=0}^{\infty} \mathbf{P}(\psi_t(A) = k)(e^{iuk} - 1)\mu(dt)\right\} =$$

$$= \exp\left\{\int_T \left(\sum_{k=0}^{\infty} \mathbf{P}(\psi_t(A) = k)e^{iuk} - 1\right)\mu(dt)\right\}$$

Denoting the characteristic function of $\psi_t(A)$ by $g(u, A, t)$ we have

THEOREM 2. 2. *If a single spreading is independent, the underlying distribution is of atomless Poisson type, and* $\mathbf{P}(\zeta(A) < \infty) = 1$, *then the distribution of* $\zeta(A)$ *is of compound Poisson type with the characteristic function*

$$f(u, A) = \exp \int_T (g(u, A, t) - 1)\mu(dt).$$

THEOREM 2. 3. *If the conditions of the preceding theorem hold, then*

$$\mathbf{D}^2\zeta(A) = \int_T \mathbf{M}\psi_t^2(A)\mu(dt)$$

provided the integral on the right hand side is finite.

PROOF. It is well-known that the variance of a random variable with compound Poisson distribution is

$$\sum_{k=0}^{\infty} k^2 c_k$$

Using (2. 3)

$$\mathbf{D}^2\zeta(A) = \sum_{k=1}^{\infty} k^2 \int_T \mathbf{P}(\psi_t(A) = k)\mu(dt) = \int_T \sum_{k=1}^{\infty} k^2 \mathbf{P}(\psi_t(A) = k)\mu(dt) =$$

$$= \int_T \mathbf{M}\psi_t^2(A)\mu(dt)$$

Q. e. d.

§ 3. The Spreading Process

We shall denote in what follows the underlying distribution by ξ and $\mathbf{M}\xi(A)$ by $\mu(A)$. In case of the single spreading each underlying point had the same life-time, and after this time the points died and generated new random point distributions independently of each other in the sense of Condition α), and the result of the spreading was the superposition of these point distributions.

In case of the spreading process the life-times are also random variables and the process contains not only one step, but the generated points also have random life-times and after it they generate random point distributions again, and so on. As the spreading process consists of single spreadings, we can use here the results obtained for the single spreading. We shall use the notations of the preceding section. (As we have said, in the case of the spreading process Y(analogously y, σ_Y) is Ω(resp. ω, σ_Ω). In the preceding section we used the notation Y to make clear the difference between the underlying distribution and the generated distributions, but from now on we use only the notation Ω.)

First we shall give the sample space of the spreading process. A realization ϑ of the spreading process is given by the history of the process. The history contains the underlying point distribution $\omega = (t_1, t_2, \ldots)$, the life-times of its points: $s_1, s_2, \ldots$ (s_{i_1} is the life-time of the point t_{i_1}), the point distributions generated by the underlying points: $\omega_1 = (t_{11}, t_{12}, \ldots)$, $\omega_2 = (t_{21}, t_{22}, \ldots)$, ... ($\omega_{i_1}$ is generated by t_{i_1}), the life-times of these points: $s_{11}, s_{12}, \ldots; s_{21}, s_{22}, \ldots; \ldots$, the point distributions generated after these life-times: $\omega_{11} = (t_{111}, t_{112}, \ldots)$, $\omega_{12} = (t_{121}, t_{122}, \ldots)$, ...; $\omega_{21} = (t_{211}, t_{212}, \ldots)$, $\omega_{22} = (t_{221}, t_{222}, \ldots)$, ...; ... and so on. The point $t_{i_1 \ldots i_n}$ has a life-time $s_{i_1 \ldots i_n}$ and thereafter it generates the point distribution $\omega_{i_1 \ldots i_n} = (t_{i_1 \ldots i_n 1}, t_{i_1 \ldots i_n 2}, \ldots)$. So the points ϑ of the sample space Θ of the spreading process have the form

$$\vartheta = (\omega; s_1, \omega_1; s_2, \omega_2; \ldots; s_{11}, \omega_{11}; s_{12}, \omega_{12}; \ldots)$$

where $\omega; \omega_1, \omega_2, \ldots; \omega_{11}, \omega_{12}, \ldots; \ldots$ are elements of Ω, $s_1, s_2, \ldots; s_{11}, s_{12}, \ldots; \ldots$ are non-negative real numbers, $t_1, t_2, \ldots; t_{11}, t_{12}, \ldots; \ldots$ are elements of T. (We remark that if $\omega_{i_1 \ldots i_n}$ has only finite number of points, for example $h_{i_1 \ldots i_n}$, where

$$h_{i_1 \ldots i_n} = \sum_{t \in T} \omega_{i_1 \ldots i_n}(t) \tag{3.1}$$

then points $t_{i_1 \ldots i_n j}$ for which $j > h_{i_1 \ldots i_n}$, do not exist in ϑ.)

Naturally the ϑ's, which differ only in the order of the t's in certain ω's, are considered as identical.

The point $t_{i_1 \dots i_m}$ is an *ancestor* of the point $t_{j_1 \dots j_n}$, if $n \geqq m$, and $i_1 = j_1, \dots, i_m = j_m$. In this case $t_{j_1 \dots j_n}$ is a *descendant* of $t_{i_1 \dots i_m}$. We say that the point $t_{i_1 \dots i_m}$ belongs to the *generation* m. We remark that in ϑ the t's are determined by the ω's. We make the following suppositions and we shall use them without mentioning:

1° The s's are independent of each other and of the ω's, with a common distribution function $U(s)$ $(U(+0)=0)$,

2° The point $t_{i_1 \dots i_m}$ determines the probability distribution of $\omega_{i_1 \dots i_m}$ in σ_Ω, that is if $C \in \sigma_\Omega$ then

$$\mathbf{P}(\omega_{i_1 \dots i_m} \in C) = \varepsilon(C, t_{i_1 \dots i_m})$$

where for fixed $t \in T$, $\varepsilon(C, t)$ is a probability measure on σ_Ω, and for fixed $C \in \sigma_\Omega$, $\varepsilon(C, t)$ is a measurable function of t with respect to the σ-algebra σ_T (See 1. 2).

If the probability measure of the underlying distribution ξ, the distribution function $U(s)$, and the measures $\varepsilon(C, t)$ are given, then the probability measure of the spreading process is defined.

The spreading process is a process of random point distributions ζ_s depending on the time parameter s $(0 \leqq s < \infty)$. The point distribution ζ_s consists of points $t_{i_1 \dots i_m}$ occurring in the realization ϑ, for which

$$s_{i_1} + s_{i_1 i_2} + \dots + s_{i_1 i_2 \dots i_{m-1}} < s \leqq s_{i_1} + s_{i_1 i_2} + \dots + s_{i_1 i_2 \dots i_{m-1} i_m}. \tag{3. 2}$$

So we can define $\zeta_s(\vartheta, A) = \zeta_s(A)$ for $A \in \sigma_T$ as the number of those $t_{i_1 \dots i_m}$, which occur in ϑ, and for which (3. 2) and

$$t_{i_1 \dots i_m} \in A$$

are satisfied. In other words, $\zeta_s(A)$ is the number of those points, which are living in the time s and fall to the set A.

We suppose that for the function $M(A, t)$ defined for the single spreading in § 2 ((2. 1))

$$M(T, t) \leqq Q \tag{3. 3}$$

independently of t $(\in T)$, that is the expected number of generated points has a finite upper bound not depending on the generating point. This supposition will be kept in the whole paper. We can define $M(A, t)$ in another manner too:

$$M(A, t) = \sum_{k=0}^{\infty} k\varepsilon(C_{A,k}, t)$$

where

$$C_{A,k} = \{\omega : \omega \in \Omega, \sum_{t \in A} \omega(t) = k\}.$$

Let us denote by α_s the number of those points which were born before the time s, i. e. the number of those $t_{i_1 \dots i_m}$ for which

$$s_{i_1} + s_{i_1 i_2} + \dots + s_{i_1 i_2 \dots i_{m-1}} < s. \tag{3. 4}$$

THEOREM 3. 1. *If* $\mu(T) < \infty$, *and* (3. 2) *holds, then* $\mathbf{M}\alpha_s < \infty$.

PROOF. For each m-tuple $i_1, \dots, i_m$ of positive integers let $\varphi_{i_1 \dots i_m}$ be 1, if for the index $i_1 \dots i_m$ there is a $t_{i_1 \dots i_m}$ in ϑ, or with other words if

$$i_1 \leqq h, \; i_2 \leqq h_{i_1}, \dots, i_m \leqq h_{i_1 \dots i_{m-1}};$$

otherwise we put $\varphi_{i_1 \dots i_m} = 0$ (about the definition of $h_{i_1 \dots i_m}$ see (3. 1)). Let $\chi_{i_1 \dots i_m}$ be 1, if (3. 4) is fulfilled, and 0 otherwise. Obviously

$$\alpha_s = \sum_{m=1}^{\infty} \sum_{i_1=1}^{\infty} \dots \sum_{i_m=1}^{\infty} \varphi_{i_1 \dots i_m} \chi_{i_1 \dots i_m}$$

and because of the independence of the $\varphi's$ from the $\chi's$

$$\mathbf{M}\alpha_s = \sum_{m=1}^{\infty} \sum_{i_1=1}^{\infty} \dots \sum_{i_m=1}^{\infty} \mathbf{M}(\varphi_{i_1 \dots i_m}) \mathbf{M}(\chi_{i_1 \dots i_m}).$$

But $\mathbf{M}(\chi_{i_1 \dots i_m}) = U^{(m)}(s)$, where $U^{(m)}$ is the m-th convolution power of U, and

$$\sum_{i_1=1}^{\infty} \dots \sum_{i_m=1}^{\infty} \mathbf{M}(\varphi_{i_1 \dots i_m}) = \sum_{i_1=1}^{\infty} \dots \sum_{i_m=1}^{\infty} \mathbf{P}(h \geqq i_1, h_{i_1} \geqq i_2, \dots, h_{i_1 \dots i_{m-1}} \geqq i_m) =$$

$$= \sum_{i_1=1}^{\infty} \dots \sum_{i_{m-1}=1}^{\infty} \sum_{i_m=1}^{\infty} \sum_{j=i_m}^{\infty} \mathbf{P}(h \geqq i_1, h_{i_1} \geqq i_2, \dots, h_{i_1 \dots i_{m-2}} \geqq i_{m-1}, h_{i_1 \dots i_{m-1}} = j) =$$

$$= \sum_{i_1=1}^{\infty} \dots \sum_{i_{m-1}=1}^{\infty} \sum_{j=1}^{\infty} j \mathbf{P}(h \geqq i_1, h_{i_1} \geqq i_2, \dots, h_{i_1 \dots i_{m-2}} \geqq i_{m-1}, h_{i_1 \dots i_{m-1}} = j) =$$

$$= \sum_{i_1=1}^{\infty} \dots \sum_{i_{m-1}=1}^{\infty} \mathbf{M}(h_{i_1 \dots i_{m-1}} | h \geqq i_1, h_{i_1} \geqq i_2, \dots, h_{i_1 \dots i_{m-2}} \geqq i_{m-1}) \cdot$$

$$\cdot \mathbf{P}(h \geqq i_1, h_{i_1} \geqq i_2, \dots, h_{i_1 \dots i_{m-2}} \geqq i_{m-1}).$$

The distribution of $h_{i_1 \dots i_{m-1}}$ depends only on $t_{i_1 \dots i_{m-1}}$, so in case of $m \geqq 2$ we can use (3. 3)

$$\mathbf{M}(h_{i_1 \dots i_{m-1}} | h \geqq i_1, h_{i_1} \geqq i_2, \dots, h_{i_1 \dots i_{m-2}} \geqq i_{m-1}) \leqq Q$$

consequently

$$\sum_{i_1=1}^{\infty} \dots \sum_{i_m=1}^{\infty} \mathbf{M}\varphi_{i_1 \dots i_m} \leqq Q \sum_{i_1=1}^{\infty} \dots \sum_{i_{m-1}=1}^{\infty} \mathbf{M}\varphi_{i_1 \dots i_{m-1}} \leqq Q^{m-1} \mu(T)$$

therefore

$$\mathbf{M}\alpha_s \leqq \sum_{m=1}^{\infty} U^{(m)}(s) Q^{m-1} \mu(T). \tag{3. 5}$$

From the following lemma it follows that the sum on the right hand side is convergent. The theorem is proved.

LEMMA. *If $\xi_1, \dots, \xi_m, \dots$ are independent, identically distributed random variables with distribution function $U(s)$, and $U(+0) = 0$, then in case of fixed s*

$$\lim_{m \to \infty} \{\mathbf{P}(\xi_1 + \dots + \xi_m < s)\}^{1/m} = 0.$$

PROOF. It is obvious that if $\lambda > 0$, then

$$\mathbf{P}(\xi_1 + \dots + \xi_m < s) = \mathbf{P}(e^{-\lambda(\xi_1 + \dots + \xi_m)} > e^{-\lambda s}).$$

If we introduce the function $q(\lambda)$ by

$$q(\lambda) = \mathbf{M}(e^{-\lambda\xi_k})$$

then $0 < q(\lambda) < 1$, and by the Markov inequality

$$\mathbf{P}(e^{-\lambda(\xi_1+\ldots+\xi_m)} > e^{-\lambda s}) < e^{\lambda s}[q(\lambda)]^m$$

so

$$\overline{\lim_{m\to\infty}} \mathbf{P}(\xi_1 + \ldots + \xi_m < s)^{1/m} \leqq q(\lambda).$$

But

$$q(\lambda) = \int_0^\infty e^{-\lambda s} dU(s) < U\left(\frac{1}{\sqrt{\lambda}}\right) + e^{-\sqrt{\lambda}}$$

thus because of $U(+0) = 0$

$$\lim_{\lambda\to\infty} q(\lambda) = 0$$

and this proves the lemma.*

If we denote by $\zeta_s^{[t_i]}$ the process of random point distributions consisting only of the descendants of the underlying point t_i, then

$$\zeta_s(A) = \sum_{t_i \in \omega} \zeta_s^{[t_i]}(A). \tag{3.6}$$

From the Theorem 3. 1 it follows that $\zeta_s^{[t_i]}(A)$ exists, and this implies the existence of $\zeta_s(A)$.

If the underlying random point distribution is of atomless Poisson type we have the following

THEOREM 3. 2 *Suppose that for an independent spreading process the underlying random point distribution is of atomless Poisson type, and that for some A and s*

$$\mathbf{P}(\zeta_s(A) < \infty) = 1.$$

Then the distribution of $\zeta_s(A)$ is of compound Poisson type.

PROOF. The random point distribution $\zeta_s^{[t]}$ depends only on the point t_i, so the spreading process can be interpreted as an independent single spreading, therefore Theorem 2. 2 is applicable. Hence the theorem.

§ 4. Expectations

We shall keep the notations and suppositions 1°, 2° and (3. 3) of § 3.

THEOREM 4. 1. *If $\mu(T) < \infty$, then $\mathbf{M}\zeta_s(A)$ exists. Let the functions $M_{s,n}(A, \mu)$ be defined by the following recursion*

$$M_{s,1}(A, \mu) = \mu(A)$$

$$M_{s,n+1}(A, \mu) = \int_T \int_0^s M_{s-r,n}(A, M_t)\, dU(r)\, \mu(dt) + (1 - U(s))\mu(A) \tag{4.1}$$

* This proof, which gives more than my original one, originates from A. RÉNYI.

where for each $t \in T$ M_t *is the measure defined by* $M_t(A) = M(A, t)$ $(A \in \sigma_T)$. *Then the sequence* $M_{s,n}(A, \mu)$ *tends to* $\mathbf{M}\zeta_s(A)$ *for* $n \to \infty$.

PROOF. We modify the spreading process supposing that the points of generation $1, 2, \ldots, n-1$ spread in the manner described above, but the points of generation n do not spread, i. e. they have an infinite life-time. So for any positive integer we get the modified process $\zeta_{s,n}$ $(0 \leqq s < \infty)$, where obviously $\zeta_{s,n}(\vartheta)$ consists of those $t_{i_1 \ldots i_m}$, for which $m < n$, and (3. 2) is fulfilled, or for which $m = n$ and

$$s_{i_1} + s_{i_1 i_2} + \ldots + s_{i_1 i_2 \ldots i_{m-1}} < s$$

holds.

We assert that $M_{s,n}(A, \mu)$ is the expectation of $\zeta_{s,n}(A)$. For $n = 1$ the assertion is obvious. Let us suppose that it is valid for some n. We apply the Theorem 2. 1 for $\mathbf{M}\zeta_{s,n+1}(A)$ therefore we need the expected number of points spread by an underlying point t to the set A, supposing that the $(n+1)-st$ generation does not spread already. The point t has a life-time r. If $0 \leqq r < s$, then in the time r the point t generates a random point distribution in T with the finite expectation measure M_t. Then we have a modified spreading process starting at time r, the expectations of which are known from the inductive assumption. If $r \geqq s$, then we have 1 or 0 point in the set A according to $t \in A$ resp. $t \notin A$. So

$$M_{s,n+1}(A, \mu) = \int_A \left(\int_0^s M_{s-r,n}(A, M_t)\, dU(r) + \big(1 - U(s)\big) \right) \mu(dt) +$$

$$+ \int_{T-A} \int_0^s M_{s-r,n}(A, M_t)\, dU(r)\, \mu(dt) = \int_T \int_0^s M_{s-r,n}(A, M_t)\, dU(r) + \big(1 - U(s)\big)\mu(A)$$

But

$$\mathbf{P}\big(\zeta_{s,n}(A) \to \zeta_s(A)\big) = 1$$

because from Theorem 3. 1 it follows that with probability one there exists an index N, such that

$$\zeta_{s,n}(A) = \zeta_s(A)$$

if $n \geqq N$. According to the definition of α_s

$$\zeta_{s,n}(A) \leqq \alpha_s$$

for each $n \geqq 1$ and $A \in \sigma_T$, but we know that $M\alpha_s < \infty$, so

$$\mathbf{M}\zeta_{s,n}(A) \to \mathbf{M}\zeta_s(A)$$

Q. e. d.

If $\mu(T) = \infty$, then we can determine $\mathbf{M}\zeta_s^{[t]}(A)$ (about $\zeta_s^{[t]}(A)$ see (3. 6)) by the help of the preceeding theorem replacing in it μ by μ_t, where

$$\mu_t(A) = \begin{cases} 1 & \text{if} \quad t \in A \\ 0 & \text{if} \quad t \notin A \end{cases}$$

so applying Theorem 2. 1

$$\mathbf{M}\zeta_s(A) = \int_T \mathbf{M}\zeta_s^{[t]}(A)\, \mu(dt) = \int_T \lim_{n \to \infty} \int_0^s M_{s-r,n}(A, M_t)\, dU(r)\, \mu(dt) + \big(1 - U(s)\big)\mu(A)$$

thus we have the

THEOREM 4. 2

$$\mathbf{M}\zeta_s(A) = \int_T \lim_{n\to\infty} \int_0^s M_{s-r,n}(A, M_t)\, dU(r)\mu(dt) + (1-U(s))\mu(A).$$

The next theorem asserts that under certain conditions the order of integration (with respect to t) and the limit can be inverted, although the finiteness of $\mu(T)$ is not supposed.

THEOREM 4. 3. If $U(s)<1$, and $\mathbf{M}\zeta_s(A)$ exists, then

$$\mathbf{M}\zeta_s(A) = \lim_{n\to\infty} \int_T \int_0^s M_{s-r,n}(A, M_t)\, dU(r)\mu(dt) + (1-U(s))\mu(A). \tag{4. 2}$$

PROOF. Let $\alpha_s(A)$ be defined as the number of those points which are elements of the set A and were born before the time s, i. e. the number of those $t_{i_1\dots i_m}$, for which $t_{i_1\dots i_m}\in A$ and

$$s_{i_1} + s_{i_1 i_2} + \dots + s_{i_1\dots i_{m-1}} < s.$$

We assert that $\mathbf{M}\alpha_s(A)<\infty$. Clearly

$$\mathbf{M}\zeta_s(A) = \sum_{k=0}^{\infty} \mathbf{M}(\zeta_s(A)|\alpha_s(A)=k)\mathbf{P}(\alpha_s(A)=k).$$

The condition $\alpha_s(A)=k$ means that k points were born in A before the time s. For each of them the probability to be alive in the time s is at least $1-U(s)$, so

$$\mathbf{M}(\zeta_s(A)|\alpha_s(A)=k) \geqq k(1-U(s))$$

therefore

$$\mathbf{M}\zeta_s(A) \geqq \sum_{k=0}^{\infty} (1-U(s))k\mathbf{P}(\alpha_s(A)=k) = (1-U(s))\mathbf{M}\alpha_s(A)$$

what means that $\mathbf{M}\alpha_s(A)<\infty$, because $U(s)<1$. The finiteness of $\mathbf{M}\alpha_s(A)$ makes possible to repeat the proof of Theorem 4. 1 for this case too.

COROLLARY. We have seen that to get the assertion of the Theorem 4. 3 we need only to prove the finiteness of $\mathbf{M}\alpha_s(A)$. What is more, if we suppose that $\mathbf{P}(\alpha_s(A)<\infty)=1$, then obviously we have (4. 2). It is also obvious that if we suppose that there exists a positive s^* such that $U(s^*)=0$, then (4. 2) will be also valid.

§ 5. The Homogeneous Spreading Process

In many practical cases the space T is a euclidian space; in this case the spreading has special properties. The most interesting property is the homogeneity. We consider this problem on topological groups. Let T_+ be a locally compact topological group (the index $_+$ marks that we assumed the group property; the other notations remain unchanged), and σ_{T_+} the σ-algebra of its Borel sets.

DEFINITION 5. 1. We say that a single spreading is *homogeneous*, if for the set $C_{A,k}$ and measure $\varepsilon(C, t)$ defined in § 2 we have

$$\varepsilon(C_{A,k}, t) = \varepsilon(C_{A-t,k}, 0)$$

where $A \in \sigma_{T_+}$, $k = 0, 1, 2, \ldots$, $t \in T_+$ and 0 is the unity of T_+ $(A - t = \{\tau : \tau \in T_+, \tau + t \in A\})$. The spreading process is homogeneous, if the single spreadings occurring in it are homogeneous.

The homogeneity of spreading means that the distribution of number of points spread by the point t to the set A is identical with the distribution of number of points spread by the origin to the set $(A - t)$.

DEFINITION 5. 2. We say that a single spreading is *homogeneous in wide sense*, if for the expectation $M(A, t)$ defined in § 2 we have

$$M(A, t) = M(A - t, 0)$$

where $A \in \sigma_{T_+}$, $t \in T_+$. The spreading process is homogeneous in the wide sense, if the single spreadings occurring in the process are homogeneous in the wide sense.

If the process is homogeneous in the wide sense, then the supposition (3. 3) means that

$$M(T_+, 0) < \infty.$$

We shall denote the measure $M(A, 0)$ by $\beta(A)$ $(A \in \sigma_{T_+})$ and generally convolution of distribution functions by $*$, and convolution of measures on T_+ by $\star$. For both types of convolution the n-th power of convolution will be denoted by upper index in brackets. We remark that

$$U^{(0)}(s) = \begin{cases} 0 & s \leqq 0 \\ 1 & s > 0 \end{cases}$$

$$\beta^{(0)}(A) = \begin{cases} 1 & 0 \in A \\ 0 & 0 \notin A \end{cases}$$

and by convention $U^{(-1)}(s) = 0$.

THEOREM 5. 1. *If in T_+ a spreading process is homogeneous in the wide sense, then*

$$\mathbf{M}\zeta_s(A) = \left[\left\{\sum_{n=0}^{\infty} [(1 - U) * U^{(n)}](s)\beta^{(n)}\right\} \star \mu\right](A).$$

PROOF. Let us denote $M_{s,n}(A, \beta)$ by $m_n(A, s)$. We shall prove by induction that

$$m_n(A, s) = \{[1 - U](s)\beta(A) + [(1 - U) * U](s)\beta^{(2)}(A) + \ldots \tag{5. 1}$$

$$\ldots + [(1 - U) * U^{(n-2)}](s)\beta^{(n-1)}(A)\} + U^{(n-1)}(s)\beta^{(n)}(A).$$

For $n = 1$

$$m_1(A, s) = \beta(A)$$

and using (4. 1)

$$m_{n+1}(A, s) = \int_{T_+} \int_0^s m_n(A - t, s - r)\, dU(r)\beta(dt) + (1 - U(s))\beta(A) =$$

$$= [(m_n(.,.) * U) \star \beta](A, s) + (1 - U(s))\beta(A) = [(1 - U) * U](s)\beta^{(n)}(A) +$$

$$+ [(1 - U) * U^{(2)}](s)\beta^{(3)}(A) + \ldots + [(1 - U) * U^{(n-1)}](s)\beta^{(n)}(A) +$$

$$+ U^{(n)}(s)\beta^{(n+1)}(A) + [1 - U](s)\beta(A)$$

and this proves 5. 1. So for fixed A and s

$$\lim_{n\to\infty} m_n(A,s) = \sum_{n=1}^{\infty} [(1-U) * U^{(n-1)}](s)\beta^{(n)}(A)$$

because $\beta^{(n)}(A) \leqq \beta^n(T_+)$, and again by the convergence of the right hand side of (3. 5)

$$U^{(n-1)}(s)\beta^{(n)}(A) \to 0.$$

Using the Theorem 4. 2 we get the assertion of the theorem.

THEOREM 5. 2. *If the spreading process is homogeneous in the wide sense, and for* $B \in \sigma_T$

$$\mu(B) \leqq K\lambda(B)$$

where λ *is the Haar measure, and* K *is a finite positive constant, then for* $\lambda(A) < \alpha$ $\mathbf{M}\zeta_s(A)$ *is finite and*

$$\mathbf{M}\zeta_s(A) = \sum_{n=0}^{\infty} [(1-U) * U^{(n)}](s)[\beta^{(n)} \star \mu](A). \tag{5. 2}$$

PROOF. If the underlying realization is $\omega = (t_1, t_2, \ldots)$, then let us denote by $\alpha_s^{[t_i]}(A)$ the number of those $t_{i_1 \ldots i_m}$'s, for which $i_1 = i$, $t_{i_1 \ldots i_m} \in A$, and

$$s_{i_1} + s_{i_1 i_2} + \ldots + s_{i_1 i_2 \ldots i_{m-1}} < s.$$

Obviously

$$\alpha_s(A) = \sum_{ti \in \omega} \alpha_s^{[t_i]}(A).$$

From Theorem 3. 1 it follows that $\mathbf{M}\alpha_s^{[t]}(T_+) < \infty$. On the basis of homogeneity

$$\mathbf{M}\alpha_s^{[t]}(A) = \mathbf{M}\alpha_s^{[0]}(A-t).$$

Using the Theorem 2. 1

$$\mathbf{M}\alpha_s(A) = \int_{T_+} \mathbf{M}\alpha_s^{[0]}(A-t)\mu(dt) \leqq$$

$$\leqq K \int_{T_+} \mathbf{M}\alpha_s^{[0]}(A-t)\lambda(dt) = K\lambda(A)\mathbf{M}\alpha_s^{[0]}(T_+) < \infty$$

thus by the Corollary of Theorem 4. 3 $\mathbf{M}\zeta_{s,n}(A) \to \mathbf{M}\zeta_s(A)$. The theorem is proved.

COROLLARY. If the life-times have exponential distribution, that is

$$U'(s) = \begin{cases} 0 & s \leqq 0 \\ \tau e^{-\tau s} & s > 0 \end{cases}$$

then

$$[(1-U) * U^{(n)}](s) = \begin{cases} 0 & s \leqq 0 \\ \dfrac{(\tau s)^n}{n!} e^{-\tau s} & s > 0 \end{cases}$$

so

$$\mathbf{M}\zeta_s(A) = \sum_{n=0}^{\infty} \frac{(\tau s)^n}{n!} e^{-\tau s}[\beta^{(n)} \star \mu](A).$$

In what follows, we prove theorems concerning the characteristic functions of the random variables $\zeta_s(A)$.

THEOREM 5. 3. *If for the homogeneous spreading process the underlying random point distribution is of atomless finite Poisson type with parameter measure μ, and the random point distribution generated by the origin is of atomless finite Poisson type with parameter measure β, then the characteristic function $f(u, A, s)$ of $\zeta_s(A)$ can be obtained in the following manner: let us introduce the notations*

$$f_{s,1}(u, A, \mu) = e^{\mu(A)(e^{iu}-1)}$$

$$f_{s,n+1}(u, A, \mu) = e^{\mu(A)(1-U(s))(e^{iu}-1)} \exp \int_{T_+} \int_0^s \left(f_{s-r,n}(u, A, \beta_t) - 1\right) dU(r)\mu(dt)$$

where β_t is the measure for which $\beta_t(A) = \beta(A-t)$. In case of $n \to \infty$ the sequence $f_{s,n}(u, A, \mu)$ tends to $f(u, A, s)$ and the convergence is uniform in every finite u-interval.

PROOF. We prove by induction that $f_{s,n}(u, A, \mu)$ is the characteristic function of $\zeta_{s,n}(A)$. For $n=1$ this assertion is trivial. Supposing that the assertion is valid for some n let us determine the characteristic function of $\zeta_{s,n+1}(A)$. We should like to apply Theorem 2. 2, so we must determine the function $g(u, A, t)$ involved in it. Because of the homogeneity the knowledge of $g(u, A, 0)$ is sufficient. The point 0 has a life time r. If $r < s$, then the origin generates a Poisson random point distribution with parameter measure β. For the spreading process starting at time r we can apply our inductive assumption. If $r \geqq s$, then the characteristic function of the number of points spread by the origin to the set A is

$$h(u, A) = \begin{cases} e^{iu} & 0 \in A \\ 1 & 0 \notin A \end{cases}$$

So

$$g(u, A, 0) = \int_0^s f_{s-r,n}(u, A, \beta)\, dU(r) + \left(1 - U(s)\right) h(u, A)$$

and in consequence of Theorem 2. 2 the characteristic function of $\zeta_{s,n+1}(A)$ is

$$\exp\left\{\int_{T_+}\left[\int_0^s \left(f_{s-r,n}(u, A-t, \beta)\, dU(r) + \left(1-U(s)\right) h(u, A-t)\right) - 1\right]\mu(dt)\right\}.$$

But if $t \in A$

$$\left(1 - U(s)\right) h(u, A-t) - 1 = \left(1 - U(s)\right)(e^{iu} - 1) - U(s)$$

while if $t \notin A$

$$\left(1 - U(s)\right) h(u, A-t) - 1 = -U(s)$$

so the characteristic function of $\zeta_{s,n+1}(A)$ is

$$e^{(1-U(s))(e^{iu}-1)\mu(A)} \exp \int_{T_+} \int_0^s \left(f_{s-r,n}(u, A-t, \beta) - 1\right) dU(r)\mu(dt)$$

which equals to $f_{s,n+1}(u, A, \mu)$, and the assertion is proved.

In Theorem 4. 1 we proved that

$$\mathbf{P}(\zeta_{s,n}(A) \to \zeta_s(A)) = 1.$$

This implies the convergence of the characteristic function $f_{s,n}(u, A, \mu)$ to the characteristic function of $\zeta_s(A)$ for $n \to \infty$.

The following theorem gives a sufficient condition for the convergence of $f_{s,n}(u, A, \mu)$ to $f(u, A, s)$, if $\mu(T_+) = \infty$.

THEOREM 5. 4. *Let us make the same assumptions as in Theorem* 5. 3 *except that instead of the finiteness of* μ *we suppose only that for* $B \in \sigma_T$

$$\mu(B) \leqq K\lambda(B)$$

where K *is a positive constant. Then the characteristic function* $f(u, A, s)$ *of* $\zeta_s(A)$ *has the form*

$$f(u, A, s) = \lim_{n\to\infty} e^{\mu(A)(1-U(s))(e^{iu}-1)} \exp \int_{T_+} \int_0^s (f_{s-r,n}(u, A, \beta) - 1)\, dU(r)\, \mu(dt).$$

PROOF. In the preceding theorem we proved that

$$f_{s,n}(u, A, \mu) = \mathbf{M}(e^{iu\,\zeta_{s,n}(A)}).$$

From the proof of Theorem 5. 2 it follows that

$$\mathbf{P}\left(\lim_{n\to\infty} \zeta_{s,n}(A) = \zeta_s(A)\right) = 1$$

so

$$\lim_{n\to\infty} f_{s,n}(u, A, \mu) = f(u, A, s)$$

and the convergence is uniform in every finite u-interval. Q. e. d.

§ 6. The Converse of the Theorems of Doob and Rényi

A. RÉNYI resp. J. L. DOOB proved the following assertions:

Let us consider on the real line a Poisson random point distribution which is stationary, that is its parameter measure has the form $\mu(A) = C\lambda(A)$, where λ is the Lebesgue-measure. The positive number C is called the *density* of the random point distribution.

A realization of the given random point distribution is $\omega = (t_1, t_2, \ldots)$. Each t_n $(\in \omega)$ is kept with probability p and is cancelled with probability $(1-p)$. The cancelling of points is independent of ω and of cancelling of other points of ω. So we have a new random point distribution about which one can easily prove that it is also a stationary Poisson type with density pC. (Interesting theorems about this type of transformation of random point distributions are contained in [8] and [9].)

Let us consider again the realization ω. Each point of ω will be transformed at random to another point of the real line independently of ω and of the transforming of other points of ω. The probability that the image of the point t belongs

to the Borel set A is $Q(A-t)$, where Q is an arbitrary probability measure on the Borel sets. J. L. DOOB has proved that the random point distribution consisting of the images is also a stationary Poisson random point distribution with unchanged density C ([11]).

In both cases an independent homogeneous single spreading happens and the new random point distribution ζ is again of Poisson type. A. RÉNYI has asked whether there are other types of single spreading in which the point distribution also remains of the Poisson type. The answer is: this can happen only in these two cases and in their combination, i. e. if the underlying points spread no more that one point. We shall treat this problem under more general conditions. We assume that the space T_+ is a locally-compact, σ-compact topological group, so the Haar-measure on T_+ is σ-finite ([10]).

THEOREM 6. 1. *In case of an independent homogeneous single spreading, the underlying distribution of which is of atomless Poisson type, the random point distribution ζ is of Poisson type if and only if with probability* 1 *the number of points spread by an underlying point is less than* 2, *that is*

$$\mathbf{P}(\psi_0(T_+)<2)=1.$$

PROOF. Let us suppose that the random point distribution ζ is of Poisson type. So if $\lambda(A)<\infty$, then there exists a constant c such that the characteristic function of $\zeta(A)$ is

$$f(u, A)=e^{c(e^{iu}-1)}.$$

But using the Theorem 2. 2

$$f(u, A)=\exp\int_{T_+}(g(u, A-t)-1)\mu(dt)$$

therefore

$$\log f(u, A)=\sum_{l=1}^{\infty}(e^{iul}-1)\int_{T_+}\mathbf{P}(\psi_0(A-t)=l)\mu(dt)=c(e^{iu}-1). \tag{6. 1}$$

By the uniqueness of Fourier expansion

$$\int_{T_+}\mathbf{P}(\psi_0(A-t)=l)\mu(dt)=0$$

if $A\in\sigma_T$, $\lambda(A)<\infty$ and $l=2, 3, \ldots$, that is

$$\int_{T_+}\mathbf{P}(\psi_0(A-t)\geqq 2)\mu(dt)=0 \tag{6. 2}$$

if $\lambda(A)<\infty$.

We assert that if $\lambda(A)<\infty$, then

$$\mathbf{P}(\psi_0(A)\geqq 2)=0. \tag{6. 3}$$

If on the contrary there were a set $\hat{A}$, such that $\lambda(\hat{A})<\infty$, and

$$\mathbf{P}(\psi_0(\hat{A})\geqq 2)=\delta>0$$

then we should choose such neighbourhood V of the origin, for which $\lambda(V)<\infty$, $\mu(V)>0$. For the set

$$\hat{A}+V=\{t+\tau: t\in\hat{A}, \tau\in V\}$$

we have

$$\int_{T_+}\mathbf{P}(\psi_0((\hat{A}+V)-t)\geqq 2)\mu(dt)\geqq\int_{V}\mathbf{P}(\psi_0((\hat{A}+V)-t)\geqq 2)\mu(dt).$$

If $t\in V$, then $(\hat{A}+V)-t\supset\hat{A}$, so

$$\int_{V}\mathbf{P}(\psi_0((\hat{A}+V)-t)\geqq 2)\mu(dt)\geqq\int_{V}\mathbf{P}(\psi_0(\hat{A})\geqq 2)\mu(dt)\geqq\delta\mu(V)>0$$

and this contradicts (6. 2), so (6. 3) is valid.

Finally we prove that

(6. 4) $$\mathbf{P}(\psi_0(T_+)\geqq 2)=0.$$

Because of the σ-finiteness of λ there exists a sequence of sets $A_1, \ldots, A_l, \ldots$ $(A_l\subset A_{l+1})$, such that $\lambda(A_l)<\infty$, and $\sum_{l=1}^{\infty}A_l=T_+$. Then

$$\{\psi_0(T_+)\geqq 2\}=\sum_{l=1}^{\infty}\{\psi_0(A_l)\geqq 2\}$$

and so

$$\mathbf{P}(\psi_0(T_+)\geqq 2)\leqq\sum_{l=1}^{\infty}\mathbf{P}(\psi_0(A_l)\geqq 2)=0$$

Thus (6. 4) is valid, and we proved the necessity of the condition.

To prove the sufficiency we remark that from (6. 4) it follows that

(6. 5) $$\mathbf{P}(\psi_0(A)=1)=\mathbf{M}\psi_0(A).$$

Let $A_1, \ldots, A_n$ be disjoint sets from σ_T, and let us denote by $g(u_1, \ldots, u_n; A_1, \ldots, A_n)$ the joint characteristic function of the random variables $\psi_0(A_1), \ldots, \psi_0(A_n)$. From (6. 5) it follows that

$$g(u_1, \ldots, u_n; A_1, \ldots, A_n)=1+\sum_{l=1}^{n}\mathbf{M}\psi_0(A_l)(e^{iu_l}-1).$$

On the basis of facts used in the proof of necessity, the joint characteristic function of random variables $\zeta(A_1), \ldots, \zeta(A_n)$ is

$$f(u_1, \ldots, u_n; A_1, \ldots, A_n)=e^{\int_{T_+}\sum_{l=1}^{n}\mathbf{M}\psi_0(A-t)(e^{iu_l}-1)\mu(dt)}=e^{\sum_{l=1}^{n}(e^{iu_l}-1)\int_{T_+}\mathbf{M}\psi_0(A_l-t)\mu(dt)}$$

that is the random point distribution ζ is of Poisson type. Q. e. d.

Connecting with this result we can easily prove the following theorem.

THEOREM 6. 2. *If for a homogeneous spreading process the underlying distribution is of atomless Poisson type, then from*

$$\mathbf{P}(\psi_0(T_+) \leqq 1) = 1 \tag{6. 6}$$

it follows that ζ_s is a Poisson random point distribution for every s.

PROOF. The condition (6. 6) implies that every underlying point may have only 0 or 1 descendant in time s, that is

$$\mathbf{P}(\zeta_s^{[r]}(T_+) \leqq 1) = 1$$

where $\zeta_s^{[r]}$ was defined in (3. 6). Taking the whole process between the times 0 and s as a single spreading, we can apply Theorem 6. 1, and the theorem is proved.

The parameter measure of ζ_s can be got from Theorem 5. 2.

DEFINITION 6. 1. We say that the random variable of compound Poisson type with characteristic function

$$\exp\left\{\sum_{k=1}^{\infty} c_k(e^{iuk} - 1)\right\}$$

is of order n, if $c_k = 0$ for $k \geqq n+1$.

THEOREM 6. 3. Under the conditions of Theorem 6. 1 the following two properties are equivalent:

(i) If $\lambda(A) < \infty$, then the distribution of $\zeta(A)$ is of compound Poisson type of order n

(ii) $\mathbf{P}(\psi_0(T_+) \leqq n) = 1$.

PROOF. On the basis of (6. 1) the logarithm of the characteristic function of $\zeta(A)$ is

$$\log f(u, A) = \sum_{l=1}^{\infty} (e^{iul} - 1) \int_{T_+} \mathbf{P}(\psi_0(A-t) = l)\mu(dt)$$

and using (i)

$$\log f(u, A) = \sum_{l=1}^{n} c_l(e^{iul} - 1).$$

Hence for $l \geqq n+1$

$$\int_{T_+} \mathbf{P}(\psi_0(A-t) = l)\mu(dt) = 0$$

and with the order of ideas followed in the proof of Theorem 6. 1

$$\mathbf{P}(\psi_0(T_+) \geqq n+1) = 1.$$

If (ii) is valid, then (i) is also valid, because in case of $l \geqq n+1$

$$\int_{T_+} \mathbf{P}(\psi_0(A-t) = l)\mu(dt) \leqq \int_{T_+} \mathbf{P}(\psi_0(T_+) = l)\mu(dt) = 0.$$

Q. e. d.

List of most frequent notations

REFERENCES

[1] HARRIS, TH. E.: *The theory of branching processes,* Berlin—Göttingen—Heidelberg, 1963.
[2] NEYMAN, J.: Sur la théorie probabiliste des amas des galaxies, *Ann. Inst. H. Poincaré* **14** (1955) 201—244.
[3] PRÉKOPA, A.: On the spreading process, *Transaction of the Second Prague Conference* (1960) 521—529.
[4] MOYAL, J. E.: The general theory of stochastic population processes, *Acta Math.* **108** (1962) 1—31.
[5] SZÁSZ, D.: Szóródási folyamatok, *Thesis,* (1966) Budapest.
[6] PRÉKOPA, A.: On secondary processes generated by a random point distribution of Poisson type, *Ann. Univ. Sci. Budapest. Eötvös Sect. Math.* **1** (1958) 153—170.
[7] RÉNYI, A.: Remarks on the Poisson process, *Studia Sci. Math. Hung.* **2** (1967) 119—123.
[8] RÉNYI, A.: A Poisson folyamat egy jellemzése, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **1** (1956) 519—527.
[9] Беляев, Ю. К.: Предельные теоремы для редеющих потоков, *Теория Вероятностей* **8** (1963) 175—184.
[10] HALMOS, P. R.: *Measure theory,* New York, 1950.
[11] DOOB, J. L.: *Stochastic processes,* New York—London, 1953.

EÖTVÖS L. UNIVERSITY, BUDAPEST

(Received September 30, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 249—256.

STATISTICS AND INFORMATION THEORY[1]

by

A. RÉNYI

§ 0. Introduction

In the present paper we deal with certain basic questions connected with the information-theoretic point of view on statistics. This paper is a continuation of the papers [1], [2], [3], [4], [5] of the author; most of the results of these previous papers are presented here in an improved (sharper or more general) form.

§ 1. On the Amount of Information in a Random Variable Concerning Another

In this section we collect the basic definitions and well known results needed in what follows.

Let $S=(\Omega, \mathscr{A}, \mathbf{P})$ be a probability space, i. e. Ω an arbitrary nonempty set, $\mathscr{A}$ a σ-algebra of subsets of Ω and $\mathbf{P}$ a probability measure on $\mathscr{A}$. In what follows θ will always denote a discrete valued random variable in S, i. e., a function $\theta=\theta(\omega)$ defined for $\omega \in \Omega$, taking on only a finite number of different values $\theta_1, \theta_2, \ldots, \theta_r$ ($r \geqq 2$) for which the set (event) $H_k=\{\omega: \theta(\omega)=\theta_k\}$ belongs to $\mathscr{A}$ for $k=1, 2, \ldots, r$. Here $\theta_1, \theta_2, \ldots, \theta_r$ may be numbers, or any distinguishable symbols: their values will be in what follows irrelevant. We shall usually interpret θ as the parameter of a probability distribution and the event H_k as the *hypothesis that the true value of the parameter θ is equal to θ_k*; we shall use the notation

$$p_k=\mathbf{P}(H_k)=\mathbf{P}(\theta=\theta_k) \qquad (k=1, 2, \ldots, r) \tag{1. 1}$$

and call the distribution $(p_1, p_2, \ldots, p_r)$ of θ (contrasting it with the conditional (or posterior) distribution of θ given certain observations, to be introduced later) the *prior distribution* of θ. The (unconditional) entropy of θ is defined by Shannon's formula [2]

$$\mathbf{H}(\theta)=\sum_{k=1}^{r} p_k \log_2 \frac{1}{p_k} \tag{1. 2}$$

where the numbers p_k ($k=1, 2, \ldots, r$) are those defined by (1. 1) $\mathbf{H}(\theta)$ will be interpreted as *the amount of missing information on θ* when nothing else is known about θ except that its prior distribution is given.

[1] This paper has been presented to the 1st European Meeting of Statisticians held in London, 5—10 September 1966.

[2] $\log_2 x$ denotes the logarithm with base 2 of the positive number x; $0 \log_2 \frac{1}{0}$ always means 0.

Let now $\xi = \xi\{\omega\} = (\xi_1(\omega\}, \ldots, \xi_n(\omega))$ be an n-dimensional vector valued random variable, i. e., an $\mathscr{A}$-measurable function defined on Ω and with values in the Euclidean space E_n of dimension n. We shall interpret ξ as an *observed sample*. As ξ and θ are random variables on the same probability space, by observing ξ we usually get some information on θ (except when θ and ξ are independent). After having observed ξ we may consider the *conditional* (or posterior) *distribution*

$$p_k(\xi) = \mathbf{P}(H_k|\xi) \tag{1. 3}$$

of H_k given the value of ξ. The conditional probability of an event $A \in \mathscr{A}$ given the observed value of ξ is as usual defined as follows: Let $\mathscr{A}_\xi$ denote the least σ-algebra of subsets of Ω on which ξ is measurable (i. e., the σ-algebra generated by ξ). By supposition $\mathscr{A}_\xi$ is a subalgebra of $\mathscr{A}$. The conditional probability $\mathbf{P}(A|\xi)$ of an event A, given the value of ξ, is defined as an $\mathscr{A}_\xi$-measurable function (random variable) such that for every $B \in \mathscr{A}_\xi$, one has

$$\int_B \mathbf{P}(A|\xi)\, dP = \mathbf{P}(AB). \tag{1. 4}$$

As well known, $\mathbf{P}(A|\xi)$ is by (1. 4) uniquely defined up to a set of measure 0 and $\{\mathbf{P}(H_1|\xi), \ldots, \mathbf{P}(H_r|\xi)\}$ is with probability one a probability distribution, i. e., $\mathbf{P}\left(\sum_{k=1}^{r} \mathbf{P}(H_k|\xi) = 1\right) = 1$. Let us consider now the entropy of the conditional (a posteriori) distribution of θ given ξ, i. e., the quantity

$$\mathbf{H}(\theta|\xi) = \sum_{k=1}^{r} p_k(\xi) \log_2 \frac{1}{p_k(\xi)}. \tag{1. 5}$$

We interpret $\mathbf{H}(\theta|\xi)$ as *the amount of information concerning* θ *still missing after having observed the sample* ξ. Clearly $\mathbf{H}(\theta|\xi)$ itself is a random variable (which is not only $\mathscr{A}$-measurable but also $\mathscr{A}_\xi$-measurable); its expectation $\mathbf{E}(\mathbf{H}(\theta|\xi))$ is interpreted as *the average amount of information still missing about* θ *after having observed* ξ. We shall call this quantity for the sake of brevity when there is no danger of misunderstanding simply „the amount of missing information", and denote it by $R(\xi, \theta)$; i. e., we put[3]

$$R(\theta, \xi) = \mathbf{E}(\mathbf{H}(\theta|\xi)). \tag{1. 6}$$

The *amount of information* $I(\theta, \xi)$ in the observed sample ξ with respect to the (unknown) parameter θ is defined as the average decrease of the entropy of θ by observing ξ; that is, we put

$$I(\theta, \xi) = \mathbf{H}(\theta) - R(\theta, \xi). \tag{1. 7}$$

Evidently the conditional (posterior) distribution $\{p_1(\xi), \ldots, p_r(\xi)\}$ of θ is identical with its prior distribution $\{p_1, \ldots, p_r\}$ if and only if ξ and θ are independent. In this case $R(\theta, \xi) = \mathbf{H}(\theta)$, i. e., $I(\theta, \xi) = 0$, that is the observation of the sample ξ does not give us any information on θ. In every other case one has $R(\theta, \xi) < \mathbf{H}(\theta)$

[3] Here and in what follows $\mathbf{E}(\eta)$ denotes the expectation of the random variable η.

and thus $I(\theta, \xi) > 0$. This can be shown by Jensen's inequality as follows. As the function $x \log_2 \frac{1}{x}$ is concave in (0, 1) and by Jensen's inequality for any concave function $f(x)$ and any random variable η the values of which are lying in the domain of definition of $f(x)$ one has

$$\mathbf{E}(f(\eta)) \leqq f(\mathbf{E}(\eta)) \tag{1.8}$$

it follows

$$R(\theta, \xi) = \sum_{k=1}^{r} \int_{\Omega} p_k(\xi) \log_2 \frac{1}{p_k(\xi)} dP \leqq \tag{1.9}$$

$$\leqq \sum_{k=1}^{r} \left(\int_{\Omega} p_k(\xi) dp \right) \frac{1}{\left(\int_{\Omega} p_k(\xi) d\xi \right)} = \mathbf{H}(\theta)$$

because by (1. 4)

$$\int_{\Omega} p_k(\xi) dP = \mathbf{P}(H_k) = p_k. \tag{1.10}$$

Evidently there is equality in (1. 9) if and only if the distribution $\{p_1(\xi), \ldots, p_r(\xi)\}$ is (with probability 1) identical to the distribution $\{p_1, \ldots, p_r\}$, i. e., if ξ and θ are independent.

Let $g(x)$ $(x \in E_n)$ be any k-dimensional vector valued Borel measurable function defined on the n-dimensional space E_n. We shall call the random variable $g(\xi)$ a *statistic*. If after observing ξ we consider the value of the statistic $g(\xi)$ only, and disregard every information (on θ) contained in the observation of ξ and not contained in $g(\xi)$, we usually loose some amount of information, i. e.,

$$I(g(\xi), \theta) \leqq I(\xi, \theta). \tag{1.11}$$

The inequality (1. 11) is clearly equivalent to

$$R(\xi, \theta) \leqq R(g(\xi), \theta). \tag{1.12}$$

To prove (1. 12) we need the following Lemma 1 which is an immediate consequence of the definition of conditional probability.

LEMMA 1. *If $f(x)$ is any Borel measurable function and $A \in \mathcal{A}$ any event, we have*

$$\mathbf{E}(f(g(\xi))\mathbf{P}(A|g(\xi))) = \mathbf{E}(f(g(\xi))\mathbf{P}(A|\xi)). \tag{1.13}$$

Using Lemma 1, we obtain

$$R(\theta, g(\xi)) - R(\theta, \xi) = \mathbf{E}\left(\sum_{k=1}^{r} \mathbf{P}(H_k|\xi) \log_2 \frac{\mathbf{P}(H_k|\xi)}{\mathbf{P}(H_k|g(\xi))} \right). \tag{1.14}$$

Now we need the following simple

LEMMA 2. *If $\{q_1, q_2, \ldots, q_r\}$ and $\{Q_1, Q_2, \ldots, Q_r\}$ are arbitrary probability distributions consisting of the same number r of terms, we have*

$$\sum_{k=1}^{r} q_k \log_2 \frac{q_k}{Q_k} \geqq 0 \tag{1.15}$$

with equality standing in (1. 15) *if and only if $q_k = Q_k$ for $k = 1, 2, \ldots, r$.*

Applying Lemma 2, we obtain from (1. 14) that (1. 12) holds and there is equality in (1. 12) if and only if with probability 1, one has

$$\mathbf{P}(H_k|\xi)=\mathbf{P}(H_k|g(\xi)) \qquad (k=1, 2, \ldots, r). \tag{1. 16}$$

If (1. 16) holds (with probability 1) we call $g(\xi)$ a *sufficient function of* ξ *for* θ (or a *sufficient statistic*). Thus a function of the observations is called sufficient for a parameter if and only if it contains all information in the observation which is relevant to the parameter, in the sense that there is equality in (1. 11).

Note that if (1. 16) holds and the random vector ξ has the conditional density $\varphi_k(x)$ under condition H_k, and $g(\varphi)$ has the density $\psi_k(g(x))$, then

$$\varphi_k(x)=\psi_k(g(x))\chi(x)$$

where the function $\chi(x)$ does not depend on k; as clearly $\varphi_k(x)$, $\psi_k(g(x))$ and $\chi(x)$ are all independent from the prior distribution $\{p_1, \ldots, p_r\}$ of θ, it follows that our definition of sufficiency is equivalent with the usual definition of a sufficient statistic in case both definitions are applicable. An advantage of our definition is that it does not depend on the existence of densities; besides it has a clear information-theoretical meaning.

Before proceeding further we prove the following

THEOREM 1. *The conditional distribution* $\Pi(\xi)=(p_1(\xi), \ldots, p_r(\xi))$ *of* θ *given* ξ, *considered as a statistic, is sufficient with respect to* θ.

To prove our theorem it is clearly enough to show that

$$p(H_k|\Pi(\xi))=p_k(\xi) \qquad (k=1, 2, \ldots, n). \tag{1. 17}$$

But (1. 17) is evidently true as $p_k(\xi)$ is $\mathscr{A}_{\Pi(\xi)}$-measurable ($p_k(\xi)$ being the k-th component of the vector $\Pi(\xi)$, we get $p_k(\xi)$ by projecting the vector $\Pi(\xi)$ to the x_k-axis.)

The statement of Theorem 1 can be expressed by saying that the *conditional distribution of* θ *given* ξ *contains all information relevant on* θ *which is present in the sample* ξ.

§ 2. A Bayesian Version of the Fundamental Lemma of Neyman and Pearson

If we have to make a *decision* concerning the parameter θ, on the basis of the observed value of the sample ξ, i. e., after observing ξ we have to select one of the possible values of θ, this decision can be described by a Borel measurable function $D(\xi)$ of ξ, the set of values of which is the set $\{\theta_1, \theta_2, \ldots, \theta_r\}$ of possible values of θ. The *error* e of such a decision is simply the probability of the decision being false, that is

$$e=\mathbf{P}(D(\xi)\neq\theta). \tag{2. 1}$$

We define the *standard decision* $\Delta(\xi)$ as follows: we decide always in favor of that hypothesis H_k (that value θ_k of θ) which has the largest conditional probability given the value of ξ; in case there is more than one value k such that $p_k(\xi)=\max_{1\leq j\leq r} p_j(\xi)$, we select in some way one among those values—say the least such value of k. If another rule is applied we call the corresponding decision a *variant of the standard decision.*

It is easy to see that it does not matter much which one of these values of k we choose (i. e., whether we use the standard decision or one of its variants) as the error of the decision is independent from this selection. As a matter of fact if ε denotes the *error of the standard decision*, we obtain by the definition (1. 4) of conditional probabilities

$$\varepsilon = \mathbf{P}(\Delta(\xi) \neq \theta) = 1 - \mathbf{P}(\Delta(\xi) = \theta) = 1 - \mathbf{E}(\mathbf{P}(\theta = \Delta(\xi)|\xi)). \tag{2. 2}$$

Clearly if we change the definition of the standard decision for some value of ξ from $\Delta(\xi) = \theta_{k_1}$ to $\Delta(\xi) = \theta_{k_2}$ where $p_{k_1}(\xi) = p_{k_2}(\xi)$, then ε remains unchanged, because $\mathbf{P}(\theta = \Delta(\xi)|\xi) = p_{\Delta(\xi)}(\xi)$ is by definition not affected by such a change.

Now let $D(\xi)$ be any other decision, and e its error. Then we get, similarly to (2. 2)

$$e = 1 - \mathbf{E}(\mathbf{P}(\theta = D(\xi)|\xi)). \tag{2. 3}$$

Thus we have

$$e - \varepsilon = \mathbf{E}(\mathbf{P}(\theta = \Delta(\xi)|\xi) - \mathbf{P}(\theta = D(\xi)|\xi)). \tag{2. 4}$$

The random variable, the expectation of which gives the difference $e - \varepsilon$, is clearly always non-negative, because for each value of ξ we have for some value of k (namely $k = D(\xi)$)

$$\mathbf{P}(\theta = \Delta(\xi)|\xi) - \mathbf{P}(\theta = D(\xi)|\xi) = \max_{1 \leqq j \leqq r} p_j(\xi) - p_k(\xi) \geqq 0. \tag{2. 5}$$

Thus we have proved the following

THEOREM 2. *No decision can have a smaller error than the standard decision.*

Clearly if the decision $D(\xi)$ is such that $\mathbf{P}(\theta = D(\xi)|\xi) \neq \mathbf{P}(\theta = \Delta(\xi)|\xi)$ with positive probability, then $e > \varepsilon$. However, if $\mathbf{P}(\theta = D(\xi)|\xi) = \mathbf{P}(\theta = \Delta(\xi)|\xi)$ with probability 1, this means that the decision $D(\xi)$ differs from the decision $\Delta(\xi)$ only in that in case a tie presents itself, i. e., if the value of k for which $p_k(\xi)$ is maximal is not unique, the decision $D(\xi)$ prescribes another choice among those values k for which $p_k(\xi)$ is maximal as $\Delta(\xi)$; thus *except for variants of the standard decision every other decision has a definitely larger error than the standard decision* (or any of its variants).

Note that the difference between Theorem 2 and the usual form of the Neyman—Pearson lemma consists in that we have supposed that the parameter θ is a random variable, i. e. we have taken the Bayesian point of view. Thus we do not distinguish between errors of the first and second kind: only one sort of error is possible. A decision is namely either correct, or wrong, and the error of a decision is the probability of it being wrong. A formal difference of minor importance is that by using the general notion of a conditional probability we did not need any supposition concerning the existence of densities.

Note that it follows from Theorem 2 that the error of the standard decision is $\leqq 1/2$ in the case $r = 2$, because if $\bar{\Delta}$ means the decision which is the opposite of Δ, $\bar{\Delta}$ has the error $1 - \varepsilon$ and thus by Theorem 2, $\varepsilon \leqq 1 - \varepsilon$.

As regards the standard decision $\Delta(\xi)$, we may compute the amount of information contained in the value of $\Delta(\xi)$ with respect to θ, i. e., the quantity $I(\Delta(\xi), \theta)$. Clearly one has $I(\Delta(\xi), \theta) \leqq I(\xi, \theta)$ with strict inequality except when $\Delta(\xi)$ is a sufficient function of ξ concerning θ; thus *even when the best possible decision is adopted some information is lost*. The explanation of this somewhat paradoxically

sounding statement is that usually the information on θ contained in the observed value of ξ is not enough to decide with certainty which is the value of the parameter, it only gives us a (conditional) probability distribution on the possible values. If, nevertheless, we insist on choosing one of the possible values and rejecting all the others, we naturally lose by this a certain amount of information.

§ 3. Estimating the Error of the Standard Decision by the Amount of Missing Information

We prove in this section the following [4]

THEOREM 3. *Let ε denote the error of the standard decision and $R=R(\theta, \xi)$ the amount of missing information, then the following inequality holds*

$$\log_2 \frac{1}{1-\varepsilon} \leqq R, \tag{3. 1}$$

or expressed otherwise

$$\varepsilon \leqq 1 - \frac{1}{2^R}. \tag{3. 2}$$

PROOF OF THEOREM 3. Let us denote for the sake of brevity the event $\Delta(\xi)=\theta_j$ by A_j $(j=1, 2, \ldots)$. Then we have clearly

$$R = \mathbf{E}(\mathbf{H}(\theta|\xi)) = \sum_{j=1}^{r} \mathbf{P}(A_j)\mathbf{E}(\mathbf{H}(\theta|\xi)|A_j). \tag{3. 3}$$

(Here and in what follows $\mathbf{E}(\eta|B)$ denotes the conditional expectation of the random variable η with respect to the condition B, when B is an event such that $\mathbf{P}(B)>0$.) Now by definition under condition A_j we have $p_k(\xi) \leqq p_j(\xi)$ for $k=1, 2, \ldots, r$; in view of (1. 5) we get that

$$R \geqq \sum_{j=1}^{r} \mathbf{P}(A_j)\mathbf{E}\left(\log_2 \frac{1}{p_j(\xi)} \,\middle|\, A_j\right). \tag{3. 4}$$

Applying now Jensen's inequality to the convex function $\log_2 \frac{1}{x}$ $(0 \leqq x \leqq 1)$, it follows that

$$R \geqq \sum_{j=1}^{r} \mathbf{P}(A_j) \log_2 \frac{1}{\mathbf{E}(p_j(\xi)|A_j)}. \tag{3. 5}$$

Now it follows from (1. 4) that

$$\mathbf{E}(p_j(\xi)|A_j) = \frac{\int_{A_j} \mathbf{P}(\theta = \theta_j|\xi)\, dP}{\mathbf{P}(A_j)} = \mathbf{P}(\Delta(\xi) = \theta|A_j). \tag{3. 6}$$

[4] In our previous paper [3] we have proved only the weaker estimate $\varepsilon \leqq R$. Clearly (3. 1) implies not only $\varepsilon \leqq R$ but also $\frac{\varepsilon}{\ln 2} \leqq R$.

Thus we obtain from (3. 5)

$$R \geqq \sum_{j=1}^{r} \mathbf{P}(A_j) \log_2 \frac{1}{\mathbf{P}(\Delta(\xi) = \theta | A_j)}. \tag{3. 7}$$

We need now Jensen's inequality, in the form that if $f(x)$ is a convex function, $x_1, \ldots, x_r$ any values in the domain of definition of $f(x)$ and $w_1, \ldots, w_r$ non-negative numbers with sum equal to one, then

$$\sum_{j=1}^{r} w_j f(x_j) \geqq f\left(\sum_{j=1}^{r} w_i x_j \right). \tag{3. 8}$$

Applying (3. 8) it follows from (3. 7) that

$$R \geqq \log_2 \frac{1}{\sum_{j=1}^{r} \mathbf{P}(A_j) \mathbf{P}(\Delta(\xi) = \theta | A_j)} = \log_2 \frac{1}{\mathbf{P}(\Delta(\xi) = \theta)} = \log_2 \frac{1}{1-\varepsilon} \tag{3. 9}$$

and this proves (3. 1).

In our previous paper [4] we have shown for the special case $r=2$ that the inequality $2\varepsilon \leqq R$ holds; for this special case this is slightly better than (3. 1). We reproduce here the proof of this inequality as it requires only a few lines. Let the possible values of θ be θ_0 and θ_1, the corresponding hypotheses $\theta = \theta_0$ and $\theta = \theta_1$ shall be denoted by H_0 and H_1 respectively. Put

$$h(x) = x \log_2 \frac{1}{x} + (1-x) \log_2 \frac{1}{1-x}. \tag{3. 10}$$

Then we have evidently $h(x) = h(1-x)$ and $h(x) \geqq 2x$ for $0 \leqq x \leqq 1/2$. Let us put

$$p^*(\xi) = \begin{cases} p_0(\xi) & \text{if } p_0(\xi) \leqq \frac{1}{2} \text{ i. e. if } \Delta(\xi) = \theta_1 \\ p_1(\xi) & \text{if } p_0(\xi) \geqq \frac{1}{2} \text{ i. e. if } \Delta(\xi) = \theta_0. \end{cases} \tag{3. 11}$$

Then we have clearly $p^*(\xi) \leqq \frac{1}{2}$ further

$$R = \mathbf{E}(h(p^*(\xi))) \geqq 2\mathbf{E}(p^*(\xi)). \tag{3. 12}$$

Denoting the event $\Delta(\xi) = \theta_0$ by B_0 and the event $\Delta(\xi) = \theta_1$ by B_1 we obtain

$$R \geqq 2\left(\int_{B_1} p_0(\xi) dP + \int_{B_0} p_1(\xi) dP \right). \tag{3. 13}$$

As by (1. 4) we have

$$\int_{B_1} p_0(\xi) dP = \mathbf{P}(H_0 B_1) \quad \text{and} \quad \int_{B_0} p_1(\xi) dP = \mathbf{P}(H_1 B_0) \tag{3. 14}$$

it follows that

$$R \geqq 2(\mathbf{P}(H_0 B_1) + \mathbf{P}(H_1 B_0)) = 2\varepsilon, \tag{3. 15}$$

which was to be proved.

Returning to the general case, we mention that one can also get an upper bound for the amount of missing information by means of the error of the standard decision. In this direction the following theorem is known (see [6] p. 35.).

THEOREM 4. *One has*

$$R \leqq h(\varepsilon) + \varepsilon \log_2(r-1) \tag{3. 16}$$

where $h(x)$ *is defined by* (3. 10).

My thanks are due to G. KATONA, who called my attention to the fact that the estimation (3.16), proved first by R. M. FANO [7], is slightly sharper than a similar estimate which I have found previously.

§ 4. Conclusion

It follows from Theorems 3 and 4 that if we have an infinite sequence of observations $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ each ξ_n being a random variable on the probability space S (it is not a restriction to suppose that each ξ_n is real valued), and $\xi^{(n)}$ denotes the sample $(\xi_1, \xi_2, \ldots, \xi_n)$ further Δ_n the standard decision concerning the true value of θ taken on the basis of observing the sample $\xi^{(n)}$ and ε_n the error of the decision Δ_n, and if finally R_n denotes the average amount of information on θ still missing after having observed the sample $\xi^{(n)}$, then $\lim_{n\to\infty} \varepsilon_n = 0$ if and only if $\lim_{n\to\infty} R_n = 0$. This shows that to get in the limit all information on θ which is needed, is equivalent with having the possibility to make decisions on the true value of θ the probability of correctness of which is in the limit equal to 1. By other words the information-theoretical point of view is in accordance with the usual point of view of statistics.

REFERENCES

[1] RÉNYI, A.: On the amount of information concerning an unknown parameter in a sequence of observations, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **9** (1964) 617—625.

[2] RÉNYI, A.: On the amount of information in a frequency count, *35th Session of the International Statistical Institute,* Beograd, 1965, 1—8.

[3] RÉNYI, A.: On the amount of missing information and the Neyman-Pearson lemma, *Festschrift for J. Neyman,* Wiley, London, 1966, 281—288.

[4] RÉNYI, A.: On the amount of information in a random variable concerning an event, *Journal of Mathematical Sciences (Delhi)* **1** (1966) 30—33.

[5] RÉNYI, A.: On some basic problems of statistics from the point of view of information theory, *Proceedings of the 5th Berkeley Symposium* (in print).

[6] FEINSTEIN, A.: *Foundations of Information Theory*, McGraw-Hill, New York, 1958.

[7] FANO, R. M.: *Statistical Theory of Communication*, MIT, Cambridge, Mass., 1954.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received October 8, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 257—263

ON THE DENSEST PACKING OF CIRCLES NOT BLOCKING EACH OTHER

by

A. HEPPES

Introduction. It is a common requirement in the design of parking areas that each car should be able to leave its place without disturbing the position of the others. In the present paper we shall approach the problem of finding the most economic parking system by giving an upper estimate for the number (or density) of congruent circles which can be packed in a given domain without blocking each other's way out of the domain[1].

Definitions, results. Consider a set of disjoint circles lying in a finite domain D. We shall say that the circles do not block each other if to each circle there is a continuous motion which carries it out of the convex hull of D without entering the other circles or disturbing their position. The upper estimate we are going to give for the number of unit discs which can be placed in D can be considered as an estimate for the packing density. The result can then be extended to unbounded domains e. g. for the whole plane[2]. Our estimate cannot be improved for certain domains. In the case of the whole plane the estimate provides a density which is somewhat greater than that of the "best expected arrangement", namely double rows of touching circles divided by narrow, slightly winding roads (Fig. 1). (For infinite domains the removal of a disc means that the distance through which it can be moved is not limited.) The density of this arrangement is $\frac{\pi}{\sqrt{12}} \cdot \frac{\sqrt{5}-1}{2} =$ $= 0.56050\ldots$, while our upper limit is $0.56518\ldots$.

Our method enables us to obtain the same density estimate under weaker requirements on the arrangement of the circles. We call a system of unit circles *approachable* if each circle can be approached and touched by a circular „vehicle" of the same size, coming from outside the convex hull of D. This condition is really weaker because it does not imply the possibility of moving each circle out of D. Moreover we shall deal with *r-approachable* packings, i. e. packings whose unit circles can be approached by a circular vehicle of radius r. Thus we can provide an estimate for the number of equal barrels which can be stored standing in a given cellar such that the brewmaster (of given circumradius) can go to each of them.

We shall prove the following

[1] This problem has been raised by G. FEJES TÓTH. He studied related questions in his paper "Über die Blockierungszahl einer Kreispackung", Elemente der Mathematik, **19** (1964) 49—53.

[2] For the exact definition of the density of a set of circles with respect to unbounded domains see L. FEJES TÓTH, *Regular Figures*, Pergamon Press, 1964, p. 161.

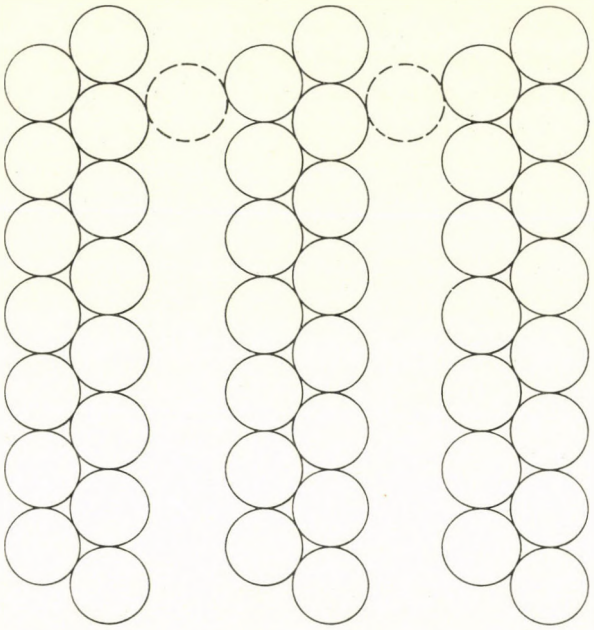


Fig. 1

THEOREM: *Let D be a simply connected bounded domain containing an r-approachable packing of n unit circles, $r \geqq \frac{2}{\sqrt{3}} - 1$. The area D_r of the outer paralleldomain $\boldsymbol{D_r}$ of radius r of the domain D satisfies the inequality*

$$D_r \geqq nA_r + 2\pi\left(\frac{(1+r)^2}{2} - \frac{\sqrt{3}}{\pi}\right)$$

where $A_r = \left(\frac{(1+r)^2}{2} - \frac{\sqrt{3}}{\pi}\right)\omega + \sqrt{3}\left(1 + \frac{\omega}{\pi}\right) + \sqrt{r^2+2r}$ and $\frac{\omega}{2} = \arcsin\frac{1}{1+r}$.

A_r is the area of the domain shown in Fig. 2. For the special case $r=1$, when $\omega = \frac{\pi}{3}$ we have

$$D_1 \geqq n\left[\frac{2}{3}\pi + 2\sqrt{3}\right] + 4\pi - 2\sqrt{3}.$$

Consequently, the density of the packing in the plane satisfies

$$\delta_r \leqq \frac{\pi}{A_r}, \quad \text{and in particular} \quad \delta_1 \leqq \frac{3\pi}{2\pi + 6\sqrt{3}} = 0.56518\ldots.$$

In the following the angles and the arcs are considered to be directed angles and arcs, and are measured in positive sense.

Lemmas. We define the angular area at 0 of a triangle AOB to be the quotient of its area and its angle at 0.

LEMMA 1: *Let AOB be a triangle with the following properties*

(a) *it contains the sector AOB of the unit circle C_1 of center O,*

(b) *neither A nor B lies in the interior of the circle C_2 of radius $2/\sqrt{3}$, concentric with C_1.*

Then the angular area of AOB with respect to its vertex O is $\geqq \sqrt{3}/\pi$. Equality holds only if the triangle is regular with sidelength $2/\sqrt{3}$.

We can suppose that the side AB meets C_2 and has a chord $A'B'$ in common with it, since otherwise the statement is trivially true. Obviously the angular area of the isosceles triangle $A'OB'$ is greater than that of AOB with the only exception that the two triangles coincide. If $A'OB'$ is not regular then the angle AOM, where M is the midpoint of the segment $A'B$, is smaller than $\pi/6$. We now compare $A'OM$ with the triangle $A'OV$ having an angle of $\pi/6$ at O and a right angle at V (Fig. 3). Clearly, the value of the angular area decreases if we replace $A'OM$ by the common part of $A'OM$ and $A'OV$ and then that by $A'OV$, which is the half of a regular triangle of sidelength $2/\sqrt{3}$. This proves Lemma 1.[3]

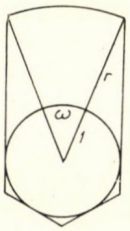


Fig. 2

LEMMA 2: *Let C_1, C_2 and C_3 be three concentric circles of radii 1, $2/\sqrt{3}$ and $1+r>2/\sqrt{3}$, respectively, with common center O. Let $A_1, A_1', \ldots, A_k, A_k'$ ($k \geqq 1$) be points on C_3 in cyclic order such that the angle $A_k'OA_1 > \frac{2\pi}{3} - \omega$, where ω denotes the central angle of a chord of length 2 of C_3.*

If P is a convex polygon with the properties

(a′) *P contains C_1,*

(b′) *P contains the arcs A_iA_i' of C_3, $i=1, \ldots, k$, and*

(c′) *all vertices of P lie outside or on the boundary of C_2,*

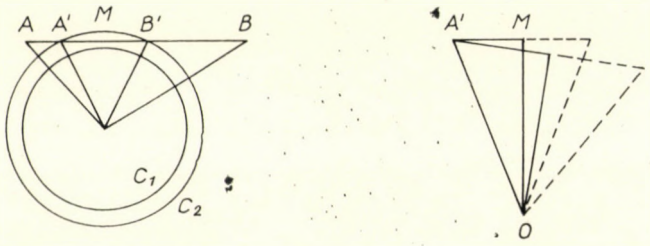


Fig. 3

[3] Similar results have been used for density estimation in a paper of J. MOLNÁR, Körelhelyezések állandó görbületű felületeken, *MTA III. Oszt. Közl.*, **12** (1962) 223—263.

then the area of the polygon satisfies

$$P \geqq \left(\frac{(1+r)^2}{2} - \frac{\sqrt{3}}{\pi}\right)\alpha + \sqrt{3}\left(1+\frac{\omega}{\pi}\right) + \sqrt{r^2+2r}$$

where α *denotes the sum of the angles* A_iOA_i', $i=1, \ldots, k$.

Let B_1 and B_k be the points of the circle C_2 for which both angles B_1OA_1 and $A_k'OB_k$ equal $\frac{\pi}{3}-\frac{\omega}{2}$. Under this condition both lines B_1A_1 and $A_k'B_k$ are tangent to C_1. Therefore, by the convexity of P, the points B_1 and B_k and also the triangles A_1OB_1 and B_kOA_k' belong to P (Fig. 4).

Now decompose P into sectors by the rays $OA_1, OA_1', \ldots, OA_k, OA_k', OB_k, OB_1$. The area of the sector S_i determined by OA_i and OA_i' clearly exceeds the area of the corresponding sector of the circle C_3, $i=1, \ldots, k$. Thus the sum S of the areas of $S_1, S_2, \ldots, S_k$ satisfies

$$S \geqq (1+r)^2 \cdot \frac{\alpha}{2}.$$

On the other hand the sum T of the areas of the triangles A_1OB_1 and $A_k'OB_k$ is

$$T = \sqrt{r^2+2r} - \frac{1}{\sqrt{3}}.$$

The rays passing through the vertices of P, lying in the interior of one of the angles $A_1'OA_2, A_2'OA_3, \ldots, A_{k-1}'OA_k$ or B_kOB_1, decompose the corresponding sectors into triangles containing the corresponding sectors of C_1 and having sides emerging from O, not shorter than $2/\sqrt{3}$. By Lemma 1 the sum U of the areas of these triangles satisfies

$$U \geqq \frac{\sqrt{3}}{\pi} \cdot \sigma$$

where σ denotes the sum of the angles of the sectors in question. But $\sigma = 2\pi - \alpha - 2\left(\frac{\pi}{3}-\frac{\omega}{2}\right)$; hence

$$U \geqq \frac{\sqrt{3}}{\pi}\left[2\pi - \alpha - 2\left(\frac{\pi}{3}-\frac{\omega}{2}\right)\right].$$

Consequently,

$$P \geqq S+T+U \geqq \left[\frac{(1+r)^2}{2} - \frac{\sqrt{3}}{\pi}\right]\alpha + \sqrt{3}\left(1+\frac{\omega}{\pi}\right) + \sqrt{r^2+2r}.$$

PROOF OF THE THEOREM. Let D be a bounded domain containing an r-approachable packing of n unit circles. The circles of radius $1+r$ concentric to the circles of the packing lie in the domain D_r, the outer paralleldomain of radius r of D. We shall refer to these circles as great circles of the unit circles of the packing. The defining property of the r-approachable packing implies that for each circle there exists a curve connecting a point of the corresponding great circle with infinity, without entering any of the great circles. For the sake of simplicity we shall suppose

that the great circles have general position i. e. that no two of them are tangent and no three of them have a common point of intersection. Since small changes of the positions of the circles do not influence the upper bound of the density this assumption imposes no restriction.

The union of the great circles consists of one or several "islands", lying in D_r, such that every great circle adjoins the "ocean" along one or more "shore arcs". It follows from our assumption that a shore arc never consists of a single point[4]. For proving the Theorem we shall consider only a single island containing m circles and show that the area $I(m)$ of this island satisfies the inequality

$$I(m) \geqq mA_r + 2\pi\left(\frac{(1+r)^2}{2} - \frac{\sqrt{3}}{\pi}\right).$$

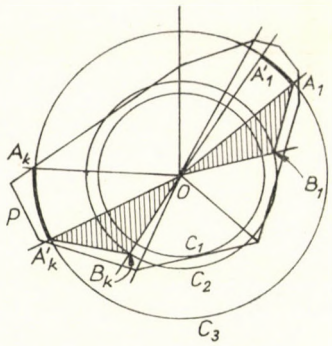


Fig. 4

The island is bounded by $m' \geqq m$ circular arcs convex outward. Going around the boundary counterclockwise we turn continuously to the left while following an arc and turn to the right at points of intersection of consecutive arcs. Because the corresponding unit discs do not overlap, their centers and the point of intersection of the great circles form an isosceles triangle with a base $\geqq 2$ and sides $1+r$, $1+r$. Consequently, the boundary of the island turns to the right at points of intersection with angles $\geqq 2 \arcsin \frac{1}{1+r} = \omega$.

Let us denote the boundary arcs of the island in cyclic order as well as their central angles, by $\alpha_1, \alpha_2, \ldots, \alpha_{m'}$, and the angle of right turn succeeding α_i by β_i, $i = 1, \ldots, m'$ ($\alpha_{m'+1} = \alpha_1$, $\beta_{m'+1} = \beta_1$) (Fig. 5). A complete turn along the boundary shows that

$$\sum_{i=1}^{m'} \alpha_i - \sum_{i=1}^{m'} \beta_i = 2\pi. \tag{1}$$

Combining this with the inequality $\beta_i \geqq \omega$, proved above, we have

$$\sum_{i=1}^{m'} \alpha_i \geqq m' \cdot \omega + 2\pi, \tag{2}$$

or with other words, the average left turn of an arc is greater than ω.

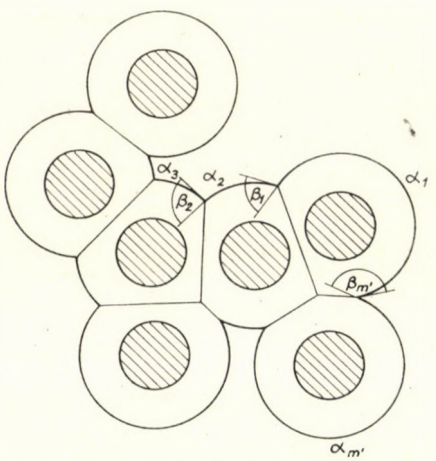


Fig. 5

[4] If a component (island) of the union of the great circles is not simply connected, then the boundaries of the "holes", not being connected with the ocean, will not be considered as shore arcs or parts of the boundary of the island, and the holes will be considered as parts of the island.

Consider now the DIRICHLET cells[5] of the circles as a decomposition of the island. It is easy to see that the cells are convex and that the non-polygonal part of the boundary of a cell consists of the shore arcs of the corresponding great circle. We shall say that the cell is good if on the great circle there exists an arc of central angle $\frac{2\pi}{3}-\omega$ which has no common point with the shore arcs.

Suppose first that all cells are good cells. Then Lemma 2 can be applied to estimate the areas of the cells. Thus the sum of the areas of the cells i. e. the area of the island satisfies

$$I(m) \geqq \left(\frac{(1+r)^2}{2}-\frac{\sqrt{3}}{\pi}\right)\left(\sum_{i=1}^{m'}\alpha_i\right)+m\left(\sqrt{3}\left(1+\frac{\omega}{\pi}\right)+\sqrt{r^2+2r}\right).$$

Hence, applying (2) we have

$$I(m) \geqq m\left[\left(\frac{(1+r)^2}{2}-\frac{\sqrt{3}}{\pi}\right)\omega+\sqrt{3}\left(1+\frac{\omega}{\pi}\right)+\sqrt{r^2+2r}\right]+2\pi\left(\frac{(1+r)^2}{2}-\frac{\sqrt{3}}{\pi}\right),$$

proving the Theorem for this case.

Unfortunately, it may happen that some of the cells are not good. Let us denote by γ an arc of such a cell, determined by the endpoint of one of the shore arcs, say α_s, and the beginning of the next shore arc on the same great circle. Finally, let δ be an arc of central angle $\frac{2\pi}{3}-\omega$, containing γ. Next we reduce the shore arcs lying on the great circle in question. For each shore arc α_j, let α_j' be the intersection of α_j and the complement of δ. Then

$$\sum\alpha_j-\sum\alpha_j' \leqq \frac{2\pi}{3}-\omega-\gamma, \tag{3}$$

where the summation extends over the shore arcs of the great circle.

We define

$$\beta_S' = \beta_S-\left(\frac{2\pi}{3}-\omega-\gamma\right). \tag{4}$$

By assumption $\gamma<\frac{2\pi}{3}-\omega$ and therefore, since the angle of the right turn after α_s is $\beta_s \geqq \pi-\gamma$ we have $\beta_s' \geqq \pi-\gamma-\left(\frac{2\pi}{3}-\omega-\gamma\right)>\omega$.

We now perform the corresponding reductions in the angles related to the other "bad" cells and then define $\alpha_k'=\alpha_k$ and $\beta_l'=\beta_l$ for the cases when α_k' and β_l' has not been defined otherwise. The total reduction in shore arcs does not exceed the

[5] The DIRICHLET cell of a circle consists of those points of the island which are nearer to the circle in question than to any other circle of the packing. As is well known in the case of packing of unit circles, each cell is the intersection of the island and a convex polygon which contains the circle and has vertices whose distances from the center of the circle are $\geqq 2/\sqrt{3}$.

total reduction of the angles β_i, i. e. it holds

$$(1') \qquad \sum_{i=1}^{m'} \alpha_i' - \sum_{i=1}^{m'} \beta_i' \geqq \sum_{i=1}^{m'} \alpha_i - \sum_{i=1}^{m'} \beta_i = 2\pi$$

and, since $\beta_i' > \omega$, we have

$$(2') \qquad \sum_{i=1}^{m'} \alpha_i' \geqq m' \cdot \omega + 2\pi.$$

But the reduced angles admit the application of Lemma 2, which was made above, for all cells. This completes the proof of the Theorem.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received December 2, 1966.)

INDEX

Printed in Hungary

A kiadásért felel az Akadémiai Kiadó igazgatója – Műszaki szerkesztő: Farkas Sándor
A kézirat nyomdába érkezett: 1967. I. 6. – Terjedelem: 23,25 (A/5) ív, 16 ábra

66-6435 Szegedi Nyomda

Die *Studia Scientiarum Mathematicarum Hungarica* ist eine Halbjahrsschrift der Ungarischen Akademie der Wissenschaften. Sie veröffentlicht Originalbeiträge aus dem Bereiche der Mathematik in deutscher, englischer, französischer oder russischer Sprache. Es erscheint jährlich ein Band.

Adresse der Redaktion: Budapest V., Reáltanoda u. 13—15, Ungarn.
Technischer Redaktor: Gy. Katona.

Abonnementspreis pro Band (pro Jahr): 165.— Ft. Bestellbar bei Buch- und Zeitungs-Aussenhandelsunternehmen *Kultúra* (Budapest 62, P. O. B. 149), oder bei den Vertretungen im Ausland.
Austauschabmachungen können mit der Bibliothek des Mathematischen Instituts (Budapest V., Reáltanoda u. 13—15) getroffen werden.
Die zur Veröffentlichung bestimmten Manuskripte sind in zwei Exemplaren an die Redaktion zu schicken.

Studia Scientiarum Mathematicarum Hungarica est une revue biannuelle de l'Académie Hongroise des Sciences publiant des essais originaux, en français, anglais, allemand ou russe, du domaine des mathématiques.

Rédaction: Budapest V., Reáltanoda u. 13—15, Hongrie.
Rédacteur téchnique: Gy. Katona

Le prix de l'abonnement: 165 Forints par an (volume). On s'abonne chez *Kultúra*, Société pour le Commerce de Livres et Journaux (Budapest 62, P. O. B. 149) ou chez ses représentants à l'étranger.
Pour établir des relations d'échange on est prié de s'adresser à la Bibliothèque de l'Institut de Mathématique (Budapest V., Reáltanoda u. 13—15).
On est prié d'envoyer les articles destinés à la publication en deux exemplaires à l'adresse de la rédaction.

Studia Scientiarum Mathematicarum Hungarica — выходит два раза в год в издании Академии наук Венгрии. Журнал публикует оригинальные исследования в области математики на немецком, английском, французском и русском языках. Отдельные выпуски составляют ежегодно один том.

Адрес редакции: Budapest V., Reáltanoda u. 13—15, Венгрия.
Технический редактор: Gy. Katona

Подписная цена на год (за один том): 165 форинтов. Подписка на журнал принимается Внешнеторговым предприятием „Культура" (Budapest 62, P. O. B. 149) или его представительствами за границей.
По поводу отношения обмена просим обращаться к Библиотеке Института Математики (Budapest V., Reáltanoda u. 13—15).
Работы, предназначенные для опубликования в журнале следует направлять по адресу редакции в двух экземплярах.

Ára: 100,— Ft

All the reviews of the Hungarian Academy of Sciences may be obtained among others from the following bookshops:

ALBANIA

Ndermarja Shtetnore e Botimeve
Tirana

AUSTRALIA

A. Keesing
Box 4886, GPO
Sidney

AUSTRIA

Globus Buchvertrieb
Salzgries 16
Wien I.

BELGIUM

Office International de Libirairie
30, Avenue Marnix
Bruxelles 5
Du Monde Entier
5, Place St. Jean
Bruxelles

BULGARIA

Raznoiznos
1 Tzar Assen
Sofia

CANADA

Pannonia Books
2 Spadina Road
Toronto 4, Ont.

CHINA

Waiwen Shudian
Peking
P. O. B. Nr. 88.

CHECHOSLOVAKIA

Artia A. G.
Ve Smeckách 30
Praha II.
Postova Novinova Sluzba
Dovoz tisku
Vinohradska 46
Praha 2
Postova Novinova Sluzba
Dovoz tlace
Leningradska 14
Bratislava

DENMARK

Ejnar Munksgaard
Nörregade 6
Kopenhagen

FINLAND

Akateeminen Kirjakauppa
Keskuskatu 2
Helsinki

FRANCE

Office International de Documentation et Libraire
48, rue Gay Lussac
Paris 5

GERMAN DEMOCRATIC REPUBLIC

Deutscher Buch-export und Import
Leninstraße 16.
Leipzig C. I.
Zeitungsvertriebsamt
Clara Zetkin Straße 62.
Berlin N. W.

GERMAN FEDERAL REPUBLIC

Kunst und Wissen
Erich Bieber
Postfach 46.
7 Stuttgart S.

GREAT BRITAIN

Collet's' Subscription Dept.
44—45 Museum Street
London W. C. I.
Robert Maxwell and Co. Ltd.
Waynflete Bldg. The Plain
Oxford

HOLLAND

Swetz and Zeitlinger
Keizersgracht 471—487
Amsterdam C.
Martinus Nijhof
Lange Voorhout 9
The Hague

INDIA

Current Technical Literature
Co. Private Ltd.
Head Office:
India House OPP.
GPO Post Box 1374
Bombay I.

ITALY

Santo Vanasia
71 Via M. Macchi
Milano
Libreria Commissionaria Sansoni
Via La Marmora 45
Firenze

JAPAN

Nauka Ltd.
2 Kanada-Zimbocho 2-chome
Chiyoda-ku
Tokyo
Maruzen and Co. Ltd.
P. O. Box 605
Tokyo
Far Eastern Booksellers
Kanada P. O. Box 72
Tokyo

KOREA

Chulpanmul
Korejskoje Obschestvo po Exportu i Importu Proizvedenij Pechati
Phenjan

NORWAY

Johan Grundt Tanum
Karl Johansgatan 43
Oslo

POLAND

Export und Import Unternehmen RUCH
ul. Wilcza 46.
Warszawa

ROUMANIA

Cartimex
Str. Aristide Briand 14—18.
Bucuresti

SOVIET UNION

Mezhdunarodnaja Kniga
Moscow
G—200

SWEDEN

Almquist and Wiksell
Gamla Brogatan 26
Stockholm

USA

Stechert Hafner Inc.
31 East 10th Street
New York 3 N. Y.
Walter J. Johnson
111 Fifth Avenue
New York 3. N. Y.

VIETNAM

Xunhasaba
Service d'Export et d'Import des Livres et Périodiques
19. Tran Quoc Toan
Hanoi

YUGOSLAVIA

Forum
Vojvode Misica broj 1.
Novi Sad
Jugoslovenska Kniga
Terazije 27.
Beograd

29. IV. 1967 **Index: 26 748**

Studia Scientiarum Mathematicarum Hungarica

AUXILIO
CONSILII INSTITUTI MATHEMATICI
ACADEMIAE SCIENTIARUM HUNGARICAE

REDIGIT
A. RÉNYI

ADIUVANTIBUS
M. ARATÓ, L. FEJES TÓTH, T. FREY,
G. FREUD, L. KALMÁR, A. PRÉKOPA,
K. TANDORI

TOMUS II.
FASC. 3–4.
1967

AKADÉMIAI KIADÓ, BUDAPEST

Studia Scientiarum Mathematicarum Hungarica

A Magyar Tudományos Akadémia matematikai folyóirata

Szerkesztőség: Budapest V., Reáltanoda u. 13—15.

Teehnikai szerkesztő: Katona Gy.

Kiadja az Akadémiai Kiadó, Budapest V., Alkotmány u. 21.

A *Studia Scientiarum Mathematicarum Hungarica* angol, német, francia vagy orosz nyelven közöl eredeti értekezéseket a matematika tárgyköréből. Félévenként jelenik meg, évi egy kötetben.

Előfizetési ára belföldre 120,— Ft, külföldre 165,— Ft. Megrendelhető a belföld számára az Akadémiai Kiadónál, a külföld számára pedig a Kultúra Könyv és Hírlap Külkereskedelmi Vállalatnál (Budapest II., Fő u. 32).

Cserekapcsolatok felvétele ügyében kérjük az MTA Matematikai Kutató Intézete Könyvtárához (Budapest V., Reáltanoda u. 13—15) fordulni.

Közlésre szánt dolgozatokat kérjük két példányban a szerkesztőség címére küldeni.

Studia Scientiarum Mathematicarum Hungarica is a journal of the Hungarian Academy of Sciences publishing original papers on mathematics, in English, German, French or Russian.

It is published semiannually, making up one volume per year.

Editorial Office: Budapest V., Reáltanoda u. 13—15, Hungary.
Technical Editor: Gy. Katona

Subscription rate: Ft 165 per volume. Orders may be placed with *Kultúra* Trading Co. for Books and Newpapers, Budapest 62, P.O.B. 149 or with its representatives abroad.

For establishing exchange relations please write to the Library of the Mathematical Institute (Budapest V., Reáltanoda u. 13—15).

Papers intended for publication should be sent to Editor in 2 copies.

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 265—273.

QUASI-CONVEXITY AND QUASI-MONOTONICITY IN NONLINEAR PROGRAMMING[1]

by

B. MARTOS

We inspect a series of theorems which have played an important role in the theory of nonlinear programming. In these theorems, the *sufficient* convexity or linearity requirements which have usually been set up for the involved functions turn out to be *unnecessary* and substitutable by weakened assumptions which are (in many case and in a defined sense) even necessary. The whole business is based upon the notion of quasi-convex (and related) functions. For objective functions of this kind, even continuity requirements are dispensable.

Significant characteristics of quasi-convex and/or quasi-monotonic functions have been investigated by DEFINETTI [5], FENCHEL [6], BERGE [3] and DEÁK [4], without reference to nonlinear programming problems. In this latter context, quasi-convexity appears first in the following publications: ARROW and ENTHOVEN [1], ARROW, HURWICZ and UZAWA [2], and KOVÁCS [9]. The first two are devoted primarily to generalizing the basic KUHN—TUCKER theorems, a subject excluded from the present paper. The third one deals with the extension of ROSEN's gradient-projection method to quasi-concave maximization, and contains a less sharp version of our Theorem 3a.

This short prehistory of the present subject must be supplemented, unfortunately enough, by a list of papers which, beside correct or partially correct theorems, happen also to contain mistakes and inaccuracies. This list consists of the articles of DEFINETTI [5], HANSON [7], HOÁNG TUY [8], and MARTOS [10]. FENCHEL [6] corrected DEFINETTI's mistake and MARTOS [11] disproved HANSON's duality theorem. The relationship of the present results to those contained in HOÁNG TUY [8] and MARTOS [10] will be made clear in what follows.

Notations and Definitions

We apply capitals to denote sets, lower case letters for vectors (also called points) and Greek letters for scalars. All the occurring sets are subsets of the Euclidian n-space, E^n.

$[x, y]$ denotes a closed straight segment connecting the points x and y; (x, y) is an open one.

[1] The present paper contains the essential part of paragraphs 4.—6. of my paper [11], published in the Hungarian language. Slight modifications in Theorem 5. and Lemma 2. (as compared with the less accurate Theorems 7. 9 and 7. 10 in the referred paper) are worth mentioning. These improvements (and a fair deal of others) are due to the highly appreciated criticism of ERVIN DEÁK (Math. Inst. Hung. Acad. Sci.). Of course, he can not be held responsible for any remaining mistakes.

$x \leqq y$ stands for the same kind of inequality in each corresponding component. An upper index like x^0, x^1, ... refers to a selected value of the variable vector x; a lower index distinguishes the components of a vector. For the reader's convenience, we recall a few well-known definitions.

Convex set. The set X is convex if $x^1, x^2 \in X$ implies $[x^1, x^2] \subset X$.

Polyhedral set. A polyhedral set $\hat{X}$ is an intersection of closed halfspaces of finite number. ($\hat{X}$ is thereby closed and convex.)

Polyhedron. A bounded polyhedral set (marked like X^{Δ}).

Edge. Let S be a straight line in E^n containing two different points of the closed convex set $X \subset E^n$. $S \cap X$ is an edge of X if $X - (S \cap X)$ is convex. (Note that X is not assumed to be polyhedral. A finite or infinite closed segment in E^1 is an edge of itself.)

Vertex. $\hat{x} \in \hat{X}$ is a vertex of the polyhedral set $\hat{X}$ if $\hat{X} - \{\hat{x}\}$ is convex.

Adjacent vertices. Two different vertices $\hat{x}^1, \hat{x}^2 \in \hat{X}$ are adjacent if $\hat{X} - [\hat{x}^1, \hat{x}^2]$ is convex. (Or, equivalently, if they lie on the same edge of $\hat{X}$. A vertex is not adjacent to itself.)

In the following definitions, let X be a convex subset of E^n and $\varphi(x)$ a scalar-valued function[2] defined for each $x \in X$.

DEFINITION 1. *Weak quasi-convexity.*[3] $\varphi(x)$ is weakly quasi-convex in X if for each $x^1, x^2 \in X$, and $x^0 \in (x^1, x^2)$ holds:

$$\varphi(x^0) \leqq \max \{\varphi(x^1), \varphi(x^2)\}. \tag{1}$$

DEFINITION 2. *Explicit quasi-convexity.*[4] $\varphi(x)$ is explicitly quasi-convex on X if it is weakly quasi-convex, and the strict inequality holds in (1) whenever $\varphi(x^1) \neq \varphi(x^2)$.

DEFINITION 3. *Weak (explicit) quasi-concavity.* $\varphi(x)$ is weakly (explicitly) quasi-concave if $[-\varphi(x)]$ is weakly (explicitly) quasi-convex.

DEFINITION 4. *Weak (explicit) quasi-monotonicity.*[5] $\varphi(x)$ is weakly (explicitly) quasi-monotonic if it is both weakly (explicitly) quasi-convex and quasi-concave.

DEFINITION 5. *Skew quasi-monotonicity.* $\varphi(x)$ is skew quasi-monotonic in the set X if it is weakly quasi-concave in X and is explicitly quasi-convex between any pair of points which do not lie on the same edge of X.[6]

[2] For a vector-valued function $f(x)$ apply these definitions to each of its component.

[3] Called „quasi-convex" by FENCHEL [6], BERGE [3], ARROW-ENTHOVEN [1] and others, „allgemein maximumlos" (i. e., general-maximumless) by DEÁK [4].

[4] Called „functionally convex" by HANSON [7], „streng allgemein maximumlos" (i. e., strictly general-maximumless) by DEÁK [4]. KOVÁCS [9] also used the narrower concept of "strict quasi-convexity" requiring strict inequality to hold in (1) even for $\varphi(x^1) = \varphi(x^2)$.

[5] Called "(streng) allgemein-intern" [i. e. (strict) general-internal] by DEÁK [4]. Weak quasi-monotonicity occurs also in ARROW-HURWICZ-UZAWA [2] without a special name. Our term is justified by the monotonicity property of the corresponding single variable function.

[6] Considering the asymmetry of the last definition, a pair of notions (reflecting each other) should have been introduced. But in this paper, where we transform all programming problems to minimization, the other part of the pair does not occur. This fact enables us to avoid a still longer expression or circumscription.

The mutual connection between convexity, linearity, and the concepts introduced above can be seen by the following propositions which we give without proof:

A. Convexity implies explicit quasi-convexity. Linear functions (including constant functions) are explicitly quasi-monotonic, consequently.

B. $\varphi(x^0) < \max\{\varphi(x^1), \varphi(x^2)\}$ for $\varphi(x^1) \neq \varphi(x^2)$ implies explicit quasi-convexity if $\varphi(x)$ is lower semi-continuous. (For such functions the weak quasi-convexity need not be stipulated in Def. 2.)

C. Weak quasi-convexity implies explicit quasi-convexity if $\varphi(x)$ can be expanded in Taylor series.[7]

D. Skew quasi-monotonicity implies weak quasi-monotonicity if X has no edges or if $\varphi(x)$ is lower semi-continuous and X does not consist merely of a single edge. If X happens to be a single edge, skew quasi-monotonicity is equivalent to weak quasi-concavity.

Instead of the well-known continuity property of convex functions, the following weaker theorem is valid:

E.[8] A weakly quasi-convex function is continuous almost everywhere.

Admissible Sets

Let $g(x) = [\gamma_1(x), \gamma_2(x), \ldots, \gamma_m(x)]$ be a vector valued function of x, defined in the closed convex set X, and b a given m-vector. The set

$$L = \{x \in X | g(x) \leqq b\} \tag{2}$$

will be called admissible set.

THEOREM 1. CONVEX ADMISSIBLE SETS. *The set L is convex for each $b \in E^m$ if and only if $g(x)$ is weakly quasi-convex on X.*

THEOREM 2. POLYHEDRAL ADMISSIBLE SETS. *Let $g(x)$ be a lower semi-continuous, weakly quasi-monotonic function in the polyhedral set $\hat{X}$. Then the admissible set*

$$\hat{L} = \{x \in \hat{X} | g(x) \leqq b\} \tag{3}$$

is a polyhedral set for each $b \in E^m$.

For proving Theorems 1 and 2 we need the following.

LEMMA 1. Let X be convex. The set

$$M = \{x \in X | \gamma(x) \leqq \beta\}$$

is convex for each real β if and only if $\gamma(x)$ is weakly quasi-convex in X.[9] The same statement holds for the set

$$N = \{x \in X | \gamma(x) < \beta\}.$$

[7] Ad libitum differentiability is insufficient, in contrast with MARTOS [10].

[8] DEÁK [4], Theorem 7, p. 120.

[9] This serves also as the most usual definition of weak quasi convexity. In the economic literature it appears often this loose way: "The level surfaces of $\varphi(x)$ should be convex". For a proof similar to ours, see FENCHEL [6], Theorem 50, p. 118.

PROOF OF LEMMA 1.

a) Sufficiency. $x^1, x^2 \in M$, (resp. N) implies $x^1, x^2 \in X$ and $x^0 \in (x^1, x^2)$ implies $x^0 \in X$, by the convexity of X. By the weak quasi-convexity of $\gamma(x)$ and the definition of M (resp. N):

$$\gamma(x^0) \leqq \max \{\gamma(x^1), \gamma(x^2)\} \leqq \beta \text{ (resp. } < \beta).$$

This is: $x^0 \in M$ (resp. N).

b) Necessity. $x^1, x^2 \in X$, $x^0 \in (x^1, x^2)$ implies $x^0 \in X$. If $\max \{\gamma(x^1), \gamma(x^2)\} = +\infty$, then $\gamma(x^0) \leqq \max \{\gamma(x^1), \gamma(x^2)\}$. Consider now the set N in case $\max \{\gamma(x^1), \gamma(x^2)\} < +\infty$, and put $\beta = \max \{\gamma(x^1), \gamma(x^2)\} + \varepsilon$, $\varepsilon > 0$. Thus $x^1, x^2 \in N$, for each $\varepsilon > 0$. By the convexity of N, $x^0 \in N$, for each $\varepsilon > 0$. This is $\gamma(x^0) < \max \{\gamma(x^1), \gamma(x^2)\} + \varepsilon$, for each $\varepsilon > 0$ and $\gamma(x^0) \leqq \max \{\gamma(x^1), \gamma(x^2)\}$, consequently. By putting $\varepsilon = 0$, the same exposition applies to the set M.

PROOF OF THEOREM 1. Both the sufficiency and necessity part of Theorem 1 result immediately from the M-part of Lemma 1 applying it component by component to $g(x)$ and b, and considering that L is the intersection of the convex sets defined this way.

PROOF OF THEOREM 2. $\hat{L}$ is closed by the lower semi-continuity of $g(x)$. Both the set

$$M^j = \{x \in X | \gamma_j(x) \leqq \beta_j\}$$

and the set

$$\overline{N}^j = \{x \in X | \gamma_j(x) > \beta_j\}$$

are convex for each β_j by Lemma 1 because $\gamma_j(x)$ is both weakly quasi-convex and quasi-concave. Moreover M^j is closed. Therefore the sets M^j and $\overline{N}^j$ are separated by a hyperplan whose intersection with $\hat{X}$ belongs to M^j. M^j is thus a closed convex polyhedral set for each β_j and so is $\hat{L} = \bigcap_j M^j$ for each b.

We have not succeeded constructing a converse (necessity) theorem to Theorem 2 as yet.

Local and Global Minima

Let us consider the non-linear programming problem[10]

$$\min \{\varphi(x) | x \in L\} \tag{4}$$

where L is a convex set in E^n.

DEFINITION 6. *Global minimum.* $x^* \in L$ is a global minimum point of $\varphi(x)$ in L, if $\varphi(x^*) \leqq \varphi(x)$ for each $x \in L$. $\varphi(x^*) = -\infty$ is allowed.

DEFINITION 7. *Local minimum.* $\tilde{x} \in L$ is a local minimum point of $\varphi(x)$ in L, if an $\varepsilon > 0$ exists so that $\varphi(\tilde{x}) \leqq \varphi(x)$, whenever $x \in L$ and $|\tilde{x} - x| < \varepsilon$.

[10] In the following part of the paper the sets L, $\hat{L}$, L^A need not be identified with the „admissible sets" as defined in (2) or (3), though in the practice of the non-linear programming they usually are.

THEOREM 3. THEOREM OF GLOBALITY.

a)[11] *If L is convex and $\varphi(x)$ is explicitly quasi-convex in L, then each local minimum point of $\varphi(x)$ is a global minimum point in L.*

b)[12] *If L is convex, $\varphi(x)$ is continuous in L and for all convex subsets K of L is valid, that each local minimum point of $\varphi(x)$ in K is a global minimum point in K, then $\varphi(x)$ is explicitly quasi-convex in L.*

PROOF OF THEOREM 3.

a) If L contains a local minimum point $\tilde{x}$ which fails to be a global minimum point, then there is a point $\bar{x} \in L$ satisfying $\varphi(\bar{x}) < \varphi(\tilde{x})$. By the explicit quasi-convexity of $\varphi(x)$ for each $x^0 \in (\bar{x}, \tilde{x})$ holds $\varphi(x^0) < \varphi(\tilde{x})$. Approaching $\tilde{x}$ by x^0 we can see that $\tilde{x}$ cannot be a local minimum point.

b) If $\varphi(x)$ is not explicitly quasi-convex it cannot be constant by proposition A. Therefore by the continuity of $\varphi(x)$ and by proposition B: $x^1, x^2 \in L$, $x^0 \in (x^1, x^2)$ must exist satisfying:

$$\varphi(x^0) \geqq \varphi(x^1) > \varphi(x^2). \tag{5}$$

If x^1 is a local minimum point in $[x^1, x^0]$ so is it in $[x^1, x^2]$ but it fails to be a global minimum point in the latter. If x^1 is not a local minimum point in $[x^1, x^0]$ then by the continuity of $\varphi(x)$ a point $x^3 \in (x^1, x^0)$ must exist satisfying:

$$\varphi(x^1) > \varphi(x^3) > \varphi(x^2). \tag{6}$$

Consider the set of points $\{y | y \in [x^3, x^0], \varphi(y) = \varphi(x^3)\}$. By the continuity this set contains an element (x^4, say) which is nearest to x^0. Then

$$\varphi(x^0) \geqq \varphi(x^1) > \varphi(x^3) = \varphi(x^4) > \varphi(x^2).$$

Accordingly x^4 is a local minimum point in the segment $[x^4, x^0]$, thus so is in the segment $[x^4, x^2]$, but fails to be a global minimum point in the latter.

Local and Global Vertex-Minima

Let $\hat{L}$ be a polyhedral set with vertices $\hat{x}^1, \hat{x}^2, \ldots, \hat{x}^r$; $R = \{1, 2, \ldots, r\}$ the set of indices (which may be empty) and $R_k \subset R$ a subset of R which consists of k and the indices of vertices which are adjacent to $\hat{x}^k$.

DEFINITION 8. *Global vertex-minimum.* The vertex $\hat{x}^k$ is a global vertex-minimum point of $\varphi(x)$ in the polyhedral set $\hat{L}$, if $\varphi(\hat{x}^k) \leqq \varphi(\hat{x}^i)$ for each $i \in R$.

DEFINITION 9. *Local vertex-minimum.* The vertex $\hat{x}^k$ is a local vertex-minimum point of $\varphi(x)$ in the polyhedral set $\hat{L}$, if $\varphi(\hat{x}^k) \leqq \varphi(\hat{x}^i)$ for each $i \in R_k$.

[11] For L polyhedron and $\varphi(x)$ continuous: MARTOS [10], Th. 2., p. 244. For L closed $\varphi(x)$ continuous, *strictly* quasi-convex: KOVÁCS [9], Lemma 3, p. 215. In this case a local minimum point is also unique. The alleged existence of such a point is valid only if L is also bounded, a condition omitted by KOVÁCS. For L polyhedron and $\varphi(x)$ *weakly* quasi-convex (erroneously): HOÁNG TUY [8]. Th. II, p. 214. (Owing to a—supposed—misprint in HOÁNG TUY's definition M_2 we cannot be quite sure what this theorem contains. My interpretation is — I believe — well meaning.)

[12] For L and K polyhedron: MARTOS [10]. Th. 2, p. 244.

First we consider theorems concerning a function $\varphi(x)$ defined in a *polyhedron* L^A, and the problem

$$\min \{\varphi(x) | x \in L^A\}.$$

THEOREM 4. THEOREM OF GLOBAL VERTEX-MINIMA.

a)[13] *If $\varphi(x)$ is a weakly quasi-concave function on the polyhedron L^A, then each global vertex-minimum point of $\varphi(x)$ in L^A is a global minimum point in L^A.*

b)[14] *If L is a convex set and for each polyhedron $L^A \subset L$ is valid that each global vertex-minimum point of $\varphi(x)$ in L^A is global minimum point in L^A, then $\varphi(x)$ is weakly quasi-concave in L.*

PROOF OF THEOREM 4.

a) Considering that each point of a polyhedron is a convex linear combination of the vertices and applying Definition 1. successively, for each $x \in L^A$ results $\varphi(x) \geqq \geqq \min \{\varphi(\hat{x}^i) | i \in R\}$.

b) Let $x^1, x^2 \in L$, and $L^A = [x^1, x^2]$. By our assumption $\varphi(x^0) \geqq \geqq \min \{\varphi(x^1), \varphi(x^2)\}$ for each $x^0 \in (x^1, x^2)$ which is just the definition of a weakly quasi-concave function in L.

From Theorem 4. results:

COROLLARY 4. *If $\varphi(x)$ is weakly quasi-concave in the polyhedron L^A then it assumes its minimum on a vertex of L^A.*

THEOREM 5. THEOREM OF LOCAL VERTEX MINIMA.

a)[15] *If $\varphi(x)$ is skew quasi-monotonic in the polyhedron L^A, then each local vertex minimum point of $\varphi(x)$ in L^A is a global minimum point in L^A.*

b)[16] *If L is convex, $\varphi(x)$ a continuous function in L and for all polyhedron $L^A \subset L$ is valid, that each local vertex minimum point of $\varphi(x)$ in L^A is a global minimum point in L^A, then $\varphi(x)$ is skew quasi-monotonic in L.*

PROOF OF THEOREM 5.

a) Assume that the vertex $\hat{x}^k$ is a local vertex-minimum point in L^A, i.e.:

$$\varphi(\hat{x}^k) = \min \{\varphi(\hat{x}^i) | i \in R_k\} \tag{7}$$

and that $\hat{x}^k$ fails to be global minimum point on L^A, i.e. there exists an $x^2 \in L^A$ satisfying

$$\varphi(x^2) < \varphi(\hat{x}^k). \tag{8}$$

Let L_k^A denote the convex polyhedron spanned by the vertices with index from R_k. If $x^2 \in L_k^A$, then by the weak quasi-concavity of $\varphi(x)$: $\varphi(x^2) \geqq \min \{\varphi(\hat{x}^i) | i \in R_k\}$ in contrast with (7) and (8). If $x^2 \notin L_k^A$, then $\hat{x}^k$ and x^2 cannot be points of the same edge of L^A. Consider a point $x^0 \in L_k^A \cap (\hat{x}^k, x^2)$. Then $\varphi(x^0) \geqq \varphi(\hat{x}^k)$ by applying Corr. 4. to L_k^A, considering that $x^0 \in L_k^A$ and $\varphi(x)$ is weakly quasi-concave. Simulta-

[13] HOÁNG TUY [8] Th. I, p. 214. For $\varphi(x)$ continuous: MARTOS [10], Th. 1, p. 244.

[14] For L polyhedron and $\varphi(x)$ continuous: MARTOS [10] Th. 1, p. 244.

[15] For $\varphi(x)$ continuous, weakly quasi-concave and explicitly quasi-convex included in MARTOS [10]. Corollary 2, p. 244. For $\varphi(x)$ weakly quasi-monotonic (erroneously): HOÁNG TUY [8], p. 214.

[16] For L polyhedron, $\varphi(x)$ weakly quasi-concave and explicitly quasi-convex included (erroneously) in MARTOS [10], Corr. 2, p. 244.

neously $\varphi(x^0) < \varphi(\hat{x}^k)$ because $x^0 \in (\hat{x}^k, x^2)$ and $\varphi(x)$ is explicitly quasi-convex in $[\hat{x}^k, x^2]$. This contradiction proves a).

b) The weak quasi-concavity of $\varphi(x)$ comes simply from applying the "local vertex minimum is global" assumption to $L^\Delta = [x^1, x^2]$, where x^1, x^2 are arbitrarily chosen from L. This property of $\varphi(x)$ finds application in the remaining part of the proof. Suppose now that there is a segment $[x^1, x^2] \subset L$, satisfying:

α) $\varphi(x^1) > \varphi(x^2)$

β) x^1 and x^2 do not lie on the same edge of L

γ) $\varphi(x)$ is not explicitly quasi-convex along this segment.

We are going to prove the theorem by constructing a convex, planar quadrangle Q, in which a local vertex minimum point fails to be global minimum point.

For the sake of brevity, let us use the notation $\varphi_i = \varphi(x^i)$, and $Q = Q[\tilde{\alpha}, \beta, \gamma, \delta]$ for the quadrangle with vertices: x^α, x^β, x^γ, x^δ, $\sim$ referring to a local vertex minimum point which is not global.

Let E be the straight line through x^1, x^2 and S a plane containing E and such that the *open* halfplanes T and V, which are separated by E, shall contain a point of L, each. The existence of such a plane results from β). All of what follows takes place in this plane. From α) and γ) comes the existence of a point $x^0 \in (x^1, x^2)$ satisfying:

$$\varphi_0 \geqq \varphi_1 > \varphi_2. \tag{9}$$

But then for each $\bar{x} \in [x^1, x^2]$ holds

$$\varphi(\bar{x}) \geqq \varphi_1 \tag{10}$$

by the weak quasi-concavity of $\varphi(x)$.

Case 1. If there is an $\bar{x} (= x^0$, maybe) for which the strict inequality holds in (10), then by the continuity of $\varphi(x)$ its neighbourhood contains a pair of points: x^3, x^4 with the following properties: $x^3 \in T \cap L$, $x^4 \in V \cap L$, $\min \{\varphi_3, \varphi_4\} \geqq \varphi_1$ and the quadrangle $Q = Q[\tilde{1}, 3, 2, 4]$ is convex.

Case 2. Consider the opposite case:

$$\varphi_1 = \varphi(\bar{x}) = \varphi_0 \quad \text{for all} \quad \bar{x} \text{ in } (x^1, x^0). \tag{11}$$

Choose a point $\bar{\bar{x}} \in (x^1, x^0)$ arbitrarily, and a pair of points x^3, x^4 such that: $[x^3, x^4]$ should be perpendicular to E in $\bar{\bar{x}}$, $x^3 \in T \cap L$, $x^4 \in V \cap L$, and

$$\varphi_3 \geqq \varphi_4 > \varphi_2. \tag{12}$$

(The left side inequality can be supposed by the symmetry, the right side results from $\varphi(\bar{\bar{x}}) = \varphi_0 > \varphi_2$ and the continuity of $\varphi(x)$ if we chose x^3, x^4 close enough to $\bar{\bar{x}}$.) From (12) by the quasi-concavity of $\varphi(x)$ along $[x^3, x^4]$:

$$\varphi(\bar{\bar{x}}) = \varphi_0 \geqq \varphi_4. \tag{13}$$

Subcase 21. If $\varphi_0 = \varphi_4$, then $Q = Q[\tilde{1}, 3, 2, 4]$.

Subcase 22. In case $\varphi_0 > \varphi_4$ the following sub-subcases are mutually excluding and collectively exhausting.

22^1. $\varphi_0 \geqq \varphi_3 > \varphi_4$. Then $Q = Q[\tilde{1}, 3, 0, 4]$.

22^2. $\varphi_0 > \varphi_3 = \varphi_4$. Approach x^3 to $\bar{\bar{x}}$ until reaching a point with a value greater than φ_4, but still not greater than φ_0. This is case 22^1.

22^3. $\varphi_3 > \varphi_0 > \varphi_4$. By continuity we have a point $x^5 \in (x^3, x^2)$ with $\varphi_5 = \varphi_0$. We should have $Q = Q[1, \tilde{5}, 0, 4]$, but it may or may not be convex. If it is not we can approach x^4 to $\bar{x}$, either prevailing $\varphi_0 > \varphi_4$, until Q becomes convex, or reaching a point with $\varphi_0 = \varphi_4$, when Subcase 21 applies.

Thereby the proof is finished.

In Theorem 4 and 5 $\varphi(x)$ is defined in a (bounded) polyhedron. The sufficiency part of these theorems however, may be extended to (not necessarily bounded) polyhedral sets depending on the following:

LEMMA 2. If $\hat{L}$ is a closed convex polyhedral set, L^A is the (non empty) polyhedron spanned by the vertices of $\hat{L}$, $\varphi(x)$ is explicitly quasi-concave in $\hat{L}$ and assumes its minimum in $\hat{L}$ then it assumes the same minimum in L^A.

PROOF OF LEMMA 2. Employing the statement a) of Theorem 4. let us suppose that $\hat{x}^k$ is one of those vertices of L^A satisfying

$$\varphi(\hat{x}^k) = \min \{\varphi(x) | x \in L^A\}. \tag{14}$$

Assume that there exists $x^0 \in \hat{L}$ so that

$$\varphi(x^0) < \varphi(\hat{x}^k). \tag{15}$$

We are going to prove that x^0 cannot be a global minimum point in $\hat{L}$. Obviously $x^0 \notin L^A$, consequently it cannot be a vertex.

Let us now choose a point $x^1 \in L^A$ so that if x^0 lies in the interior of $\hat{L}$ then x^1 be any point of L^A, but if x^0 is a boundary point, then x^1 be such a point of L^A that all those bounding hyperplanes of $\hat{L}$ which contain x^0 should contain x^1, too. This choice enables us to find a third point x^2 such that $[x^1, x^2] \subset \hat{L}$, $x^1 \in L^A$ and $x^0 \in (x^1, x^2)$.

$\varphi(x^1) \leqq \varphi(x^2)$ is impossible, because by weak quasi-concavity in $[x^1, x^2]$ and by (14) $\varphi(x^0) \geqq \varphi(x^1) \geqq \varphi(\hat{x}^k)$ would result in contrast with (15). Thus $\varphi(x^1) > \varphi(x^2)$. By the explicit quasi-concavity $\varphi(x^0) > \varphi(x^2)$, accordingly x^0 fails to be a global minimum point on $\hat{L}$. Summed up: if the global minimum of $\varphi(x)$ on L^A fails to be the same in $\hat{L}$, then $\varphi(x)$ does not assume its minimum on $\hat{L}$, what is just what the lemma alleges.

From Lemma 2 immediately results:

COROLLARY 4—5. *The statement* a) *of Theorems* 4 *and* 5 *remains valid even if we replace the polyhedron L^A with a polyhedral set $\hat{L}$ supposed that $\varphi(x)$ is explicitly (rather than weakly) quasi-concave and that $\varphi(x)$ assumes its minimum in $\hat{L}$.*[17]

The Set of Optimum Points

Let us consider the problem $\min \{\varphi(x) | x \in L\}$, where L is supposed to be convex. Let L^* be the set of the optimum points of the problem and $\tilde{L} = L - L^*$ the complement of L^*. Whether L^* or $\tilde{L}$ is allowed to be empty.

[17] In case $\hat{L}$ has no vertices the referred statements become meaningless. This case is, however, excluded as a rule, $\hat{L}$ being defined as in (3) and $\hat{X}$ identified with the non-negative orthant of E^n.

THEOREM 6.[18] THE SET OF OPTIMUM POINTS. *If $\varphi(x)$ is weakly quasi-convex (resp. quasi-concave) in the convex set L, then the set L^* of the optimum points (resp. the complementary set $\tilde{L}=L-L^*$) is convex, maybe empty.*

* * *

A practical importance may also be attached to the above results. They enable us to extend and in some cases delimit the power of some well-known programming techniques. We published a fair deal of such results in [11].

REFERENCES

[1] ARROW, K. J., ENTHOVEN, A. C.: Quasi-concave Programming, *Econometrica* **29** (1961) 779—800.

[2] ARROW, K. J., HURWICZ, L., UZAWA, H.: Constraint Qualifications in Maximization Problems, *Naval Res. Logist. Quart.* **8** (1961) 175—191.

[3] BERGE, C.: *Topological Spaces*, Oliver & Boyd, Edinburgh, 1963.

[4] DEÁK, E.: Über konvexe und interne Funktionen, sowie eine gemeinsame Verallgemeinerung von beiden. *Ann. Univ. Sci. Budapest. Sect. Math.* **5** (1962) 109—154.

[5] DEFINETTI, B.: Sulla stratificazioni convesse, *Ann. Mat. Pura Appl.* **30** (1949) 173—183.

[6] FENCHEL, W.: *Convex Cones, Sets and Functions*, Princeton Univ. 1953.

[7] HANSON, M. A.: Duality and Self-Duality in Mathematical Programming, *SIAM J. Appl. Math.* **12** (1964) 446—449.

[8] HOÁNG TUY: Sur une classe des programmes non linéaires, *Bull Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.* **12** (1964) 213—215.

[9] KOVÁCS, L. B.: Gradient Projection Method for Quasi-Concave Programming, *Colloqu. on Appl. of Math. to Economics, Budapest, 1963.* Akadémiai Kiadó, Budapest, 1965.

[10] MARTOS, B.: The Direct Power of Adjacent Vertex Programming Methods, *Management Sci.* **12** (1965) 241—252.

[11] MARTOS, B.: Nem-lineáris programozási módszerek hatóköre (The Power of Non-linear Programming Methods), *A Magyar Tudományos Akadémia Közgazdaságtudományi Intézetének Közleményei*, No. 20. Budapest (1966).

INSTITUTE OF ECONOMICS OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received February 18, 1966)

(Revised December 20, 1966)

[18] For L polyhedron: MARTOS [10] Lemma 1. and 2., p. 249. The proof remains unchanged. HOÁNG TUY [8] Th. I., p. 214 erroneously alleges that L^* is convex if $\varphi(x)$ is weakly quasi-concave (instead of quasi-convex). For a counterexample see: MARTOS [11], Example 10. 8, p. 58.

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 275—283.

ANWENDUNG DER HYPERMATRIZEN FÜR DIE UNTERSUCHUNG EINES WIDERSTANDNETZES

von

Z. PERJÉS

1. Einführung. Für die Behandlung der räumlichen Systeme periodischer Struktur erwies sich die Anwendung der Hypermatrizen als ein vorteilhaftes Mittel. Zum Beispiel sind diese für die Untersuchung der Kristallgitter in der Festkörperphysik sehr geeignet [1]. In dieser Arbeit wird gezeigt, daß mit Hilfe von Hypermatrizen auch der resultierende Widerstand periodischer Widerstandssysteme bestimmt werden kann.

Wir werden unseren Gedankengang am Beispiel eines unendlich ebenen *Q*uadratgitternetzes vorführen. Jedes Element des Netzes hat den Widerstand r (Abb. 1). Binden wir zur Untersuchung der Eigenschaften des Netzes eine Spannungsquelle an die Endpunkte eines Elements. Wir werden ausrechnen, wie großes Potential sich in irgendeinem Endpunkt des beliebigen Elements ausbilden wird. In Kenntnis dieser Potentiale wollen wir die Größe des Widerstandes bestimmen, den man zwischen den Endpunkten eines Netzelementes messen kann.

Zu diesem Zwecke nehmen wir zuerst ein aus endlich vielen Elementen bestehendes Quadratgitternetz, und zwar ein solches, in dem die Elemente in n-zahligen Reihen und Spalten der gleichen Zahl angeordnet sind. Es vereinfacht die Verhandlung, wenn man das unendlich ebene Quadratgitternetz als den Grenzfall eines solchen Netzes betrachtet, das sich an einer Torusfläche befindet. Strebt der Parallel- und Meridiankreis des Torus dem Unendlichen zu, und strebt inzwischen auch n zum Unendlichen, dann erreichen wir im Grenzfall ein unendliches Flachgitter. Wir werden also so vorgehen, daß wir, die erste Reihe des Flachgitters mit der n-ten koppelnd, zu einem Zylinder kommen, dann die erste Spalte mit der n-ten koppelnd einen Torus erreichen. Somit wird das System in ein zyklisches umgeändert werden.

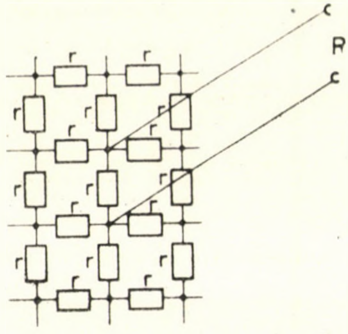


Abb. 1

Dann werden wir mit Hilfe der Kirchhoffschen Gesetze die Gleichungen anschreiben, die einen Zusammenhang unter den Spannungen in den einzelnen Knotenpunkten des Netzes geben. Das so erhaltene Gleichungssystem wird mit Anwendung der Algebra von Hypermatrizen gelöst. Das Gleichungssystem wird ergänzt, um die Symmetrie des so erhaltenen Systems ausnützend, die Potentiale bestimmen zu können. Zerlegen wir das System auf geeignete Weise, so bekommen wir für die Potentiale ein inhomogenes lineares Gleich-

ungssystem. Die Inverse der Koeffizientenmatrix dieses Gleichungssystems wird aus der ursprünglichen Koeffizientenmatrix berechnet.

Es ist naheliegend, daß auch andere periodische Widerstandsstrukturen auf ähnliche Weise zu behandeln sind.

2. Zusammenfassung der benützten Sätze. Setzt man beliebige Matrizen (Blöcke) an Stelle der Elemente einer aus skalaren Elementen bestehenden Matrix, erhält man eine Hypermatrix. Es ist bekannt, daß das direkte Produkt der Matrizen $\mathbf{A}$ und $\mathbf{B}=[b_{jk}]$ die Hypermatrix

$$\mathbf{A}\cdot\times\mathbf{B}=[\mathbf{A}b_{jk}]$$

ist. Die Spektralzerlegung der Hypermatrizen, die man als direktes Produkt von symmetrischen Matrizen erzeugen kann, wird auf Grund des Satzes von Egerváry [2] durchgeführt. Hier werden wir diesen Satz in der folgenden Abfassung (von verengter Gültigkeit) verwenden:

SATZ A. *Sind* $\mathbf{A}$ *und* $\mathbf{B}$ *symmetrische Matrizen, dann sind die Eigenwerte von* $\mathbf{A}\cdot\times\mathbf{B}$ *die Produkte* a_jb_k *der Eigenwerte von* $\mathbf{A}$ *und* $\mathbf{B}$ *und die Eigenvektoren von* $\mathbf{A}\cdot\times\mathbf{B}$ *sind die direkten Produkte* $\mathbf{u}_j\cdot\times\mathbf{v}_k$ *der Eigenvektoren von* $\mathbf{A}$ *und* $\mathbf{B}$.

SATZ B. *Es seien die Blöcke* $\mathbf{A}$ *und* $\mathbf{D}$ *der Hypermatrix*

$$\begin{array}{|c|c|}\hline \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \\ \hline \end{array}$$

quadratische Matrizen beliebiger Ordnung. Wenn diese Hypermatrix nicht-singulär ist, und in aufgeteilter Form folgend aufgeschrieben wird

$$\begin{array}{|c|c|}\hline \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \\ \hline \end{array}^{-1}=\begin{array}{|c|c|}\hline \mathbf{X} & \mathbf{Y} \\ \hline \mathbf{Z} & \mathbf{W} \\ \hline \end{array}$$

weiterhin der Block $\mathbf{X}$ *umkehrbar ist, so existiert auch* $\mathbf{D}^{-1}$ *und kann, wie folgt, ausgedrückt werden:*

$$\mathbf{D}^{-1}=\mathbf{W}-\mathbf{Z}\mathbf{X}^{-1}\mathbf{Y}.$$

Dieser Satz war in einer etwas anderen Abfassung von EGERVÁRY, RÓZSA und SIEBER erkannt (Siehe z. B. [3]).

BEWEIS. Nach der Definition der inversen Matrix gilt die Beziehung:

$$\begin{array}{|c|c|}\hline \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \\ \hline \end{array}\begin{array}{|c|c|}\hline \mathbf{X} & \mathbf{Y} \\ \hline \mathbf{Z} & \mathbf{W} \\ \hline \end{array}=\begin{array}{|c|c|}\hline \mathbf{E} & \mathbf{O} \\ \hline \mathbf{O} & \mathbf{E} \\ \hline \end{array}$$

Wählen wir die Gleichungen aus, die sich auf die untenstehenden Blöcke der Einheitsmatrix beziehen:

$$\mathbf{C}\mathbf{X}+\mathbf{D}\mathbf{Z}=\mathbf{O},$$

$$\mathbf{C}\mathbf{Y}+\mathbf{D}\mathbf{W}=\mathbf{E}.$$

Aus der ersten Gleichung kann man $\mathbf{C}$ ausdrücken, wenn man mit $\mathbf{X}^{-1}$ von rechts multipliziert:

$$\mathbf{C}=-\mathbf{D}\mathbf{Z}\mathbf{X}^{-1}.$$

Setzt man diesen Ausdruck in die untere Gleichung ein, ergibt sich:

$$\mathbf{D}(-\mathbf{ZX}^{-1}\mathbf{Y}+\mathbf{W}) = \mathbf{E},$$

woraus der zu beweisende Zusammenhang unmittelbar folgt.

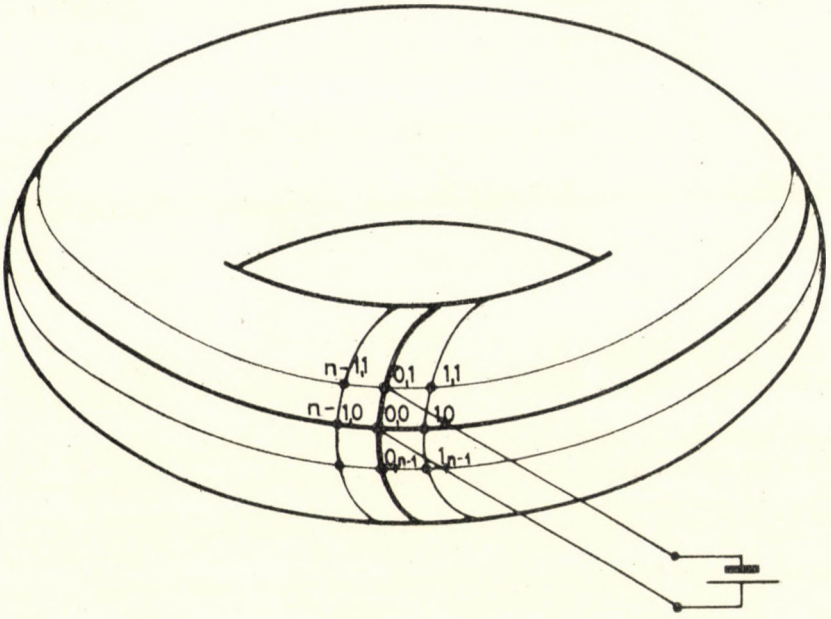


Abb. 2

3. Das Potential in den Netzpunkten. Wir bezeichnen das Potential der Netzpunkte, wie in der Abb. 2 gezeigt wird. Wir binden die zwei Pole einer Stromquelle von gegebener Spannung an die (0, 0) und (0, 1) Punkte. So sind U_{00} und U_{01} bekannt und zwar

$$U_{00} = 0,$$
$$U_{01} = U.$$

Am Torus sind $n \cdot n$ Netzpunkte, also die Anzahl der unbekannten Spannungswerte ist $n^2 - 2$, sie sind also durch gleich so viel, $n^2 - 2$ Gleichungen zu bestimmen. Den Wert von U_{jk} ($(j, k) \neq (0, 0)$ und $(0, 1)$) können wir aus den Spannungswerten in den Nachbarpunkten mit Hilfe der Kirchhoffschen Sätze ermitteln (Abb. 3). Da die Summe der Ströme in Punkt (j, k) gleich Null ist:

$$\sum_{\nu=1}^{4} I_\nu = 0,$$

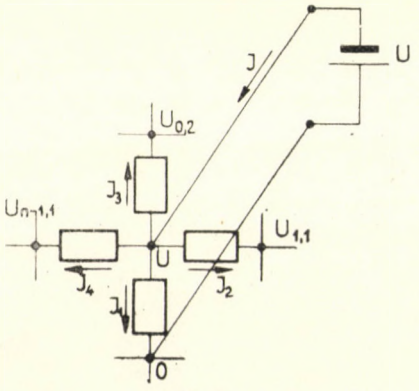


Abb. 3

hieraus folgt

$$\frac{U_{j,k}-U_{j-1,k}}{r}+\frac{U_{j,k}-U_{j+1,k}}{r}+\frac{U_{j,k}-U_{j,k-1}}{r}+\frac{U_{j,k}-U_{j,k+1}}{r}=0.$$

Es ist daher

$$U_{j,k}=\frac{U_{j-1,k}+U_{j+1,k}+U_{j,k-1}+U_{j,k+1}}{4},$$

oder

$$4U_{j,k}-U_{j-1,k}-U_{j+1,k}-U_{j,k-1}-U_{j,k+1}=0.$$

Fügen wir die zu den Knotenpunkten (0, 0) und (0, 1) gehörenden Gleichungen:

$$4U_{0,0}-U_{0,1}-U_{1,0}-U_{n-1,0}-U_{0,n-1}=0,$$

$$4U_{0,1}-U_{0,0}-U_{0,2}-U_{1,1}-U_{n-1,1}=U$$

dem Gleichungssystem hinzu. Wir erhalten das System

$$\mathbf{AX}'=\mathbf{b},$$

wo

$$\mathbf{b}=\begin{bmatrix}0\\U\\0\\0\\\vdots\\0\end{bmatrix};\qquad \mathbf{X}'=\begin{bmatrix}U_{0,0}\\U_{0,1}\\\cdots\\U_{0,n-1}\\\hline U_{1,0}\\\cdots\\U_{1,n-1}\\\hline - - -\\- - -\\- - -\\\hline U_{n-1,0}\\\cdots\\U_{n-1,n-1}\end{bmatrix}.$$

$\mathbf{X}'$ besteht nicht aus lauter Unbekannten, weil die ersten zwei Elemente 0 bzw. U sind. Diese Schreibweise hat den Vorteil, daß man die Matrix $\mathbf{A}$ als das direkte Polynom der wohlbekannten zyklischen Matrix n-ter Ordnung

$$\mathbf{\Omega}=\begin{bmatrix}&1&&&\\&&1&&\\&&&\ddots&\\&&&&1\\1&&&&\end{bmatrix}$$

aufschreiben kann:

$$\mathbf{A} = \mathbf{E} \cdot \times \mathbf{K} + \mathbf{K} \cdot \times \mathbf{E}; \qquad \mathbf{K} = 2\mathbf{E} - \boldsymbol{\Omega} - \boldsymbol{\Omega}^{-1}.$$

Die Spektralzerlegung von $\boldsymbol{\Omega}$ läßt sich nämlich in geschlossener Form anschreiben:

$$\boldsymbol{\Omega} = \mathbf{U}\boldsymbol{\Lambda}\overline{\mathbf{U}}^{*}, \qquad \mathbf{U}^{-1} = \overline{\mathbf{U}}^{*}$$

$$\boldsymbol{\Lambda} = \langle \lambda_k \rangle = \langle e^{\frac{2\pi i}{n} k} \rangle; \qquad k = 0, 1, \ldots, n-1,$$

$$\mathbf{U} = [u_{jk}] = \left[\frac{1}{\sqrt{n}} e^{\frac{2\pi i}{n} j \cdot k} \right]; \qquad j = 0, 1, \ldots, n-1.$$

So ist

$$K = \mathbf{U}(2\mathbf{E} - \boldsymbol{\Lambda} - \boldsymbol{\Lambda}^{-1})\overline{\mathbf{U}}^{*},$$

und die Eigenwerte von $\mathbf{K}$ sind:

$$2 - e^{\frac{2\pi i}{n} k} - e^{-\frac{2\pi i}{n} k} = 4 \sin^2 \frac{\pi}{n} k, \qquad k = 0, 1, \ldots, n-1.$$

Wenden wir jetzt den Satz A an, so können wir die Elemente von $\mathbf{A}$ aufschreiben:

$$\mathbf{A}_{(j,k),\,(l,m)} = (\mathbf{K} \cdot \times \mathbf{E} + \mathbf{E} \cdot \times \mathbf{K})_{(j,k),\,(l,m)} = \sum_{r=0}^{n-1} \sum_{s=0}^{n-1} u_{jr} u_{ks} \left(4 \sin^2 \frac{r\pi}{n} + 4 \sin^2 \frac{s\pi}{n} \right) \bar{u}_{lr} \bar{u}_{ms}.$$

Zerlegen wir jetzt das Gleichungssystem folgenderweise:

$$\begin{array}{|c|c|}\hline & \mathbf{V}^{*} \\ \hline \mathbf{V} & \mathbf{B} \\ \hline \end{array} \begin{array}{|c|}\hline 0 \\ U \\ \hline \mathbf{x} \\ \hline \end{array} = \begin{array}{|c|}\hline 0 \\ U \\ \hline 0 \\ \hline \end{array};$$

wir erhalten:

$$\begin{array}{|c|}\hline \\ \\ \hline \end{array} \begin{array}{|c|}\hline 0 \\ U \\ \hline \end{array} + \mathbf{V}^{*}\mathbf{x} = \begin{array}{|c|}\hline 0 \\ U \\ \hline \end{array},$$

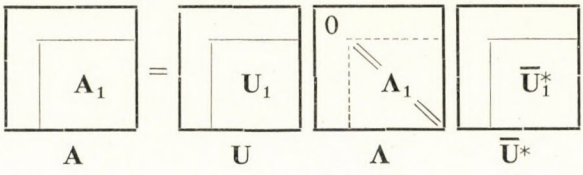
$$\mathbf{B} \cdot \mathbf{x} = -\mathbf{V} \begin{array}{|c|}\hline 0 \\ U \\ \hline \end{array}; \qquad -\mathbf{V} \begin{array}{|c|}\hline 0 \\ U \\ \hline \end{array} = \mathbf{c}.$$

Aus diesen zwei Systemen brauchen wir nur das Zweite zur Bestimmung der Un-

bekannten. Aus diesem folgt:

$$\text{(1)} \qquad \mathbf{x} = \mathbf{B}^{-1}\mathbf{c}; \qquad \mathbf{c} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \hline 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ \hline -- \\ -- \\ \hline 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \cdot U. \qquad \begin{matrix} \\ \\ \\ \\ \\ \leftarrow (1,1) \\ \\ \\ \\ \\ \\ \\ \leftarrow (n-1,1) \\ \\ \\ \end{matrix} \quad \leftarrow (0,2)$$

Hier dürften wir den Satz B anwenden, wenn $\mathbf{A}$ eine Inverse hätte. Aber $\mathbf{A}$ ist eine singuläre Matrix, deswegen werden wir die gesuchte Inverse $\mathbf{B}^{-1}$ in zwei Teilen bestimmen. Durch das Weglassen der ersten Spalte und Zeile von $\mathbf{A}$ bekommen wir $\mathbf{A}_1$. Zuerst bestimmen wir die Inverse von $\mathbf{A}_1$, indem wir die bekannte Spektralzerlegung der Matrix $\mathbf{A}$ benützen. Mit Hilfe der Spektralzerlegung der Matrix $\mathbf{A}$ erhalten wir:



Also:

$$\mathbf{A}_1^{-1} = (\overline{\mathbf{U}}_1^*)^{-1}\mathbf{\Lambda}_1^{-1}\mathbf{U}_1^{-1}.$$

$\mathbf{U}_1^{-1}$ und $(\overline{\mathbf{U}}_1^*)^{-1}$ lassen sich mit Hilfe des Satzes B ausrechnen:

$$\mathbf{U}^{-1} = \overline{\mathbf{U}}^* - \frac{1}{u_{11}}\mathbf{v}_1\mathbf{u}_1^*,$$

$$\overline{\mathbf{U}}^{*-1} = \mathbf{U} - \frac{1}{u_{11}}\mathbf{u}_1\mathbf{v}_1^*.$$

Hier haben wir mit $\mathbf{v}_1^*$ die erste Zeile, mit $\mathbf{u}_1$ die erste Spalte von $\mathbf{U}$ bezeichnet und

$$\mathbf{u}_1 = \mathbf{v}_1.$$

Aus der Inversen von $\mathbf{A}_1$ können wir die Inverse von $\mathbf{B}$ nunmehr ausrechnen:

$$\begin{array}{|c|c|}\hline 0 & 0 \\ \hline 0 & \mathbf{B}^{-1} \\ \hline \end{array} = \underbrace{\begin{array}{|c|c|}\hline \gamma & \alpha^* \\ \hline \beta & \\ \hline \end{array}^{-1}}_{\mathbf{A}_1^{-1}} - \begin{array}{|c|}\hline 0 \\ \hline \beta \\ \hline \end{array} \begin{array}{|c|}\hline \gamma \\ \hline \end{array}^{-1} \begin{array}{|c|c|}\hline 0 & \alpha^* \\ \hline \end{array},$$

$$\begin{array}{|c|}\hline \beta \\ \hline \end{array} = \left(\mathbf{U} - \frac{1}{u_{11}}\mathbf{u}_1\mathbf{v}_1^*\right)\mathbf{\Lambda}_1^{-1}\left(\bar{\mathbf{v}}_2 - \frac{u_{21}}{u_{11}}\mathbf{v}_1\right),$$

$$\begin{array}{|c|}\hline \alpha^* \\ \hline \end{array} = \left(\mathbf{v}_2^* - \frac{u_{21}}{u_{11}}\mathbf{v}_1^*\right)\mathbf{\Lambda}_1^{-1}\left(\bar{\mathbf{U}}^* - \frac{1}{u_{11}}\mathbf{v}_1\mathbf{u}_1^*\right),$$

$$\begin{array}{|c|}\hline \gamma \\ \hline \end{array} = \left(\mathbf{v}_2^* - \frac{u_{21}}{u_{11}}\mathbf{v}_1^*\right)\mathbf{\Lambda}_1^{-1}\left(\bar{\mathbf{v}}_2 - \frac{u_{21}}{u_{11}}\mathbf{v}_1\right).$$

Durch $\mathbf{v}_1^*$ wird die erste, durch $\mathbf{v}_2^*$ die zweite Zeile in $\mathbf{U}$ bezeichnet. Beachten wir, daß $\frac{1}{u_{11}}\mathbf{u}_1^* = (1, 1, \ldots, 1) = \mathbf{e}^*$ ist, dann bekommen wir das folgende Ergebnis für die Inverse der Matrix $\mathbf{B}$:

$$(2) \qquad \mathbf{B}^{-1} = (\mathbf{U} - \mathbf{e}\mathbf{v}_1^*)\left\{\mathbf{\Lambda}_1^{-1} - \frac{\mathbf{\Lambda}_1^{-1}(\mathbf{v}_2 - \mathbf{v}_1)\cdot(\mathbf{v}_2^* - \mathbf{v}_1^*)\mathbf{\Lambda}_1^{-1}}{(\mathbf{v}_2^* - \mathbf{v}_1^*)\mathbf{\Lambda}_1^{-1}(\bar{\mathbf{v}}_2 - \bar{\mathbf{v}}_1)}\right\}(\bar{\mathbf{U}}^* - \mathbf{v}_1\mathbf{e}^*).$$

Substituiert man die Matrizen $\mathbf{U}$, $\mathbf{\Lambda}_1^{-1}$ usw. in (2), dann ergibt sich für das (j, j'); (k, k')-te Element der Inversen:

$$[\mathbf{B}^{-1}]_{(j,j'),(k,k')} = \sum_{\substack{r \\ r' \\ s \\ s'}}{}' \frac{1}{n^2}\left\{e^{\frac{2\pi i}{n}(jr+j'r')} - 1\right\} a_{(r,r'),(s,s')}\left\{e^{-\frac{2\pi i}{n}(sk+s'k')} - 1\right\};$$

wo

$$a_{(r,r'),(s,s')} = \frac{\delta_{(s,s'),(r,r')}}{N_n(r,r')} - \frac{1}{c_n}\frac{\left[e^{-\frac{2\pi i}{n}(r+r')} - 1\right]\left[e^{\frac{2\pi i}{n}(s+s')} - 1\right]}{N_n(r,r')\cdot N_n(s,s')};$$

$$N_n(r,r') = 4\left(\sin^2\frac{r\pi}{n} + \sin^2\frac{r'\pi}{n}\right),$$

$$\delta_{(s,s'),(r,r')} = \delta_{sr}\cdot\delta_{s'r'}, \qquad c_n = \sum_{y,z=0}^{n-1}{}' \frac{\sin^2\frac{y+z}{n}\pi}{\sin^2\frac{y\pi}{n} + \sin^2\frac{z\pi}{n}}.$$

Das Zeichen $\sum'$ zeigt, daß das Glied mit dem Index (0, 0, 0, 0) aus der Summe wegzulassen ist.

Mit Hilfe der Matrix $\mathbf{B}^{-1}$ und der Gleichung (2) schreiben wir die Potentiale auf:

$$\mathbf{x} = \left[U \cdot \frac{1}{n^2} \sum_{\substack{r \\ r' \\ s \\ s'}}^{n-1\,\prime} \left[e^{\frac{2\pi i}{n}(jr+j'r')} - 1 \right] a_{(r,r'),\,(s,s')} \left[e^{-\frac{4\pi i}{n}s'} + e^{-\frac{2\pi i}{n}(s+s')} + e^{\frac{2\pi i}{n}(s-s')} - 3 \right] \right]_{=0}.$$

4. Bestimmung des resultierenden Widerstandes. Der im vorigen Punkt für $U_{(j,k)}$ gewonnene Ausdruck wird verwendet, um den resultierenden Widerstand zwischen den beiden Endpunkten eines Elements auszurechnen. Binden wir die Stromquelle eben an dieses Element. Dann ist der resultierende Widerstand (Abb. 3):

$$R = \frac{U}{I}.$$

I ist der durch die Stromquelle fließende Strom. Der verzweigt sich im Punkt (0, 1) in vier Teile:

$$I = \sum_{\nu=1}^{4} I_\nu;$$

$$I_1 = \frac{U}{r}; \quad I_2 = \frac{U - U_{1,1}}{r}; \quad I_3 = \frac{U - U_{0,2}}{r}; \quad I_4 = \frac{U - U_{n-1,1}}{r}.$$

Der resultierende Widerstand ist also:

$$R = \frac{1}{4 - \dfrac{U_{1,1} + U_{0,2} + U_{n-1,1}}{U}} r.$$

Wenden wir den Ausdruck für die Potentiale an:

$$\frac{U_{1,1} + U_{0,2} + U_{n-1,1}}{U} = \frac{1}{n^2} \sum_{\substack{r \\ r' \\ s \\ s'}}^{n-1\,\prime} \left(e^{\frac{4\pi i}{n}r'} + 2e^{\frac{2\pi i}{n}r'} \cos\frac{2\pi}{n} r - 3 \right) \cdot$$

$$\cdot\, a_{(r,r'),\,(s,s')} \left(e^{-\frac{4\pi i}{n}s'} + 2e^{-\frac{2\pi i}{n}s'} \cos\frac{2\pi}{n} s - 3 \right).$$

Bilden wir jetzt den Grenzübergang $n \to \infty$, so erhalten wir:

$$\frac{U_{1,1} + U_{0,2} + U_{n-1,1}}{U} = I_1 + I_2,$$

wo

$$I_1 = \frac{1}{4} \frac{1}{(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} \frac{(e^{2ix'} + 2e^{ix'} \cos x - 3)(e^{-2ix'} + 2e^{-ix'} \cos x - 3)}{\sin^2 \dfrac{x}{2} + \sin^2 \dfrac{x'}{2}} \, dx\, dx' = 4,$$

ferner

$$I_2 = -\frac{1}{c}\frac{1}{(2\pi)^4}\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi}\frac{1}{16}(e^{2ix'}+2e^{ix'}\cos x-3)\cdot$$

$$\cdot\frac{\left[e^{-i\frac{x+x'}{n}}-1\right]\left[e^{i\frac{y+y'}{n}}-1\right]}{\left(\sin^2\frac{x}{2}+\sin^2\frac{x'}{2}\right)\left(\sin^2\frac{y}{2}+\sin^2\frac{y'}{2}\right)}(e^{-2iy'}+2e^{-iy'}\cos y-3)\,dx\,dx'\,dy\,dy',$$

$$c=\frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{2\pi}\frac{\sin^2\frac{x+y}{2}}{\sin^2\frac{x}{2}+\sin^2\frac{y}{2}}\,dx\,dy$$

ist.

Der Integrand im Ausdruck von c ist nichtnegativ, so ist $c>0$. Das vierfache Integral zerfällt in das Produkt von zwei doppelten Integralen. Das erste von diesen kann man nach einer kleinen Umänderung mit dem Nenner vereinfachen. Der Ausdruck von I_2 hat somit folgende Gestalt:

$$I_2 = -\frac{1}{\pi^2}\int_0^{2\pi}\int_0^{2\pi}\frac{1-\cos(x+y)}{2-\cos x-\cos y}\,dx\,dy=-\frac{8}{\pi},$$

wie es leicht zu beweisen ist.

Also lautet der resultierende Widerstand des Systems:

$$R=\frac{r}{4-I_1-I_2}=\frac{\pi}{8}r;$$

$$R\approx 0{,}394r.$$

Herrn Dr. P. Rózsa danke ich für die wertvolle Förderung dieser Arbeit.

LITERATURVERZEICHNIS

[1] Litzman, O. und Rózsa P.: Allgemeine Behandlung primitiver idealer u. nichtidealer Kristallgitter mit Anwendung der Theorie der Hypermatrizen. *Phys. Status Sol.* **2** (1962) 28—41.

[2] Egerváry, E.: On hypermatrices whose blocks are commutable in pairs and their application in lattice-dinamics, *Acta Sci. Math.* (Szeged) **15** (1954) 211—222.

[3] Egerváry, E.: Über eine Methode zur numerischen Lösung der Poissonschen Differenzengleichung für beliebige Gebiete, *Acta Math. Acad. Sci. Hungar.* **11** (1960) 350—355.

ZENTRALFORSCHUNGSINSTITUT FÜR PHYSIK, BUDAPEST

(Eingegangen: 20. Februar, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 285—290.

KARTESISCHES PRODUKT VON MENGENSYSTEMEN UND GRAPHEN

von

W. IMRICH

Mengensysteme sind eine Verallgemeinerung von Graphen. In dieser Arbeit wird der Begriff des Kartesischen Produkts von Graphen auf Mengensysteme erweitert und gezeigt, daß die Primfaktorzerlegung zusammenhängender Mengensysteme bezüglich des Kartesischen Produkts eindeutig ist, falls überhaupt eine Primfaktorzerlegung existiert. Eine einfache Folgerung davon ist, daß idempotente zusammenhängende Mengensysteme keine Primfaktorzerlegung haben. Weiters wird gezeigt, daß die Automorphismengruppe eines zusammenhängenden Mengensystems mit Primfaktorzerlegung isomorph zur Automorphismengruppe der Summe der Faktoren ist.

SABIDUSSI [1] erzielte ähnliche Ergebnisse. Er zeigte mit einer anderen Methode, daß die Primfaktorzerlegung von zusammenhängenden Graphen finiten Typs oder mit einem Punkt endlichen Grades eindeutig ist, und die Automorphismengruppe solcher Graphen isomorph zur Automorphismengruppe der Summe der Faktoren ist. Außerdem konstruierte er zusammenhängende idempotente Graphen und vermutete, daß sie keine Primfaktorzerlegung haben.

Es ist zweckmäßig mehrere graphentheoretische Begriffe auf Mengensysteme zu verallgemeinern:

DEFINITION 1. Unter einem Mengensystem X verstehen wir ein geordnetes Paar $\langle g, G\rangle$, bestehend aus einer Menge g und einer aus Teilmengen von g zusammengesetzten Menge G. Wir lassen allerdings nur solche Teilmengen zu, die mindestens zwei Punkte aus g enthalten. Die Elemente von g heißen Punkte, wir bezeichnen sie mit kleinen lateinischen Buchstaben $a, b, c, \ldots$. Die Elemente von G nennen wir Kanten. Sie werden mit großen lateinischen Buchstaben bezeichnet. Die Kardinalzahl der Menge der Kanten, denen ein Punkt angehört, ist der Grad dieses Punktes.

Besteht G nur aus Paaren (a, b), so liegt ein ungerichteter Graph ohne Mehrfachkanten und Schlingen vor.

DEFINITION 2. Wir sagen die Kantenfolge $\{C_1, C_2, \ldots, C_n\}$ verbindet die Punkte a und b, falls $a \in C_1$ und $b \in C_n$ ist, sowie $C_i \cap C_{i+1} \neq \emptyset$ für $i = 1, 2, \ldots, n-1$. Liegen a und b in einer Kante, so sagen wir a und b seien durch eine Kante verbunden. Die Zahl der Kanten einer Kantenfolge nennen wir die Länge der Kantenfolge, das Minimum der Längen der Kantenfolgen, die zwei Punkte verbinden, den Abstand dieser Punkte. Ein Mengensystem ist zusammenhängend, falls je zwei Punkte durch eine endliche Kantenfolge verbunden sind.

DEFINITION 3. Es seien $X = \langle g, G\rangle$ und $Y = \langle h, H\rangle$ Mengensysteme. Unter einem Isomorphismus von X auf Y verstehen wir eine eineindeutige Abbildung φ von g auf h,

sodaß $\varphi A \in H$ ist für alle $A \in G$ und $\varphi^{-1} B \in G$ für alle $B \in H$, symbolisch $\varphi X = Y$. Für isomorphe X und Y führen wir außerdem die Schreibweise $X \cong Y$ ein. Ein Automorphismus ist ein Isomorphismus von X auf X selbst.

Sind $X = \langle g, G \rangle$ und $Y = \langle h, H \rangle$ Mengensysteme mit $g \subset h$ und $G \subset H$, so sagen wir X sei in Y enthalten, $X \subset Y$. Unter dem Durchschnitt zweier Mengensysteme X und Y verstehen wir das Mengensystem $X \cap Y = \langle g \cap h, F \rangle$, wobei F aus allen jenen Elementen von G und H besteht, die in $g \cap h$ enthalten sind. Ist $\{X_\iota | \iota \in I\}$ eine Menge von Mengensystemen $X_\iota = \langle g_\iota, G_\iota \rangle$ mit $g_\iota \cap g_\varkappa = \emptyset$ für $\iota \neq \varkappa$, so sagen wir das Mengensystem $X = \langle \bigcup_{\iota \in I} g_\iota, \bigcup_{\iota \in I} G_\iota \rangle$ sei die Summe $\sum_{\iota \in I} X_\iota$ der Mengensysteme X_ι.

Ist $\{g_\iota | \iota \in I\}$ eine Menge von Mengen, so bezeichne $g = \prod_{\iota \in I} g_\iota$ das Kartesische Produkt der Mengen g_ι. Für $A \subset g$ bezeichne weiters $p_\iota A$ die Projektion von A auf die ι-te Koordinate von g.

DEFINITION 4. Es sei $\{X_\iota | \iota \in I\}$ eine Menge von Mengensystemen $X_\iota = \langle g_\iota, G_\iota \rangle$. Unter dem Kartesischen Produkt $\prod_{\iota \in I} X_\iota$ der Mengensysteme X_ι verstehen wir das Mengensystem $\langle g, G \rangle$, wobei $g = \Pi g_\iota$ ist, und G aus allen Teilmengen A von g besteht, für die es ein $\varkappa$ gibt, sodaß $p_\varkappa A \in G_\varkappa$ und $p_\iota A \in g_\iota$ für alle $\iota \neq \varkappa$. Da wir nur ein Produkt betrachten, werden wir oft statt „Kartesisches Produkt" nur „Produkt" sagen.

Offensichtlich ist das Kartesische Produkt von Mengensystemen assoziativ und kommutativ, wenn man isomorphe Mengensysteme identifiziert. Außerdem gibt es eine Einheit, nämlich das triviale Mengensystem $E = \langle g, \emptyset \rangle$, wobei g nur aus einem Punkt besteht. Das Produkt endlich vieler zusammenhängender Mengensysteme ist zusammenhängend, jedoch ist das Produkt nicht zusammenhängend, wenn unendlich viele Faktoren auftreten, oder ein Faktor nicht zusammenhängend ist.

DEFINITION 5. Unter einer $X_\varkappa$-Schicht von $X = \langle g, G \rangle = \Pi X_\iota$ verstehen wir ein in X enthaltenes Mengensystem $Y = \langle h, H \rangle$, für das $p_\varkappa h = g_\varkappa$, $p_\varkappa H = G_\varkappa$ ist und $p_\iota h \in g_\iota$ für alle $\iota \neq \varkappa$. H besteht also aus allen Elementen von G, die in h enthalten sind. Die $X_\varkappa$-Schicht, die den Punkt a enthält bezeichnen wir mit $X_\varkappa^a$.

LEMMA 1. *Sind $a, b \in X_\iota^c$ und liegen a und b in einer Kante A aus ΠX_ι, so ist A in X_ι^c enthalten.*

LEMMA 2. *Sind X_ι^a und X_ι^b zwei verschiedene Schichten des Mengensystems ΠX_ι, so ist jeder Punkt aus X_ι^a mit genau einem Punkt aus X_ι^b durch eine Kante verbunden, oder kein Punkt aus X_ι^a ist mit Punkten aus X_ι^b durch eine Kante verbunden.*

LEMMA 3. *Ist $a \in X_\iota^a$ mit zwei verschiedenen Punkten aus X_ι^b durch je eine Kante verbunden, so ist $X_\iota^a = X_\iota^b$.*

LEMMA 4. *Es sei φ eine isomorphe Abbildung des zusammenhängenden Mengensystems $R \times S = T$ auf $X \times Y = Z$, bei der jede R-Schicht von T auf eine X-Schicht von Z abgebildet wird. Dann gibt es einen Isomorphismus ϱ von R auf X und σ von S auf Y mit $\varphi(r, s) = (\varrho r, \sigma s)$.*

BEWEIS. Lemma 1 und 2 folgen direkt aus der Definition des Kartesischen Produkts. Lemma 3 ist eine unmittelbare Folgerung aus Lemma 2. Um Lemma 4

zu beweisen, zeigen wir zuerst, daß aus $\varphi(r, s)=(x, y)$ und $\varphi(r, t)=(x', z)$ folgt $x=x'$. Da wir $s\neq t$ voraussetzen, liegen (r, s) und (r, t) in verschiedenen R-Schichten aus T. Nach Voraussetzung wird jede R-Schicht auf eine X-Schicht abgebildet, also ist $y\neq z$. Weiters ist s zusammenhängend, also gibt es eine Kantenfolge in einer S-Schicht, die (r, s) mit (r, t) verbindet. Diese Kantenfolge wird durch φ auf eine Kantenfolge in Z abgebildet, die (x, y) mit (x', z) verbindet. Ist die Länge der Kantenfolge eins, so liegen (x, y) und $(x' z)$ in einer Kante. Da $y\neq z$ ist, muß laut Definition des Kartesischen Produkts $x=x'$ sein. Durch Induktion ergibt sich die Richtigkeit der Behauptung $x=x'$ auch für beliebig lange Kantenfolgen. Es gibt also eine von s unabhängige Abbildung ϱ, mit $\varphi(r, s)=(\varrho r, y(r, s))$. Das heißt, daß jede S-Schicht von T auf eine Y-Schicht von Z abgebildet wird. Es gibt daher eine von r unabhängige Abbildung σ mit $\varphi(r, s)=(x, \sigma s)$, also $\varphi(r, s)=$ $=(\varrho r, \sigma s)$. Man sieht leicht, daß ϱ ein Isomorphismus von R auf X und σ ein Isomorphismus von S auf Y ist.

DEFINITION 6. Ein Mengensystem X ist prim, wenn aus $X\cong Y\times Z$ folgt $Y\cong E$ oder $Z\cong E$, wobei E das triviale Mengensystems ist. Zwei Mengensysteme X, Y sind relativ prim, wenn aus $X\cong U\times Z$ und $Y\cong V\times Z$ folgt $Z\cong E$.

Ist $X=\langle g, G\rangle$ ein zusammenhängendes Mengensystem mit einem Punkt endlichen Grades, so hat X eine Primfaktorzerlegung. Denn sind $X=\langle g, G\rangle$ und $Y=\langle h, H\rangle$ zwei Mengensysteme und ist (a, b) ein Punkt aus $X\times Y$, wobei a den Grad α in X hat und b den Grad β in Y, so hat (a, b) den Grad $\alpha+\beta$ in $X\times Y$. Ebenso hat jedes zusammenhängende Mengensystem, für das die Distanz von je zwei beliebigen Punkten unter einer Schranke m liegt eine Primfaktorzerlegung.

SATZ 1. *Es sei $Z=U\times V=X\times Y$ ein zusammenhängendes Mengensystem. Dann ist jede U-Schicht von Z das Produkt $P\times Q$ zweier Mengensysteme $P\subset X$ und $Q\subset Y$.*

BEWEIS. Es sei $X=\langle g, G\rangle$ mit $G=\{A_\iota|\iota\in I\}$, $Y=\langle h, H\rangle$ mit $H=\{B_\varkappa|\varkappa\in K\}$ und $U^a=\langle w, W\rangle$ eine U-Schicht von Z. Für die Punkte d mit $p_X d\in A_\iota$ bezeichne ferner A_ι^d die Kante aus Z mit $p_X A_\iota^d=A_\iota$ und $d\in A_\iota^d$. Analog für $B_\varkappa^d$. Wir führen den Beweis in mehreren Schritten.

(A) Sind A_ι^d und $B_\varkappa^d$ zwei Kanten aus U^a, so zeigen wir zuerst, daß alle Punkte z mit $p_X z\in A_\iota$ und $p_Y z\in B_\varkappa$ in U^a liegen. Denn sei etwa $p_X z\in A_\iota$ und $p_Y z\in B_\varkappa$. Dann ist $e=(p_X z, p_Y d)\in A_\iota^d$, $f=(p_X d, p_Y z)\in B_\varkappa^d$ und z und e liegen in der Kante $B_\varkappa^z$, sowie z und f in der Kante A_ι^z. Liegt z also nicht schon in A_ι^d oder $B_\varkappa^d$, so ist z mit zwei verschiedenen Punkten von U^a durch je eine Kante verbunden und liegt daher nach Lemma 3 in U^a.

Da $U^a=\langle w, W\rangle$ alle Kanten von Z enthält, die Teilmengen von w sind, liegen also auch alle Kanten A_ι^z und $B_\varkappa^z$ mit $p_X z\in A_\iota$ und $p_Y z\in B_\varkappa$ in U^a. Das bedeutet, daß das Kartesische Produkt der Mengensysteme A_ι und $B_\varkappa$ in U^a liegt.

(B) Liegt die die Punkte b und c verbindende Kantenfolge $\{B_\varkappa^d, A_\iota^d\}$ in U^a, so liegt auch $\{A_\iota^b, B_\varkappa^c\}$ in U^a und verbindet b und c, wie gerade gezeigt wurde. Liegen also $b, c\in U^a$ nicht in derselben Y-Schicht, so kann man annehmen, daß von zwei Kanten aus U^a, die b und c verbinden, die erste in einer X-Schicht und die zweite in einer Y-Schicht von Z liegt.

(C) Ist $\{C_1, C_2, \ldots, C_n\}$ eine Kantenfolge F in U^a mit $C_i\cap C_{i+1}\neq\emptyset$ für $1\leqq i<n$, so zeigen wir durch Induktion, daß alle z mit $p_X z\in p_X C_i$ für ein C_i aus F und $p_Y z\in$

$\in p_Y C_k$ für ein C_k aus F in U^a liegen. Der Fall $n=2$ wurde soeben behandelt. Die Behauptung sei also für $2 \leqq k < n$ richtig. Liegen C_1 und C_n in X-Schichten, so liegt $p_Y z$ in $p_Y C_k$ für ein k mit $1 < k < n$. Je nachdem ob $p_X z \in p_X C_i$ mit $1 \leqq i < n$, oder $i = n$ ist, wenden wir die Induktionsvoraussetzung auf $\{C_1, \ldots, C_{n-1}\}$ oder auf $\{C_2, \ldots, C_n\}$ an. Ebenso geht man vor, wenn C_1 und C_2 beide in Y-Schichten von Z liegen.

Es bleibt also der Fall zu betrachten, wo von den Kanten C_1 und C_n eine in einer X-Schicht und eine in einer Y-Schicht von Z liegt. Falls nicht schon C_1 in einer X-Schicht liegt, so kann man dies durch Umkehrung der Numerierung der C_i erreichen. Es sei also $p_X C_1 = A_\iota$ und $p_Y C_n = B_\varkappa$. Ist $p_X z \in p_X C_i$ und $p_Y z \in p_Y C_k$ mit $1 \leqq i, k < n$ oder $1 < i, k \leqq n$, so wenden wir die Induktionsvoraussetzung auf $\{C_1, \ldots, C_{n-1}\}$ oder $\{C_2, \ldots, C_n\}$ an. Es sei nun $p_X z \in A_\iota$ und $p_Y z \in B_\varkappa$. Da $C_i \cap C_{i+1} \neq \emptyset$ ist für $1 \leqq i < n$, gibt es einen Punkt $c \in C_1 \cap C_2$ und einen Punkt $d \in C_{n-1} \cap C_n$. Laut Voraussetzung liegt dann der Punkt $e = (p_X c, p_Y d)$ in U^a. Ebenso sind alle Punkte $(p_X c, s)$ mit $s \in B_\varkappa$ und $(r, p_Y d)$ mit $r \in A_\iota$ in U^a. Laut Definition der Schicht liegen dann die Kanten A_ι^e und $B_\varkappa^e$ in U^a. Also liegen alle Punkte z mit $p_X z \in p_X A_\iota^e = A_\iota = p_X C_1$ und $p_Y z \in p_Y B_\varkappa^e = B_\varkappa = p_Y C_n$ in U^a.

(D) Sind b und c zwei beliebige Punkte aus U^a, die in verschiedenen X- und Y-Schichten liegen, so gibt es eine Kantenfolge $\{C_1, C_2, \ldots, C_n\}$ in U^a, die b mit c verbindet, und zwar derart, daß für ein bestimmtes l die Kanten $C_1, \ldots, C_l$ in einer X-Schicht von Z liegen und $C_{l+1}, \ldots, C_n$ in einer Y-Schicht von Z.

Da U^a zusammenhängend ist, gibt es sicher Kantenfolgen in U^a, die b und c verbinden, es bleibt zu zeigen, daß eine mit den gewünschten Eigenschaften darunter ist. Für $n=2$ wurde dies in (B) gezeigt. Die Aussage sei also richtig für $2 \leqq k < n$. Ist C_1 in einer X-Schicht enthalten, so wenden wir die Induktionsvoraussetzung auf $\{C_2, \ldots, C_n\}$ an und alles ist bewiesen. Es liege also C_1 in einer Y-Schicht. Da b und c in verschiedenen X- und Y-Schichten liegen, gibt es sicher C_i, die in einer X-Schicht liegen. Es sei C_m die erste Kante dieser Art. Man kann $m < n$ voraussetzen, andernfalls wende man die Induktionsvoraussetzung auf $\{C_{n-1}, C_n\}$ an. Durch Anwendung der Induktionsvoraussetzung auf $\{C_1, \ldots, C_m\}$ erhalten wir dann eine Kantenfolge $\{D_1, \ldots, D_m, C_{m+1}, \ldots, C_n\}$, die b und c verbindet, wobei D_1 in einer X-Schicht liegt. Dieser Fall wurde aber gerade betrachtet.

(E) Wir zeigen nun, daß $U^a = \langle w, W \rangle$ das Produkt zweier Mengensysteme ist. Es sei $p_X W$ die Menge aller Kanten A aus X, für die es ein d gibt mit $A^d \in W$. Analog definieren wir $p_Y W$. Setzt man $P = \langle p_X w, p_X W \rangle$ und $Q = \langle p_Y w, p_Y W \rangle$, so folgt aus (C) und (D) sofort $U^a = P \times Q$.

SATZ 2. *Ist φ ein Isomorphismus des zusammenhängenden Mengensystems $P \times Q = R$ auf $X \times Y = Z$, wobei P prim ist, so liegen die φ-Bilder der P-Schichten von R entweder alle in X-Schichten von Z, oder alle in Y-Schichten.*

BEWEIS. Es genügt zu zeigen, daß für $P \times Q = X \times Y = Z$ jede P-Schicht in einer X-Schicht, oder jede P-Schicht in einer Y-Schicht von Z liegt. Wegen Satz 1 ist jede P-Schicht ganz in einer X-Schicht oder ganz in einer Y-Schicht enthalten. Wir führen den Beweis nun indirekt. Angenommen es sei $P^a \subset X^a$ und $P^b \subset Y^b$. Sind a und b nicht schon in einer Kante enthalten, so gibt es eine Kantenfolge, die a und b verbindet. Jeder Punkt dieser Kantenfolge liegt in einer P-Schicht. Da die Kantenfolge nur endlich viele Kanten hat, gibt es eine Kante, in der zwei Punkte liegen, von denen einer einer P-Schicht angehört, die

in einer X-Schicht enthalten ist, und der andere einer P-Schicht, die in einer Y-Schicht enthalten ist. Es sei also $P^a \subset X^a$, $P^b \subset Y^b$ und o.B.d.A. a, $b \in A_i^a \subset X^a$ und $c = (p_P c, p_Q b)$ ein Punkt aus P^b, der mit b in einer Kante liegt. Offenbar ist $c \in Y^b$, $c \notin X^a$. Weiters liegt der Punkt $(p_P c, p_Q a) \neq b$ in P^a und ist mit c durch eine Kante verbunden. Dann ist also $c \notin X^a$ mit zwei Punkten aus X^a durch je eine Kante verbunden, im Widerspruch zu Lemma 2.

FOLGERUNG. *Liegt unter den Voraussetzungen von Satz 2 das Bild einer P-Schicht von $P \times Q$ in X^a, so ist $X^a = \varphi(P \times Q')$ für ein Mengensystem $Q' \subset Q$.*

BEWEIS. Nach Satz 2 liegen dann alle Bilder von P-Schichten von $P \times Q$ in X-Schichten von $X \times Y$. Ist $P = \langle g, G \rangle$ und $Q = \langle h, H \rangle$, so bezeichnen wir die Menge aller $q \in h$, für die $\varphi\{(p, q) | p \in g\} \subset X^a$ ist mit h' und das Mengensystem $\langle h', H' \rangle$, wobei $H' = \{A | A \in H, A \subset h'\}$ ist, mit Q'. Dann ist offensichtlich $\varphi(P \times Q') = X^a$.

SATZ 3. *Ist $\varphi(P \times R) = P_1 \times P_2 \times \ldots \times P_n$, wobei P und $P_1, \ldots, P_n$ zusammenhängende prime Mengensysteme sind, so gibt es ein P_k, sodaß jede P-Schicht von $P \times R$ auf eine P_k-Schicht von $P_1 \times \ldots \times P_n$ abgebildet wird.*

BEWEIS durch Induktion. Für $n = 1$ ist der Satz sicher richtig. Es sei nun $P_2 \times \ldots \times P_n = S$, also $\varphi(P \times R) = P_1 \times S$. Nach Satz 2 sind zwei Fälle möglich: 1. Jede P-Schicht von $P \times R$ wird auf eine P_1-Schicht von $P_1 \times S$ abgebildet, dann ist nichts mehr zu zeigen. 2. Alle Bilder von P-Schichten liegen in S-Schichten. Jede Schicht S^a von $P_1 \times S$ ist isomorph zu $P_2 \times \ldots \times P_n$, und nach der Folgerung zu Satz 2 gibt es ein Mengensystem $R' = \langle h', H' \rangle \subset R$ mit $\varphi(P \times R') = S^a$. Laut Induktionsvoraussetzung gibt es also ein P_i, sodaß jede P-Schicht von $P \times R'$ auf eine P_i-Schicht von S^a abgebildet wird. Für ein festes $r \in h'$ und beliebiges $p \in P$ gilt also

$$\varphi(p, r) = (p_1, p_2, \ldots, p_{i-1}, \pi p, p_{i+1}, \ldots, p_n).$$

Ist $P^{(q, s)}$ eine andere P-Schicht von $P \times R$, so gilt analog

$$\varphi(q, s) = (q_1, q_2, \ldots, q_{k-1}, \varrho q, q_{k+1}, \ldots, q_n).$$

Wir brauchen nur noch zu zeigen, daß $k = i$ ist und $\pi p = \varrho p$. Für $p_1 = q_1$ ist dies laut Induktionsvoraussetzung der Fall. Liegen r und s in einer Kante von R, so liegen $\varphi(p, r)$ und $\varphi(p, s)$ ebenfalls in einer Kante. Nach der Definition des Kartesischen Produktes muß dann auch für $p_1 \neq q_1$ $k = i$ sein und $\pi p = \varrho p$. Da R zusammenhängend ist, und jede zwei Punkte verbindende Kantenfolge nur endlich viele Kanten hat, ergibt sich die Richtigkeit der Behauptung.

SATZ 4. *Ist X ein zusammenhängendes Mengensystem und hat X eine Primfaktorzerlegung, so ist sie eindeutig bis auf die Reihenfolge und Isomorphie der Faktoren.*

BEWEIS. Da X zusammenhängend ist, kann X nur als Produkt von höchstens endlich vielen Faktoren dargestellt werden. Wir führen den Beweis durch Induktion nach der Minimalzahl der Faktoren aller möglichen Primfaktorzerlegungen von X. Für $n = 1$ ist der Satz sicher richtig, er gelte also für $k < n$. Es sei $P_1 \times \ldots \times P_n$ eine minimale Primfaktorzerlegung von X und $Q_1 \times \ldots \times Q_m$, $m \geqq n$ eine beliebige andere. Setzen wir $Q_2 \times \ldots \times Q_m = R$, so ist $Q_1 \times R = P_1 \times \ldots \times P_n$. Nach Satz 3 gibt es daher ein P_i, sodaß jede Q_1-Schicht von X eine P_i-Schicht von X ist. Klarerweise

ist $P_i \cong Q_1$. Nach Lemma 4 ist weiters $R \cong P_1 \times \ldots \times P_{i-1} \times P_{i+1} \times \ldots \times P_n$. Da nach Induktionsvoraussetzung die Q_k, $2 \leqq k \leqq m$ mit den P_l, $1 \leqq l \leqq n$, $l \neq i$ bis auf Reihenfolge und Isomorphie übereinstimmen, ist der Satz bewiesen.

FOLGERUNG. *Ist X ein zusammenhängendes idempotentes Mengensystem, d.h. ist $X \times X \cong X$, so hat X keine Primfaktorzerlegung.*

BEWEIS. Hätte X eine Primfaktorzerlegung, so wäre sie nicht eindeutig.

SATZ 5. *Ist $P_1 \times \ldots \times P_n$ eine Primfaktorzerlegung des zusammenhängenden Mengensystems X, und sind die Punktmengen der P_i paarweise disjunkt, so ist die Automorphismengruppe von X isomorph zur Automorphismengruppe der Summe der P_i.*

BEWEIS. Aus dem Beweis von Satz 4 geht hervor, daß es zu jedem Automorphismus φ von X und jedem P_i ein P_k gibt, sowie einen Isomorphismus π von P_i auf P_k mit

$$\varphi(p_1, \ldots, x_i, \ldots, p_n) = (q_1, \ldots, q_{k-1}, \pi x_i, q_{k+1}, \ldots, q_n).$$

Man sieht, daß jeder Automorphismus von X einen Automorphismus von $\sum P_i$ liefert. Auch die Umkehrung zeigt man leicht.

LITERATURVERZEICHNIS

[1] SABIDUSSI, G.: Graph Multiplication, *Math. Z.* **72** (1960) 446—457.

TECHNISCHE HOCHSCHULE 3. INSTITUT FÜR MATHEMATIK, WIEN

(Eingegangen: 4. August, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 291—292.

AN INFORMATION THEORETICAL IDENTITY AND A PROBLEM INVOLVING CAPACITY

by

F. TOPSØE

Let $X=\{x_i: i=1, 2, \ldots, n\}$ and $Y=\{y_j: j=1, 2, \ldots, m\}$ be finite sets. A probability distribution over X will be denoted by the letter $\mathbf{p}$ and one over Y by the letter $\mathbf{q}$.

Suppose that the distribution of Y is changed from $\mathbf{q}^*$ to $\mathbf{q}$. Then we define the amount of information one gains by knowing that the distribution has changed from $\mathbf{q}^*$ to $\mathbf{q}$ by

$$I(\mathbf{q}^* : \mathbf{q}) = \sum_j q_j \log (q_j/q_j^*). \tag{1}$$

In order that this quantity be finite we must assume that q_j vanishes whenever q_j^* vanishes (no new events have been born). The quantity just defined has been studied by many authors; in the book [1] the notation $I(\cdot\|\cdot)$ is used. $I(\mathbf{q}^* : \mathbf{q})$ is non-negative, and it is zero iff $\mathbf{q}^*=\mathbf{q}$.

Now suppose that a joint probability distribution on the product space $X\times Y$ is given. Let $\mathbf{p}$ and $\mathbf{q}$ denote the marginal distributions on X and Y respectively, and denote by $\mathbf{q}_i$ the conditional distribution on Y given x_i. Clearly, X and Y can also be regarded as random variables. The amount of information X contains about Y can be defined by the equation

$$I(X, Y) = \sum_i p_i I(\mathbf{q} : \mathbf{q}_i). \tag{2}$$

This quantity is symmetrical, so that $I(X, Y)=I(Y, X)$.

If $\mathbf{q}^*$ is any probability distribution over Y such that $I(\mathbf{q}^* : \mathbf{q})$ is finite then the identity

$$I(X, Y) = \sum_i p_i I(\mathbf{q}^* : \mathbf{q}_i) - I(\mathbf{q}^* : \mathbf{q}) \tag{3}$$

holds.

This identity is a simple consequence of the additive property of the logarithmic function.

We now want to apply our identity to the problem of finding the capacity for a memoryless channel. Let (X, P, Y) be the channel. X and Y are finite sets as before and $P=(p_{ij})_{i=1,\ldots,n;\, j=1,\ldots,m}$ a stochastic matrix. The i'th row in P is denoted $\mathbf{q}_i$ i.e. $\mathbf{q}_i=(p_{i1}, \ldots, p_{im})$. We assume that no column in P is identically zero. Clearly, any distribution $\mathbf{p}$ on X induces a distribution on $X\times Y$; if we call the corresponding marginal distribution on Y for $\mathbf{q}$ then $\mathbf{q}=\sum_i p_i\mathbf{q}_i$. A distribution on Y is called admissible if it is a convex combination of the $\mathbf{q}_i$'s. An optimal source is a distribution on X that maximizes $I(X, Y)$; the maximal value of $I(X, Y)$ is the capacity of the channel.

THEOREM. *Let* $\mathbf{p}^*$ *be any distribution on* X *with all* p_i^* *positive and denote by* $\mathbf{q}^*$ *the corresponding distribution on* Y *i.e.* $\mathbf{q}^* = \sum_i p_i^* \mathbf{q}_i$. *Then a necessary and sufficient condition that* $\mathbf{p}^*$ *is an optimal source is that the value of* $I(\mathbf{q}^* : \mathbf{q}_i)$ *is independent of* i; $i = 1, 2, \ldots, n$.

If the condition is fulfilled then the common value of $I(\mathbf{q}^* : \mathbf{q}_i)$ *is the capacity of the channel.*

The necessity is well known and can for instance be proved by introducing Lagrange-multipliers. The sufficiency, as well as the last part of the theorem, is an easy concequence of the identity (3) and the non-negativity of $I(\mathbf{q}^* : \mathbf{q})$.

It is easy to prove that a necessary and sufficient condition for the existence of a distribution $\mathbf{q}^*$ with all $\mathbf{q}^*$ positive and such that $I(\mathbf{q}^* : \mathbf{q}_i)$ is independent of i, is that the matrix equation $P\mathbf{x} = \mathbf{h}$ has a solution; here h_i is the entropy of $\mathbf{q}_i$ ($h_i = -\sum_j p_{ij} \log p_{ij}$). A complicated problem arises, however, since we cannot be sure that such a $\mathbf{q}^*$ is admissible (compare the example in [2]). Unfortunately, we have not been able to describe in informationtheoretical terms when this unpleasant situation arises.

REFERENCES

[1] RÉNYI, A.: *Wahrscheinlichkeitsrechnung mit einem Anhang über Informationstheorie*, Berlin, VEB. Deutscher Verlag der Wissenschaften, 1962.

[2] SILVERMAN, R. A.: On Binary Channels and Their Cascades, *IRE Trans. on Inform. Theory* **1** (1955) 19—27.

UNIVERSITY OF COPENHAGEN

(Received September 26, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 293—298.

ON THE ZEROS OF THE SOLUTIONS OF THE DIFFERENTIAL EQUATION $y''+q(x)y=0$, WHERE $[q(x)]^\nu$ IS CONCAVE

by

Á. ELBERT

A. S. GALBRAITH proved the following theorem [1]:

In the differential equation

$$y''+q(x)y=0 \tag{1}$$

suppose $q(x)$ to be nonnegative, monotonic and concave (no point of an arc lies below its chord) in some closed interval $[a, b]$. If

$$(b-a)\int_a^b q(x)\,dx \geqq \frac{9}{8}n^2\pi^2,$$

where n is an integer, then every solution of (1) has at least n zeros in $[a, b]$. The number 9/8 cannot be replaced by a smaller one.

In this paper we shall generalize this theorem in two directions and at the same time give another proof of the theorem of GALBRAITH. At first we make a

DEFINITION. The function $q(x)$ belongs to $C_\nu[a, b]$ $(0<\nu<\infty)$ if $q(x)$ is a nonnegative continuous function in the closed interval $[a, b]$ and $[q(x)]^\nu$ is concave (in sense mentioned above).

It is obvious that the set $C_1[a, b]$ is a little more general as the set of those functions which are taken into account in the theorem of GALBRAITH.

Now we can formulate our

THEOREM. *If* $q(x)\in C_\nu[a, b]$ *and*

$$(b-a)\int_a^b q(x)\,dx \geqq \frac{(2\nu+1)^2}{4\nu(\nu+1)}n^2\pi^2, \tag{2}$$

where n *is an integer, then every solution of* (1) *has at least* n *zeros in* $[a,b]$. *The coefficient of* $n^2\pi^2$ *on the right side of* (2) *cannot be replaced by any smaller one.*

Throughout this paper we shall use the notation $A(\nu)=\dfrac{(2\nu+1)^2}{4\nu(\nu+1)}$.

We set out from a result of E. MAKAI [2]. His result implies the following

LEMMA I. If $p(x)$ is a nonnegative continuous function in the closed interval $[a, b]$ with continuous first and second derivatives and the inequalities

$$\frac{5}{4}p'^2(x)-p(x)p''(x)\geqq 0$$

and

$$\int_a^b \sqrt{p(x)}\,dx \geqq n\pi$$

hold, where n is an integer, then every solution of the differential equation

$$y''+p(x)y=0$$

has at least n zeros in $[a, b]$.

We need yet two other Lemmas too.

LEMMA II. If $\gamma(u)$ is a non-negative continuous function in the interval $[0, \tau]$ which satisfies the equalities

$$\int_0^\tau \gamma(u)\,du = \alpha$$

and

$$\int_0^\tau \gamma(u)\frac{\tau^{1+\frac{1}{2\nu}}-u^{1+\frac{1}{2\nu}}}{\tau-u}\,du = \beta \qquad (0<\nu<\infty)$$

then the inequality

$$\int_0^\tau \gamma(u)\frac{\tau^{1+\frac{1}{\nu}}-u^{1+\frac{1}{\nu}}}{\tau-u}\,du < 2\beta\tau^{\frac{1}{2\nu}}-\alpha\tau^{\frac{1}{\nu}}$$

holds.

PROOF: Taking into account the meaning of the quantities α and β we obtain

$$\int_0^\tau \gamma(u)\frac{\tau^{1+\frac{1}{\nu}}-u^{1+\frac{1}{\nu}}}{\tau-u}\,du = \int_0^\tau \gamma(u)\left[\tau^{\frac{1}{\nu}}+u\frac{\left(\tau^{\frac{1}{2\nu}}+u^{\frac{1}{2\nu}}\right)\left(\tau^{\frac{1}{2\nu}}-u^{\frac{1}{2\nu}}\right)}{\tau-u}\right]du <$$

$$< \alpha\tau^{\frac{1}{\nu}}+2\tau^{\frac{1}{2\nu}}\int_0^\tau \gamma(u)u\frac{\tau^{\frac{1}{2\nu}}-u^{\frac{1}{2\nu}}}{\tau-u}\,du =$$

$$= \alpha\tau^{\frac{1}{\nu}}+2\tau^{\frac{1}{2\nu}}\int_0^\tau \gamma(u)\left[\frac{\tau^{1+\frac{1}{2\nu}}-u^{1+\frac{1}{2\nu}}}{\tau-u}-\tau^{\frac{1}{2\nu}}\right]du = 2\beta\tau^{\frac{1}{2\nu}}-\alpha\tau^{\frac{1}{\nu}}\,.$$

LEMMA III. If $q(x)\in C_\nu[a, b]$ and

$$\int_a^b \sqrt{q(x)}\,dx = J,$$

then

$$I=(b-a)\int_a^b q(x)\,dx \leqq A(\nu)J^2.$$

PROOF. We denote the maximum of the function $[q(x)]^\nu$ in the interval $[a, b]$ by τ, and the length of that subinterval in $[a, b]$, where the inequality $[q(x)]^\nu \geqq t$ holds $(0\leqq t\leqq\tau)$, by $\varrho(t)$. The function $\varrho(t)$ is decreasing and concave in $[0, \tau]$ and, of course, $\varrho(0)=b-a$.

By this definition of $\varrho(t)$ it is clear that

$$J=\int_a^b \sqrt{q(x)}\,dx=\int_0^{\tau^{\frac{1}{2\nu}}}\varrho(s^{2\nu})ds=\frac{1}{2\nu}\int_0^{\tau}\varrho(t)t^{\frac{1}{2\nu}-1}\,dt \tag{3}$$

and

$$I=(b-a)\int_a^b q(x)dx=(b-a)\int_0^{\tau^{\frac{1}{\nu}}}\varrho(s^{\nu})ds=\frac{b-a}{\nu}\int_0^{\tau}\varrho(t)t^{\frac{1}{\nu}-1}\,dt. \tag{4}$$

We extend the domain of the definition of $\varrho(t)$ from $[0,\tau]$ to $(-\infty,\tau]$ by making the convention $\varrho(t)=\varrho(0)=b-a$ for $t\leqq 0$. Let the function $\varrho_\varepsilon(t)$ be defined by

$$\varrho_\varepsilon(t)=\frac{1}{\varepsilon^2}\int_{t-\varepsilon}^{t}\left(\int_{s-\varepsilon}^{s}\varrho(\sigma)d\sigma\right)ds \qquad (\varepsilon>0)$$

and I_ε, J_ε by putting $\varrho_\varepsilon(t)$ instead of $\varrho(t)$ in (3) and (4). The function $\varrho_\varepsilon(t)$ is also concave $\varrho_\varepsilon(0)=\varrho(0)=b-a$, its first and second derivatives exist and are continuous.

Let us write

$$r(t,u)=\begin{cases}1 & \text{for} \quad 0\leqq t\leqq u\\ \dfrac{\tau-t}{\tau-u} & \text{for} \quad u\leqq t\leqq\tau.\end{cases}$$

A simple calculation shows that

$$\varrho_\varepsilon(t)=\varrho_\varepsilon(\tau)-\varrho_\varepsilon'(0)(\tau-t)-\int_0^{\tau}\varrho_\varepsilon''(u)(\tau-u)r(t,u)\,du \quad \text{for} \quad 0\leqq t\leqq\tau$$

and

$$\int_0^{\tau}r(t,u)t^{\frac{1}{\nu}-1}dt=\frac{\nu^2}{\nu+1}\,\frac{\tau^{1+\frac{1}{\nu}}-u^{1+\frac{1}{\nu}}}{\tau-u},$$

hence

$$J_\varepsilon=\frac{1}{2\nu}\int_0^{\tau}\varrho_\varepsilon(t)t^{\frac{1}{2\nu}-1}dt=\varrho_\varepsilon(\tau)\tau^{\frac{1}{2\nu}}-\frac{2\nu}{2\nu+1}\varrho_\varepsilon'(0)\tau^{1+\frac{1}{2\nu}}-$$

$$-\frac{2\nu}{2\nu+1}\int_0^{\tau}\varrho_\varepsilon''(u)(\tau-u)\frac{\tau^{1+\frac{1}{2\nu}}-u^{1+\frac{1}{2\nu}}}{\tau-u}\,du \tag{5}$$

and

$$I_\varepsilon\cdot(b-a)^{-1}=\frac{1}{\nu}\int_0^{\tau}\varrho_\varepsilon(t)t^{\frac{1}{\nu}-1}\,dt=\varrho_\varepsilon(\tau)\tau^{\frac{1}{\nu}}-\frac{\nu}{\nu+1}\varrho_\varepsilon'(0)\tau^{\frac{1}{\nu}+1}-$$

$$-\frac{\nu}{\nu+1}\int_0^{\tau}\varrho_\varepsilon''(u)(\tau-u)\frac{\tau^{1+\frac{1}{\nu}}-u^{1+\frac{1}{\nu}}}{\tau-u}\,du. \tag{6}$$

Let now be

$$\gamma(u) = -\varrho_\varepsilon''(u)(\tau - u),$$

$$\alpha = \int_0^\tau \gamma(u)\,du = \varrho_\varepsilon(0) - \varrho_\varepsilon(\tau) + \varrho_\varepsilon'(0)\tau$$

and

$$\beta = \frac{2\nu+1}{2\nu} J_\varepsilon - \frac{2\nu+1}{2\nu}\varrho_\varepsilon(\tau)\tau^{\frac{1}{2\nu}} + \varrho_\varepsilon'(0)\tau^{1+\frac{1}{2\nu}},$$

then by Lemma II and (5) we get from (6)

$$I_\varepsilon(b-a)^{-1} \leqq \varrho_\varepsilon(\tau)\tau^{\frac{1}{\nu}} - \frac{\nu}{\nu+1}\varrho_\varepsilon'(0)\tau^{1+\frac{1}{\nu}} + \frac{\nu}{\nu+1}\left\{2\tau^{\frac{1}{2\nu}}\left[\frac{2\nu+1}{2\nu}J_\varepsilon - \right.\right.$$

$$\left.\left. - \frac{2\nu+1}{2\nu}\varrho_\varepsilon(\tau)\tau^{\frac{1}{2\nu}} + \varrho_\varepsilon'(0)\tau^{1+\frac{1}{2\nu}}\right] - \tau^{\frac{1}{\nu}}[\varrho_\varepsilon(0) - \varrho_\varepsilon(\tau) + \varrho_\varepsilon'(0)\tau]\right\} =$$

$$= \frac{2\nu+1}{\nu+1}\tau^{\frac{1}{2\nu}}J_\varepsilon - \frac{\nu}{\nu+1}\varrho_\varepsilon(0)\tau^{\frac{1}{\nu}} = \frac{(2\nu+1)^2}{4\nu(\nu+1)}J_\varepsilon^2(b-a)^{-1} -$$

$$- \frac{\nu}{\nu+1}(b-a)\left[\frac{2\nu+1}{2\nu}J_\varepsilon(b-a)^{-1} - \tau^{\frac{1}{2\nu}}\right]^2 \leqq A(\nu)J_\varepsilon^2(b-a)^{-1},$$

i. e.

$$I_\varepsilon \leqq A(\nu)J_\varepsilon^2. \tag{7}$$

If ε tends to 0, then $\varrho_\varepsilon(t)$ tends uniformly to $\varrho(t)$, therefore from (7) we obtain

$$I \leqq A(\nu)J^2,$$

as it was stated.

PROOF OF THE THEOREM. Applying Lemma III we get from (2) that

$$\int_a^b \sqrt{q(x)}\,dx \geqq n\pi. \tag{8}$$

Let $q_{\varepsilon,\Delta}(x)$ be defined for $\varepsilon, \Delta > 0$ by

$$q_{\varepsilon,\Delta}(x) = \left[\Delta + \frac{1}{\varepsilon^2}\int_x^{x+\varepsilon}\left(\int_\xi^{\xi+\varepsilon} q^\nu(\eta)\,d\eta\right)d\xi\right]^{\frac{1}{\nu}} \qquad (a \leqq x \leqq b-2\varepsilon).$$

The function $q_{\varepsilon,\Delta}(x) \in C_\nu[a, b-2\varepsilon]$ and there exist its first and second derivatives, too, hence

$$[q_{\varepsilon,\Delta}^\nu(x)]'' = \nu q_{\varepsilon,\Delta}^{\nu-2}(x)[(\nu-1)q_{\varepsilon,\Delta}'^2(x) + q_{\varepsilon,\Delta}(x)q_{\varepsilon,\Delta}''(x)] \geqq 0,$$

therefore

$$\frac{5}{4}q_{\varepsilon,\Delta}'^2(x) - q_{\varepsilon,\Delta}(x)q_{\varepsilon,\Delta}''(x) = \left(\frac{1}{4}+\nu\right)q_{\varepsilon,\Delta}'^2(x) + [(1-\nu)q_{\varepsilon,\Delta}'^2(x) - q_{\varepsilon,\Delta}(x)q_{\varepsilon,\Delta}''(x)] \geqq 0. \tag{9}$$

Let $y=y(x)$ be any solution of (1). We will prove that the inequality (8) implies the existence of at least n zeros of $y(x)$ in $[a, b]$. Let $y_{\varepsilon,\Delta}(x)$ be the solution of the differential equation

$$y''+q_{\varepsilon,\Delta}(x)y=0$$

with the initial conditions $y_{\varepsilon,\Delta}(a)=y(a)$ and $y'_{\varepsilon,\Delta}(a)=y'(a)$. If ε is small enough, then by Lemma I, (8) and (9) the function $y_{\varepsilon,\Delta}(x)$ has at least n zeros in $[a, b-2\varepsilon]$ since

$$\lim_{\varepsilon\to 0}\int_a^{b-2\varepsilon}\sqrt{q_{\varepsilon,\Delta}(x)}\,dx=\int_a^b[\Delta+q^\nu(x)]^{\frac{1}{2\nu}}\,dx>\int_a^b\sqrt{q(x)}\,dx\geqq n\pi.$$

But

$$\lim_{\substack{\varepsilon\to 0\\ \Delta\to 0}} q_{\varepsilon,\Delta}(x)=q(x),$$

hence

$$\lim_{\substack{\varepsilon\to 0\\ \Delta\to 0}} y_{\varepsilon,\Delta}(x)=y(x),$$

and $y(x)$ has at least n zeros in the interval $[a, b]$, too.

Finally we shall show that the quantity $A(\nu)$ cannot be replaced by a smaller one. Let $q(x)$ be $x^{\frac{1}{\nu}}\in C_\nu[0,\infty)$ and consider the differential equation

$$y''+x^{\frac{1}{\nu}}y=0.$$

Let $y(x)$ be the solution of this equation with the initial conditions $y(0)=0$, $y'(0)\neq 0$, and denote the nth zero of $y(x)$ by x_n, where $x_0=0$. Then from the result of the paper [3] and of Lemma I we obtain

$$n\pi-\frac{\pi}{2}<\int_0^{x_n}x^{\frac{1}{2\nu}}\,dx=\frac{2\nu}{2\nu+1}x_n^{1+\frac{1}{2\nu}}\leqq n\pi,$$

hence

$$A(\nu)\left(n-\frac{1}{2}\right)^2\pi^2<x_n\int_0^{x_n}x^{\frac{1}{\nu}}\,dx=\frac{\nu}{\nu+1}x_n^{2+\frac{1}{\nu}}\leqq A(\nu)n^2\pi^2,$$

therefore

$$\lim_{n\to\infty}\frac{1}{n^2\pi^2}x_n\int_0^{x_n}x^{\frac{1}{\nu}}\,dx=A(\nu),$$

i.e. the coefficient of $n^2\pi^2$ in (2) cannot be smaller than $A(\nu)$.

REFERENCES

[1] GALBRAITH A. S.: On the zeros of solutions of ordinary differential equations of second order, *Proc. Amer. Math. Soc.* **17** (1966) 333—337.

[2] MAKAI E.: Über die Nullstellen von Funktionen, die Lösungen Sturm-Liouville'scher Differentialgleichungen sind, *Comment. Math. Helv.* **16** (1943—44) 153—199.

[3] ELBERT Á.: On the zeros of solutions of ordinary differential equations of second order, *Publ. Math. Debrecen.* (1967) in the press.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received November 10, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 299—318.

INFORMATION-TYPE MEASURES OF DIFFERENCE OF PROBABILITY DISTRIBUTIONS AND INDIRECT OBSERVATIONS

by

I. CSISZÁR

This paper deals with the implications of sufficiency and ε-sufficiency of observation channels defined in terms of certain information-type measures of difference of probability distributions, called f-divergences. The results published here are taken from the author's thesis[1] [1]; except for Theorem 3. 1, which — although a direct consequence of the ideas developed in [1] — explicitely does not occur there. The results published earlier in [2] and [3] are considerably improved (the concept of f-divergence has been introduced in [2]). The first two sections are of introductory character; the main results are contained in § 3 while § 4 contains corollaries and comments.

§ 1. The Concept of f-divergence

Let $f(u)$ denote an arbitrary convex function defined in the interval $(0, +\infty)$. In order to avoid meaningless expressions in the sequel, let us agree in the following notational conventions:

$$f(0) = \lim_{u \to +0} f(u)$$

$$0 \cdot f\left(\frac{0}{0}\right) = 0 \tag{1.1}$$

$$0 \cdot f\left(\frac{a}{0}\right) = \lim_{\varepsilon \to +0} \varepsilon f\left(\frac{a}{\varepsilon}\right) = a \lim_{u \to +\infty} \frac{f(u)}{u} \qquad (0 < a < +\infty).$$

If $(X, \mathcal{X}, \lambda)$ is a measure space, i.e. X an arbitrary set, $\mathcal{X}$ a σ-algebra of subsets of X and λ a measure on $(X, \mathcal{X})$ (all measures will be assumed either finite or σ-finite), the symbol $[\lambda]$ will stand, as usual, for "almost everywhere with respect to λ."

LEMMA 1. 1. *Let $\alpha(x)$ and $\beta(x)$ be two nonnegative measurable functions on a measure space $(X, \mathcal{X}, \lambda)$; then $\int_E \beta(x) f\left(\frac{\alpha(x)}{\beta(x)}\right) \lambda(dx)$ is well-defined for all $E \in \mathcal{X}$*

[1] Other results of [1] — concerning topological properties of f-divergences — will be published in another paper.

on which both $\alpha(x)$ *and* $\beta(x)$ *are integrable, and for such an* E

$$\int_E \beta(x) f\left(\frac{\alpha(x)}{\beta(x)}\right)\lambda(dx) \geqq \int_E \beta(x)\lambda(dx)\cdot f\left(\frac{\int_E \alpha(x)\lambda(dx)}{\int_E \beta(x)\lambda(dx)}\right) > -\infty. \tag{1.2}$$

If $\int_E \beta(x)\lambda(dx) > 0$ *and* $f(u)$ *is strictly convex at* $u_0 = \int_E \alpha(x)\lambda(dx) \Big/ \int_E \beta(x)\lambda(dx)$, *there is strict inequality in* (1. 2), *except for the case*

$$\alpha(x) = u_0\beta(x) \ [\lambda] \quad \textit{on the set } E. \tag{1.3}$$

PROOF. One may assume $\int_E \beta(x)\lambda(dx) > 0$ and $\int_E \alpha(x)\lambda(dx) > 0$, since otherwise, according to the conventions (1. 1), the statement is trivially true. Let b denote the arithmetic mean of the left and right derivatives of $f(u)$ at the point $u_0 = \int_E \alpha(x)\lambda(dx) \Big/ \int_E \beta(x)\lambda(dx)$ $(0 < u_0 < +\infty)$. Then b is finite and the convexity of $f(u)$ implies

$$f(u) \geqq f(u_0) + b(u - u_0) \qquad (0 \leqq u < +\infty), \tag{1.4}$$

whence, substituting $u = \dfrac{\alpha(x)}{\beta(x)}$, we obtain for $\beta(x) > 0$

$$\beta(x) f\left(\frac{\alpha(x)}{\beta(x)}\right) \geqq \beta(x) f(u_0) + b\big(\alpha(x) - u_0\beta(x)\big). \tag{1.5}$$

According to (1. 1), relation (1. 5) clearly holds even for $\beta(x) = 0$, as the convexity of $f(u)$ implies $b \leqq \lim\limits_{u\to\infty} \dfrac{f(u)}{u}$. The right hand side of (1. 5) is integrable on E by assumption and thus the integral over E of the left hand side is also well-defined; since $\int_E \alpha(x)\lambda(dx) - u_0 \int_E \beta(x)\lambda(dx) = 0$ by the definition of u_0, (1. 5) gives rise to (1. 2). If $f(u)$ is strictly convex at $u = u_0$, the inequality in (1. 4) and (1. 5) is strict, except for $u = u_0$ and $\alpha(x) = u_0\beta(x)$, respectively. Thus also the condition (1. 3) of the equality in (1. 2) is proved.

Let now μ_1 and μ_2 be two probability distributions (i.e. measures with $\mu_1(X) = \mu_2(X) = 1$) on a measurable space $(X, \mathcal{X})$. Let λ denote an arbitrary dominating measure of μ_1 and μ_2, i.e. a finite or σ-finite measure on $(X, \mathcal{X})$ such that both μ_1 and μ_2 are absolutely continuous with respect to λ (one may choose e.g. $\lambda = \mu_1 + \mu_2$). Denote the density (Radon—Nikodym-derivative) of μ_i with respect to λ by $p_i(x)$:

$$p_i(x) = \frac{\mu_i(dx)}{\lambda(dx)} \qquad (i = 1, 2). \tag{1.6}$$

Integrals without specifying the domain of integration will always mean integrals over the whole space in question.

DEFINITION 1. 1. The quantity

$$\mathcal{I}_f(\mu_1, \mu_2) = \int p_2(x) f\left(\frac{p_1(x)}{p_2(x)}\right) \lambda(dx) \tag{1.7}$$

is called the f-divergence of the distributions μ_1 and μ_2.

By Lemma 1. 1, the integral (1. 7) is always well-defined and $\mathcal{I}_f(\mu_1, \mu_2) \geqq f(1)$ with equality only for $\mu_1 = \mu_2$, provided that $f(u)$ is strictly convex at $u_0 = 1$. It is also clear that the value of $\mathcal{I}_f(\mu_1, \mu_2)$ does not depend on the choice of λ.

The concept of f-divergence has been introduced in [2] as a generalisation of KULLBACK's "information for discrimination" or I-divergence[2] [4], [5] and of RÉNYI's "information gain" (I-divergence) of order α [9]; in fact, the I-divergence (of order 1) equals

$$I(\mu_1 \| \mu_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} \lambda(dx) = \mathcal{I}_{u \log u}(\mu_1, \mu_2) \tag{1.8}$$

while the I-divergence of order α

$$I_\alpha(\mu_1 \| \mu_2) = \frac{1}{\alpha - 1} \int p_1^\alpha(x) p_2^{1-\alpha}(x) \lambda(dx) \tag{1.9}$$

is a strictly monotone function of an f-divergence:

$$I_\alpha(\mu_1 \| \mu_2) = \frac{1}{\alpha - 1} \log |\mathcal{I}_{f_\alpha}(\mu_1, \mu_2)|; \quad f_\alpha(u) = u^\alpha \operatorname{sgn}(\alpha - 1). \tag{1.10}$$

Let us also mention, that the choice $f(u) = (u-1)^2$ yields again a familiar measure of difference of distributions, that may be called χ^2-divergence (having in mind the occurrence of its discrete special case in the χ^2-statistic):

$$\chi^2(\mu_1, \mu_2) = \int \frac{(p_1(x) - p_2(x))^2}{p_2(x)} \lambda(dx) = \mathcal{I}_{(u-1)^2}(\mu_1, \mu_2) \tag{1.11}$$

and that the variation distance $|\mu_1 - \mu_2| = \int |p_1(x) - p_2(x)| \lambda(dx)$ is equal to $\mathcal{I}_{|u-1|}(\mu_1, \mu_2)$.

As the minimum of $\mathcal{I}_f(\mu_1, \mu_2)$ equals $f(1)$, (attained only for $\mu_1 = \mu_2$, provided that $f(u)$ is strictly convex at $u_0 = 1$) it may seem convenient to require $f(1) = 0$ if we want to consider the f-divergence as a measure of difference of probability distributions; since, however, an additive constant is irrelevant in all the following considerations, I prefer to make no such restriction on $f(u)$.

The f-divergence is in general asymmetric in μ_1 and μ_2. Nevertheless, the convexity of $f(u)$ implies that of

$$f^*(u) = uf\left(\frac{1}{u}\right) \tag{1.12}$$

and with this function we have

$$\mathcal{I}_f(\mu_2, \mu_1) = \mathcal{I}_{f^*}(\mu_1, \mu_2). \tag{1.13}$$

[2] Often called also generalised entropy, as in [6], [7] and [8].

Hence follows, in particular, that the symmetrised f-divergence $\mathscr{I}_f(\mu_1, \mu_2) + \mathscr{I}_f(\mu_2, \mu_1)$ is again an f-divergence, with respect to the convex function $f(u) + f^*(u)$. E.g. for the so called J-divergence (symmetrised I-divergence) of order 1 we have

$$J(\mu_1, \mu_2) = I(\mu_1 \| \mu_2) + I(\mu_2 \| \mu_1) = \mathscr{I}_{(u-1)\log u}(\mu_1, \mu_2). \tag{1. 14}$$

The concept of f-divergence can be used also to define a class of measures of dependence of random variables. Let namely ξ and η be two random variables (in the most general sense) i.e. measurable mappings of a probability space $(\Omega, \mathscr{F}, \mathrm{P})$ into measurable spaces $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$, respectively, and consider the probability distributions $\mu_{\xi\eta}$ and $\mu_\xi \times \mu_\eta$ on the product space $(X \times Y, \mathscr{X} \times \mathscr{Y})$, where $\mu_{\xi\eta}$, μ_ξ and μ_η stand for the joint distribution of ξ and η and the marginal distributions, respectively. Then, intuitively, one can say that ξ and η are the less dependent, the closer are the distributions $\mu_{\xi\eta}$ and $\mu_\xi \times \mu_\eta$, thus $\mathscr{I}_f(\mu_{\xi\eta}, \mu_\xi \times \mu_\eta)$ can be considered as a measure of dependence of ξ and η.

A set of properties required of a measure of dependence (for real-valued random variables) has been proposed in [10]. The above quantity obviously possesses the properties A), B) and F) listed there (with Borel-measurable functions replaced by measurable mappings) and it is easy to transform it in such a way as to possess also the properties C), D)[3] and, if $\lim_{u\to\infty} \frac{f(u)}{u} = +\infty$, also E). For $f(u) = u \log u$ we obtain the mutual information of ξ and η:

$$I(\xi, \eta) = I(\mu_{\xi\eta} \| \mu_\xi \times \mu_\eta) = \mathscr{I}_{u \log u}(\mu_{\xi\eta}, \mu_\xi \times \mu_\eta) \tag{1. 15}$$

while $f(u) = (u-1)$ yields the mean square contingency:

$$\varphi^2(\xi, \eta) = \mathscr{I}_{(u-1)^2}(\mu_{\xi\eta}, \mu_\xi \times \mu_\eta). \tag{1. 16}$$

REMARK. In [10] the mean square contingency is defined only for $\mu_{\xi\eta} \ll \mu_\xi \times \mu_\eta$; in order that (1. 16) be generally valid, one has to extend the definition by the obvious setting $\varphi^2(\xi, \eta) = +\infty$ for $\mu_{\xi\eta} \not\ll \mu_\xi \times \mu_\eta$.

Of course, the f-divergences have unpleasent properties, too. The f-divergence is, in general, not even a quasimetric, i.e. it does not satisfy the triangle inequality; moreover, the "neighbourhoods" $U_f(\mu, \varepsilon) = \{\mu' : \mathscr{I}_f(\mu', \mu) - f(1) < \varepsilon\}$ need not define a topology in the space of distributions. As to the purposes of this paper, we need not worry about these embarrassing facts, present already in the case of the I-divergence; they will be considered in another paper.

§ 2. Indirect Observations

A measurable space $(X, \mathscr{X})$ can be interpreted as an experiment, the outcome x being governed by some probability distribution μ on $(X, \mathscr{X})$. In mathematical statistics the distribution μ is usually assumed to belong to some family $\{\mu_\theta\}_{\theta \in \Theta}$ of distributions, where θ is an unknown parameter.

[3] Provided that $f(u)$ is strictly convex at $u_0 = 1$.

DEFINITION 2. 1. By an *indirect observation* we shall mean a measurable space $(Y, \mathcal{Y})$ related to $(X, \mathcal{X})$ in one of the following ways, including also a correspondence of probability distributions μ on $(X, \mathcal{X})$ to distributions $\bar{\mu}$ on $(Y, \mathcal{Y})$:

(i) $Y = X$, $\mathcal{Y} \subset \mathcal{X}$, $\bar{\mu}$ = restriction of μ to $\mathcal{Y}$;

(ii) there is given a measurable mapping $y = Tx$ of $(X, \mathcal{X})$ into $(Y, \mathcal{Y})$; $\bar{\mu} = \mu T^{-1}$;

(iii) for each $x \in X$ there is given a probability measure $\nu(\cdot|x)$ on $(Y, \mathcal{Y})$ such that $\nu(B|\cdot)$ is $\mathcal{X}$-measurable for each fixed $B \in \mathcal{Y}$; $\bar{\mu}$ is defined by

$$\bar{\mu}(B) = \int \nu(B|x)\mu(dx). \tag{2. 1}$$

In mathematical statistics indirect observations of types (i) and (ii) are the most familiar (grouping of data, using a statistic); they are well known to be essentially identical (in case (i) the identity mapping can be introduced as a statistic, while in case (ii) $(Y, \mathcal{Y})$ can be replaced by the isomorphic measure algebra $(X, T^{-1}\mathcal{Y})$. In practice, however, most of the indirect observations are subject to random effects (e.g. measurement errors), in which case (iii) is the appropriate model. In terms of information theory, one can speak of an *observation channel* (X, ν, Y), the outcome y of the indirect observation being interpreted as the output of the channel (cf. [7]).

Clearly, indirect observations of type (ii) — and thus, essentially, of type (i) too — are special cases of the more general model (iii); one has only to set $\nu(B|x) = 1$ for $Tx \in B$ and 0 otherwise. On the other hand, (iii) can be reduced to (i) if the space $(X, \mathcal{X})$ is replaced by the product space $(X \times Y, \mathcal{X} \times \mathcal{Y})$. Indeed, to each distribution (and, in general, to each σ-finite measure) μ on $(X, \mathcal{X})$ there corresponds a measure μ^* on $(X \times Y, \mathcal{X} \times \mathcal{Y})$ defined by

$$\mu^*(A \times B) = \int_A \nu(B|x)\mu(dx) \qquad (A \in \mathcal{X},\ B \in \mathcal{Y}) \tag{2. 2}$$

and the routine extension to the whole $\mathcal{X} \times \mathcal{Y}$; the restriction of μ^* to the σ-algebra

$$\mathcal{Y}' = \{X \times B\colon B \in \mathcal{Y}\} \tag{2. 3}$$

is essentially identical with $\bar{\mu}$ as defined by (2. 1) (in the sense of the isomorphism of the corresponding measure algebras).

The indirect observation will be called *sufficient* with respect to a family $\{\mu_\theta\}_{\theta \in \Theta}$ of distributions on $(X, \mathcal{X})$ if the conditional distribution given the outcome of the indirect observation does not depend on θ; restated more formally:

DEFINITION 2. 2. An indirect observation of type (i), (ii) or (iii) is sufficient with respect to a family $\{\mu_\theta\}_{\theta \in \Theta}$ of distributions on $(X, \mathcal{X})$ if for each $A \in \mathcal{X}$ there exists a $\mathcal{Y}$-measurable function $\pi_A(y)$ on Y such that for every $B \in \mathcal{Y}$ and $\theta \in \Theta$

$$\int \pi_A(y)\bar{\mu}_\theta(dy) = \mathsf{P}_\theta(x \in A, y \in B), \tag{2. 4}$$

where $\mathsf{P}_\theta(x \in A, y \in B)$ stands for $\mu_\theta(A \cap B)$, $\mu_\theta(A \cap T^{-1}B)$ or $\mu_\theta^*(A \times B)$, respectively.

This definition can be considered as a special case of the more general concept of sufficiency of an experiment $(Y, \mathcal{Y})$ with respect to another experiment $(X, \mathcal{X})$, introduced by BLACKWELL [11]; in cases (i) and (ii) it reduces to the familiar concept

of *sufficient σ-algebras* and *statistics*, respectively. In case (iii) we shall speak of *sufficient channels*.

In the sequel we shall concentrate mainly on the case $\Theta=\{1, 2\}$, i.e. in terms of mathematical statistics, the problem of discrimination between two simple hypotheses.

Let μ_1 and μ_2 be two probability distributions on $(X, \mathcal{X})$. Let $\mathcal{X}_0$ be a sub-σ-algebra of $\mathcal{X}$ and $\bar{\mu}_i$ the restriction of μ_i to $\mathcal{X}_0$ $(i=1, 2)$; let $C\in\mathcal{X}_0$ denote a set with $\bar{\mu}_2(C)=0$ such that $\bar{\mu}_1\ll\bar{\mu}_2$ on $X-C$ (if $\mu_1\ll\mu_2$, we choose $C=\emptyset$). We introduce an auxiliary distribution μ_{12} defined by

$$\mu_{12}(A)=\int_{X-C}\mu_2(A|\mathcal{X}_0)\mu_1(dx)+\mu_1(A\cap C) \qquad (A\in\mathcal{X}), \tag{2. 5}$$

where $\mu_2(A|\mathcal{X}_0)$ denotes the conditional μ_2-probability of A with respect to the σ-algebra $\mathcal{X}_0$. It is clear from the properties of conditional probabilities, that μ_{12} is uniquely defined and it is in fact a probability distribution on $(X, \mathcal{X})$. For $A\in\mathcal{X}_0$ (2. 5) implies $\mu_{12}(A)=\mu_1(A)$; thus μ_{12} is an extension of $\bar{\mu}_1$ to $\mathcal{X}$, in general different from μ_1.

LEMMA 2. 1. *$\mathcal{X}_0\subset\mathcal{X}$ is a sufficient σ-algebra with respect to the pair (μ_1, μ_2) if and only if $\mu_{12}=\mu_1$.*

PROOF. The statement follows directly from definition 2. 2 and (2. 5).

Let $p_i(x)$ and $\bar{p}_i(x)$ denote the density of μ_i and $\bar{\mu}_i$ $(i=1, 2)$ with respect to some dominating measure λ and its restriction $\bar{\lambda}$ to $\mathcal{X}_0$, respectively. ($\bar{\lambda}$ is also assumed to be σ-finite).

LEMMA 2. 2. *μ_{12} is also dominated by λ and its density function equals*

$$p_{12}(x)=\begin{cases}\dfrac{\bar{p}_1(x)}{\bar{p}_2(x)}p_2(x) & \text{if} \quad \bar{p}_2(x)>0\\ p_1(x) & \text{if} \quad \bar{p}_2(x)=0.\end{cases} \tag{2. 6}$$

PROOF. The absolute continuity of μ_{12} with respect to λ is obvious. To prove (2. 6) we may and do assume that λ is a probability measure. We set $C=\{x: \bar{p}_2(x)=0\}$ and denote the indicator function of a set $A\subset X$ by $\chi_A(x)$. Utilizing the well-known properties of conditional expectations, in particular that for any $\mathcal{X}_0$-measurable function h $\mathsf{E}_\lambda(gh|\mathcal{X}_0)=h\mathsf{E}_\lambda(g|\mathcal{X}_0)$, we obtain for arbitrary $A\in\mathcal{X}$

$$\int_A p_{12}(x)\lambda(dx)=\mathsf{E}_\lambda\left(\frac{\bar{p}_1}{\bar{p}_2}p_2\chi_{A-C}\right)+\mathsf{E}_\lambda(p_1\chi_{A\cap C})=\mathsf{E}_\lambda\left(\frac{\bar{p}_1}{\bar{p}_2}\chi_{X-C}\mathsf{E}_\lambda(\chi_A p_2|\mathcal{X}_0)\right)+$$

$$+\mu_1(A\cap C)=\mathsf{E}_\lambda\left(\frac{\bar{p}_1}{\bar{p}_2}\chi_{X-C}\mu_2(A|\mathcal{X}_0)\bar{p}_2\right)+\mu_1(A\cap C)= \tag{2. 7}$$

$$=\int_{X-C}\mu_2(A|\mathcal{X}_0)\bar{p}_1\lambda(dx)+\mu_1(A\cap C)=\mu_{12}(A).$$

This means that (2. 6) is indeed the density function of μ_{12}.

LEMMA 2.3. *The σ-algebra $\mathscr{X}_0 \subset \mathscr{X}$ is sufficient with respect to the pair (μ_1, μ_2) if and only if $p_1(x)/p_2(x)$ equals almost everywhere an $\mathscr{X}_0$-measurable function, both with respect to μ_1 and μ_2 (here we understand $\frac{a}{0} = +\infty$ $(a>0)$ but $\frac{0}{0}$ remains undefined)*; *moreover in this case* $\frac{p_1(x)}{p_2(x)} = \frac{\bar{p}_1(x)}{\bar{p}_2(x)}$ $[\mu_1+\mu_2]$.

PROOF.[4] If $\mathscr{X}_0$ is sufficient, from Lemmas 2. 1 and 2. 2 follows $\frac{p_1(x)}{p_2(x)} = \frac{\bar{p}_1(x)}{\bar{p}_2(x)}$ $[\mu_1+\mu_2]$, the exceptional set being $\{x: p_1(x)=\bar{p}_2(x)=0\}$ and, possibly, some λ-null-set. On the other hand, if $p_1(x)/p_2(x)$ is $[\mu_1+\mu_2]$ equal to an $\mathscr{X}_0$-measurable function, the μ_2-null-set $D = \{x: \frac{p_1(x)}{p_2(x)} = +\infty\}$ differs at most by a $(\mu_1+\mu_2)$-null-set from a set in $\mathscr{X}_0$. Hence follows that D is contained in $C = \{x: \bar{p}_2(x)=0\} \in \mathscr{X}_0$ up to a $(\mu_1+\mu_2)$-null-set, and thus on $X-C$ we have $\mu_1 \ll \mu_2$, $\frac{\mu_1(dx)}{\mu_2(dx)} = \frac{p_1(x)}{p_2(x)}$ $[\mu_2]$. (Of course, μ_1 need not be a probability measure on $X-C$). Since, by assumption, $p_1(x)/p_2(x)$ is $[\mu_1+\mu_2]$ equal to an $\mathscr{X}_0$-measurable function and on $X-C$ $\frac{\bar{\mu}_1(dx)}{\bar{\mu}_2(dx)} = \frac{\bar{p}_1(x)}{\bar{p}_2(x)}$ the relation

$$\mathsf{E}_\mu\left(\frac{\nu(dx)}{\mu(dx)} \,\middle|\, \mathscr{X}_0\right) = \frac{\bar{\nu}(dx)}{\bar{\mu}(dx)} \tag{2.8}$$

valid by the definition of conditional expectations for any $\nu \ll \mu$ (ν need not be a probability measure), implies

$$\frac{p_1(x)}{p_2(x)} = \frac{\bar{p}_1(x)}{\bar{p}_2(x)} \quad [\mu_2] \quad \text{on} \quad X-C. \tag{2.9}$$

From (2. 8) and $\mu_1 \ll \mu_2$ on $X-C$ the sufficiency of $\mathscr{X}_0$ follows by Lemmas 2. 1 and 2. 2.

Let us now be given an observation channel (X, ν, Y) and let μ_1 and μ_2 be two probability distributions on $(X, \mathscr{X})$. Consider the product space $(X \times Y, \mathscr{X} \times \mathscr{Y})$ and the distributions μ_1^* and μ_2^* on it, defined according to (2. 2). Let λ be a dominating measure of μ_1 and μ_2; then the corresponding λ^* on the product space is also σ-finite.

LEMMA 2. 4. *The density of μ_i^* with respect to λ^* equals the density of μ_i with respect to λ*:

$$p_i^*(x, y) = p_i(x) \quad [\lambda^*] \qquad (i=1, 2). \tag{2.10}$$

[4] Lemma 2. 3 is a variant of the well known factorisation criterion of sufficiency; the proof — though presumably new, at least in form — is given mainly for the sake of completeness (in lack of a direct reference for the required form of the statement).

PROOF. It suffices to prove $\int_E p_i(x)\lambda^*(dx, dy) = \mu_i^*(E)$ only for $E = A \times B$ ($A \in \mathscr{X}$, $B \in \mathscr{Y}$). But this follows directly from the definition of λ^* and μ^* (cf. (2. 2)):

$$\int_{A\times B} p_i(x)\lambda^*(dx, dy) = \int_A p_i(x)\nu(B|x)\lambda(dx) = \int_A \nu(B|x)\mu_i(dx) = \mu_i^*(A \times B).$$

LEMMA 2. 5. *The channel* (X, ν, Y) *is sufficient with respect to the pair* (μ_1, μ_2) *if and only if the* σ*-algebra* $\mathscr{Y}' = \{X \times B : B \in \mathscr{Y}\}$ *is sufficient with respect to the pair* (μ_1^*, μ_2^*).

PROOF. The „if" part is trivial; to prove the „only if" part, we have to show (cf. Definition 2. 2) that for every $E \in \mathscr{X} \times \mathscr{Y}$ there exists a $\mathscr{Y}'$-measurable function $\pi_E^*(x, y)$ i.e. a $\mathscr{Y}$-measurable $\pi_E^*(y)$ such that

$$\int_{X\times B} \pi_E^*(y)\mu_i^*(dx, dy) = \mu_i^*\big(E \cap (X \times B)\big) \qquad (i = 1, 2) \tag{2. 11}$$

for each $B \in \mathscr{Y}$.

If $E = E_1 \times E_2$, the function $\pi_E^*(y) = \pi_{E_1}(y)\chi_{E_2}(y)$ obviously possesses this property, where $\pi_{E_1}(y)$ denotes a function, existing by assumption, such that

$$\int_B \pi_{E_1}(y)\bar{\mu}_i(dy) = \mu_i^*(E_1 \times B) \quad (i = 1, 2) \quad \text{for all} \quad B \in \mathscr{Y}.$$

This already proves the lemma, as the class of sets $E \subset X \times Y$ for which $\pi_E^*(y)$ satisfying (2. 11) exists is obviously a σ-algebra.

We define the auxiliary distribution μ_{12}^* on $(X \times Y, \mathscr{X} \times \mathscr{Y})$ by a transliteration of (2. 5) to the present case. Then, as it is easy to see, μ_{12}^* is defined by

$$\mu_{12}^*(A \times B) = \int_{B-C} \mu_2(A|y)\bar{\mu}_1(dy) + \mu_1^*\big(A \times (B \cap C)\big) \tag{2. 12}$$

(where $\bar{\mu}_i$ $(i = 1, 2)$ is defined by (2. 1) and $C \in \mathscr{Y}$ stands for a set such that $\bar{\mu}_2(C) = 0$ and on $Y - C$ $\bar{\mu}_1 \ll \bar{\mu}_2$) and the usual extension to the whole $\mathscr{X} \times \mathscr{Y}$. Here $\mu_2(A|y)$ is an abbreviation for the conditional probability $\mu_2^*(A \times Y|\mathscr{Y}')$. The density of μ_{12}^* (with respect to λ^* connected with λ according to (2. 2)) is given, according to lemmas 2. 2 and 2. 4 by

$$p_{12}^*(x, y) = \begin{cases} \dfrac{\bar{p}_1(y)}{\bar{p}_2(y)} p_2(x) & \text{if} \quad \bar{p}_2(y) > 0 \\ p_1(x) & \text{if} \quad \bar{p}_2(y) = 0, \end{cases} \tag{2. 13}$$

where $\bar{p}_i(y)$ denotes the density of $\bar{\mu}_i$ with respect to $\bar{\lambda}$, defined[5] according to (2. 1). The following characterisation of sufficient channels is also of some independent interest.

LEMMA 2. 6. *The channel* (X, ν, Y) *is sufficient with respect to the pair* (μ_1, μ_2) *if and only if the output space* Y *can be represented in the form*

$$Y = \bigcup_{0 \le s \le +\infty} B_s \quad (B_{s_1} \cap B_{s_2} = \emptyset \quad \textit{for} \quad s_1 \neq s_2) \tag{2. 14}$$

[5] We consider only such σ-finite λ's for which $\bar{\lambda}$, too, is σ-finite.

where for every $0 \leqq a \leqq b \leqq +\infty$ $\bigcup_{a \leqq s \leqq b} B_s \in \mathscr{Y}$ *and* $\frac{p_1(x)}{p_2(x)} = s$ *implies* $\nu(B_s|x) = 1$ $[\mu_1 + \mu_2]$ $\left(\frac{a}{0} = +\infty \text{ if } a > 0\right)$. *Here one may choose*

$$B_s = \left\{y : \frac{\bar{p}_1(y)}{\bar{p}_2(y)} = s\right\} \quad (s > 0), \quad B_0 = \{y : \bar{p}_1(y) = 0\}. \tag{2.15}$$

PROOF. We set

$$\begin{aligned} A_k^n &= \left\{x : \frac{k-1}{2^n} p_2(x) \leqq p_1(x) < \frac{k}{2^n} p_2(x)\right\} \\ B_k^n &= \left\{y : \frac{k-1}{2^n} \bar{p}_2(y) \leqq \bar{p}_1(y) < \frac{k}{2^n} \bar{p}_2(y)\right\} \end{aligned} \qquad \begin{pmatrix} k = 1, 2, \ldots \\ n = 1, 2, \ldots \end{pmatrix} \tag{2.16}$$

and

$$\begin{aligned} A_\infty &= \{x : p_1(x) > p_2(x) = 0\} \\ B_\infty &= \{y : \bar{p}_1(y) > \bar{p}_2(y) = 0\}. \end{aligned} \tag{2.17}$$

If the channel is sufficient with respect to (μ_1, μ_2), we have, according to (2. 2) and (2. 4), for each k and n

$$\frac{k-1}{2^n} \mu_2^*(A_k^n \times B_l^n) \leqq \mu_1^*(A_k^n \times B_l^n) \leqq \frac{k}{2^n} \mu_2^*(A_k^n \times B_l^n) \tag{2.18}$$

$$\frac{l-1}{2^n} \mu_2^*(A_k^n \times B_l^n) \leqq \mu_1^*(A_k^n \times B_l^n) \leqq \frac{l}{2^n} \mu_2^*(A_k^n \times B_l^n); \tag{2.19}$$

both in (2. 18) and (2. 19) the second inequality is strict except for $\mu_i^*(A_k^n \times B_l^n) = 0$ $(i = 1, 2)$. (2. 18) and (2. 19) imply $\mu_i^*(A_k^n \times B_l^n) = 0$ $(i = 1, 2)$ for $k \neq l$, and by a similar reasoning, we also have $\mu_i^*(A_\infty \times B_l^n) = \mu_i^*(A_k^n \times B_\infty) = 0$ $(i = 1, 2)$. This means, by (2. 2), that $x \in A_k^n$ implies $\nu(B_k^n|x) = 1$ $[\mu_1 + \mu_2]$ for all n and k, including also $k = +\infty$ if we write $A_\infty^n = A_\infty$, $B_\infty^n = B_\infty$ for all n. Hence follows, if we set in accordance with (2. 15)

$$B_s = \bigcap_{n=1}^{\infty} B_{[2^n s]+1}^n \quad (0 < s \leqq +\infty), \quad B_0 = Y - \bigcup_{s>0} B_s, \tag{2.20}$$

that for $\frac{p_1(x)}{p_2(x)} = s$ one has $\nu(B_s|x) = 1$ $[\mu_1 + \mu_2]$, proving the "only if" part of the statement. On the other hand, if Y can be represented in the form (2. 14), let $s(y)$ denote the value s for which $y \in B_s$. Then $s(y)$ is $\mathscr{Y}$-measurable and

$$s(y) = \frac{p_1(x)}{p_2(x)} \quad [\mu_1^* + \mu_2^*], \tag{2.21}$$

thus, utilising Lemmas 2. 3, 2. 4 and 2. 5, the channel (X, ν, Y) is sufficient.

§ 3. (f, ε)-sufficiency of Indirect Observations and its Implications

It is intuitively clear, that in indirect observations the discernability of two probability distributions cannot increase. This means, that if the f-divergence is an appropriate measure of difference of probability distributions, the inequality

$$\mathscr{I}_f(\overline{\mu}_1, \overline{\mu}_2) \leqq \mathscr{I}_f(\mu_1, \mu_2) \tag{3. 1}$$

has to be valid for any pair of distributions on $(X, \mathscr{X})$. It has been shown in [2] that this inequality holds, indeed, and for strictly convex f and finite $\mathscr{I}_f(\overline{\mu}_1, \overline{\mu}_2)$ the condition of equality in (3. 1) is just the sufficiency[6] of the indirect observation with respect to the pair (μ_1, μ_2). (The proof will be reproduced in the sequel.) This theorem is a generalisation of the theorem of KULLBACK and LEIBLER [4] stating that for indirect observations of type (i) or (ii) (Definition 2. 1) the I-divergence does not increase, i.e.

$$I(\overline{\mu}_1 \| \overline{\mu}_2) \leqq I(\mu_1 \| \mu_2) \tag{3. 2}$$

with equality only for sufficient σ-algebras or statistics, respectively (provided that the I-divergences in question are finite). The difference of the two sides in (3. 2) can be interpreted as the loss of information in the indirect observation and it provides a measure of degree of non-sufficiency of the indirect observation, c.f. [12], [13].

Indirect observations of type (iii) do not seem to have been considered in this context; of course, as it has been pointed out in § 2, this is rather a question of interpretation, though to recognize the fact that the theory covers also indirect observations of type (iii) seems to be essential at least from the point of view of applications. From the pure mathematical point of view, the generalisation from I-divergences to arbitrary f-divergences is more important; it turns out that the convexity of f alone leads to interesting results. It should be noted however, that the restriction to $f(u) = u \log u$ (i.e. I-divergences) does have certain advantages, e.g. that the "loss of information" can be expressed again as an I-divergence, namely[7], with the notation (2. 5),

$$I(\mu_1 \| \mu_2) - I(\overline{\mu}_1 \| \overline{\mu}_2) = I(\mu_1 \| \mu_{12}); \tag{3. 3}$$

obviously no similar equation can be hoped for in the general case.

DEFINITION 3. 1. An indirect observation (of any of the types (i), (ii), and (iii)) will be called (f, ε)-sufficient with respect to the pair of distributions (μ_1, μ_2) if $\mathscr{I}_f(\overline{\mu}_1, \overline{\mu}_2) < +\infty$ and

$$\mathscr{I}_f(\mu_1, \mu_2) - \mathscr{I}_f(\overline{\mu}_1, \overline{\mu}_2) \leqq \varepsilon \qquad (\varepsilon \geqq 0). \tag{3. 4}$$

REMARK. The term „ε-sufficiency" has been introduced by A. PÉREZ, in connection with decision problems (cf. e.g. [13]). Here we use the term in a somewhat different context, but the idea is the same.

[6] In [2] the concept of sufficient channels did not occur but the condition of equality given there is just that of the sufficiency.

[7] Equation (3. 3) is an immediate consequence of (1. 8) and Lemma 2. 2.

According to the result of [2], referred to above, for strictly convex f (f, 0)-sufficiency is equivalent to the usual sufficiency. Our aim is to deduce estimates for the deviation of μ_1 and the auxiliary measure μ_{12} (or of μ_1^* and μ_{12}^*) in case of ε-sufficiency.

First we consider indirect observations of type (i).

THEOREM 3. 1.[8] *Let E be a compact subset of the interval $(0, +\infty)$ such that $f(u)$ is strictly convex at each point of E. Then there exists a positive function $\varphi(v)$ $(0<v<+\infty)$ depending only on f and E, with $\varphi(0)=\lim_{v\to+0}\varphi(v)=0$ such that*

$$|\mu_1-\mu_{12}|(A)\leqq\varphi(\mathscr{I}_f(\mu_1,\mu_2)-\mathscr{I}_f(\bar{\mu}_1,\bar{\mu}_2)), \tag{3. 5}$$

provided that the right hand side is meaningful, where

$$A=\left\{x:\frac{\bar{p}_1(x)}{\bar{p}_2(x)}\in E\right\}. \tag{3. 6}$$

PROOF. We start from the inequality (1. 4) valid for all $0\leqq u<+\infty$ and $0<u_0<+\infty$ (with b depending on u_0). The condition that $f(u)$ is strictly convex at the point u_0 means that for $|u-u_0|\geqq\varepsilon$ there holds even a stronger inequality, namely

$$f(u)\geqq f(u_0)+b(u-u_0)+\varepsilon_1|u-u_0|, \tag{3. 7}$$

where ε_1 is a sufficiently small positive number (depending on u_0 and ε). If $f(u)$ is strictly convex at each point of the compact set E, it is easily seen that $\varepsilon_1=\psi(\varepsilon)>0$ can be chosen uniformly for $u_0\in E$ and (3. 7) gives rise to

$$f(u)\geqq f(u_0)+b(u_0)(u-u_0)+\psi(\varepsilon)|u-u_0|(1-\chi_{u_0,\varepsilon}(u))\chi_E(u_0)\qquad(u_0>0); \tag{3. 8}$$

here $\chi_{u_0,\varepsilon}$ and χ_E stand for the indicator function of the interval $(u_0-\varepsilon, u_0+\varepsilon)$ and of E, respectively. Substituting $u=\dfrac{p_1(x)}{p_2(x)}$, $u_0=\dfrac{\bar{p}_1(x)}{\bar{p}_2(x)}$ and multiplying formally by $p_2(x)$ we obtain an inequality valid everywhere on the set $X_0=\{x:\bar{p}_1(x)\bar{p}_2(x)>0\}$; integrating with respect to λ over X_0, we obtain

$$\begin{aligned}\int_{X_0}p_2(x)f\left(\frac{p_1(x)}{p_2(x)}\right)\lambda(dx)\geqq\\ \geqq\int_{X_0}\bar{p}_2(x)f\left(\frac{\bar{p}_1(x)}{\bar{p}_2(x)}\right)\lambda(dx)+\psi(\varepsilon)\left(\int_A|p_1(x)-p_{12}(x)|\lambda(dx)-\varepsilon\right).\end{aligned} \tag{3. 9}$$

In fact, Lemma 2. 2 implies that on X_0

$$\left(\frac{p_1(x)}{p_2(x)}-\frac{\bar{p}_1(x)}{\bar{p}_2(x)}\right)p_2(x)=p_1(x)-p_{12}(x)\,[\lambda];$$

[8] $|\mu_1-\mu_{12}|(A)$ stands for the total variation of $\mu_1-\mu_{12}$ on the set A, i. e.

$$|\mu_1-\mu_{12}|(A)=\int_A|p_1(x)-p_{12}(x)|\lambda(dx).$$

now the vanishing of the integral of the second term follows from the $\mathscr{X}_0$-measurability of $b\left(\frac{\bar{p}_1(x)}{\bar{p}_2(x)}\right)$ and the fact that on $\mathscr{X}_0$ the measures μ_1 and μ_{12} coincide, while the estimate in brackets of the integral of the third term follows directly from the definition of the set A and of $\chi_{u_0,\varepsilon}$.

We also have for $X_1=\{x\colon \bar{p}_1(x)=0\}$, where also $p_1(x)=0$ $[\lambda]$,

$$\int_{X_1} p_2(x) f\left(\frac{p_1(x)}{p_2(x)}\right)\lambda(dx)=\int_{X_1}\bar{p}_2(x) f\left(\frac{\bar{p}_1(x)}{\bar{p}_2(x)}\right)\lambda(dx)=\mu_2(X_1)\cdot f(0); \tag{3.10}$$

further, by Lemma 1. 1 and the convention (1. 1)

$$\int_{X_2} p_2(x) f\left(\frac{p_1(x)}{p_2(x)}\right)\lambda(dx)\geqq\int_{X_2} p_2(x)\lambda(dx)\cdot f\left(\frac{\int_{X_2} p_1(x)\lambda(dx)}{\int_{X_2} p_2(x)\lambda(dx)}\right)= \tag{3.11}$$

$$=\mu_1(X_2)\lim_{u\to+\infty}\frac{f(u)}{u}=\int_{X_2}\bar{p}_2(x) f\left(\frac{\bar{p}_1(x)}{\bar{p}_2(x)}\right),$$

where $X_2=\{x\colon \bar{p}_1(x)>\bar{p}_2(x)=0\}=X-X_0-X_1$. From (3. 9), (3. 10) and (3. 11) follows

$$\psi(\varepsilon)\big(|\mu_1-\mu_{12}|(A)-\varepsilon\big)\leqq\mathscr{I}_f(\mu_1,\mu_2)-\mathscr{I}_f(\bar{\mu}_1,\bar{\mu}_2) \tag{3.12}$$

and hence, choosing to given $0<v<+\infty$ such an $\varepsilon=\varepsilon(v)>0$ that for $v\to 0$ $\varepsilon(v)+\frac{\varepsilon}{\varepsilon(v)}=\varphi(v)\to 0$, we obtain (3. 5). The proof is complete.

Thus, under the conditions of Theorem 3. 1, (f,ε)-sufficiency (definition 3. 1) implies $|\mu_1-\mu_{12}|(A)\leqq\varphi(\varepsilon)$. The result that $(f,0)$-sufficiency is equivalent (for strictly convex f) to the usual sufficinency is a direct corollary of Theorem 3. 1. In fact, $(f,0)$-sufficiency implies $|\mu_1-\mu_{12}|(A_n)=0$ for $A_n=\left\{x\colon \frac{1}{n}\leqq\frac{\bar{p}_1(x)}{\bar{p}_2(x)}\leqq n\right\}$, $n=1,2,\ldots$; for $A_0=\{x\colon \bar{p}_1(x)\bar{p}_2(x)=0\}$ we also have $|\mu_1-\mu_{12}|(A)=0$ as it follows at once from Lemma 2. 2. But $X=\bigcup_{i=0}^{\infty}A_i$, thus $|\mu_1-\mu_{12}|=0$ and $\mathscr{X}_0$ is sufficient by Lemma 2. 1. The converse is obvious (Lemma 2. 3).

REMARK. If $\frac{\bar{p}_1(x)}{\bar{p}_2(x)}\in E$ $[\mu_1+\mu_2]$, (3. 5) can be rewritten as $|\mu_1-\mu_{12}|\leqq\leqq\varphi\big(\mathscr{I}_f(\mu_1,\mu_2)-\mathscr{I}_f(\bar{\mu}_1,\bar{\mu}_2)\big)$; under appropriate restrictions of $f(u)$, we shall obtain an estimate of this type also for the general case (see Theorem 3. 3).

THEOREM 3. 2. *Let E be a Borel subset of the interval $[0,+\infty)$ such that in the neighbourhood of radius r_0 of each point of E the function $f(u)$ is twice differentiable and $f''(u)\geqq a>0$. Then*

$$|\mu_1-\mu_{12}|(A)\leqq\sqrt{\frac{8}{a}}\sqrt{\mathscr{I}_f(\mu_1,\mu_2)-\mathscr{I}_f(\bar{\mu}_1,\bar{\mu}_2)} \tag{3.13}$$

provided that $\mathscr{I}_f(\mu_1,\mu_2)-\mathscr{I}_f(\bar{\mu}_1,\bar{\mu}_2)\leqq\frac{1}{2}ar_0^2$, where A is defined in the same way as in (3. 6).

PROOF. From the assumption follows that in (3. 7) one may choose $\varepsilon_1 = \psi(\varepsilon) = \frac{a}{2}\varepsilon$ for $\varepsilon \leqq r_0$. Then, setting $\varepsilon = \sqrt{\frac{2\delta}{a}}$, $\delta = \mathscr{I}_f(\mu_1, \mu_2) - \mathscr{I}_f(\bar{\mu}_1, \bar{\mu}_2)$, from (3. 12) we obtain the desired estimate.

Let us remark that there are more general inequalities underlying to Theorems 3. 1 and 3. 2. We formulate the one corresponding to Theorem 3. 2 as

LEMMA 3. 1. *Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, $\mathscr{F}_0$ a sub- σ-algebra of $\mathscr{F}$ and ξ a random variable on $(\Omega, \mathscr{F}, \mathsf{P})$ with finite expectation. Let $f(u)$ be a convex function defined in a real interval I[9] that contains the range of ξ and let B be a Borel subset of I such that in the r_0-neighbourhood of each point of B $f(u)$ is twice differentiable and $f''(u) \geqq a > 0$. Then*

$$\int_{\mathsf{E}(\xi|\mathscr{F}_0) \in B} |\xi - \mathsf{E}(\xi|\mathscr{F}_0)| \, \mathsf{P}(d\omega) \leqq \sqrt{\frac{8}{a}} \sqrt{\mathsf{E}(f(\xi) - f(\mathsf{E}(\xi|\mathscr{F}_0)))} \tag{3.14}$$

provided that $\mathsf{E}f(\xi) - \mathsf{E}f(E(\xi|\mathscr{F}_0)) \leqq \frac{1}{2} a r_0^2$.

The proof is the same as that of Theorems 3. 1 and 3. 2 except that in (3. 8) we have to substitute $u = \xi$, $u_0 = \mathsf{E}(\xi|\mathscr{F}_0)$, and to show that the integral of the second term vanishes we have to use the well-known properties of conditional expectations.

Lemma 3. 1 can be considered as the generalisation of the lemma of [3] on the "stability of Jensen's inequality" for conditional expectations, announced there.

It should be noted, however, that a direct application of Lemma 3. 1 to obtain Theorem 3. 2 is possible only in the case $\mu_1 \ll \mu_2$, when one may set $(\Omega, \mathscr{F}, \mathsf{P}) = (X, \mathscr{X}, \mu_2)$, $\mathscr{F}_0 = \mathscr{X}_0$, $\xi(\omega) = \frac{p_1(x)}{p_2(x)}$, $\mathsf{E}(\xi|\mathscr{F}_0) = \frac{\bar{p}_1(x)}{\bar{p}_2(x)}$ (cf. (2. 8)).

The following variant of Lemma 3. 1 will be also useful.

LEMMA 3. 2. *Under the conditions of Lemma 3. 1, if $f''(u) \geqq \min\left(a, \frac{b}{|u|}\right)$ with some positive constants a and b, then*

$$\mathsf{E}|\xi - \mathsf{E}(\xi|\mathscr{F}_0)| \leqq \sqrt{\frac{8}{a} + \frac{16\mathsf{E}|\xi|}{b}} \sqrt{\mathsf{E}(f(\xi) - f(\mathsf{E}(\xi|\mathscr{F}_0)))} \tag{3.15}$$

provided that the right hand side is meaningful.

PROOF. We split Ω into the disjoint subsets

$$\Omega_n = \left\{\omega : (n-1)\frac{b}{2a} \leqq \mathsf{E}(|\xi||\mathscr{F}_0) < n\frac{b}{2a}\right\} \qquad (n = 1, 2, \ldots)$$

and apply Lemma 3. 1 to the conditional probabilities with respect to the events Ω_n. In this way we get for $\mathsf{P}(\Omega_n) > 0$

$$\mathsf{E}(|\xi - \mathsf{E}(\xi|\mathscr{F}_0)||\Omega_n) \leqq \sqrt{\frac{8n}{a}} \sqrt{\mathsf{E}(f(\xi) - f(\mathsf{E}(\xi|\mathscr{F}_0))|\Omega_n)}\,; \tag{3.16}$$

[9] Unlike to the rest of the paper, in Lemmas 3. 1 and 3. 2 we do not assume that $I = (0, +\infty)$.

here we used the fact, that on the set Ω_n the conditional expectation with respect to $\mathscr{F}_0$ under the conditional probability measure with respect to Ω_n is the same as under the original P, further that in the neighbourhood of radius $r_n = n\frac{b}{2a}$ of each point of the interval $\left(-n\frac{b}{2a}, n\frac{b}{2a}\right)$ we have $f''(u) \geqq \frac{a}{n}$ and also that

$$\mathsf{E}(|\xi - \mathsf{E}(\xi|\mathscr{F}_0)||\Omega_n) = \frac{1}{\mathsf{P}(\Omega_n)} \int_{\Omega_n} |\xi - \mathsf{E}(\xi|\mathscr{F}_0)| \mathsf{P}(d\omega) \leqq$$

$$\leqq \frac{1}{\mathsf{P}(\Omega_n)} \cdot 2 \int_{\Omega_n} \mathsf{E}(|\xi||\mathscr{F}_0) \mathsf{P}(d\omega) \leqq n\frac{b}{a} \tag{3.17}$$

in any case, thus (3. 16) holds also for

$$\mathsf{E}(f(\xi) - f(\mathsf{E}(\xi|\mathscr{F}_0))|\Omega_n) > \frac{1}{2} \cdot \frac{a}{n} \cdot r_n^2 = \frac{nb^2}{8a}.$$

From (3. 16) follows by the Cauchy inequality ($\sum_n'$ denotes summation for the indices n with $\mathsf{P}(\Omega_n) > 0$)

$$\mathsf{E}|\xi - \mathsf{E}(\xi|\mathscr{F}_0)| = \sum_n' \mathsf{P}(\Omega_n) \mathsf{E}(|\xi - \mathsf{E}(\xi|\mathscr{F}_0)||\Omega_n) \leqq$$

$$\leqq \sqrt{\frac{8}{a}} \sum_n' \mathsf{P}(\Omega_n) \sqrt{n\mathsf{E}(f(\xi) - f(\mathsf{E}(\xi|\mathscr{F}_0))|\Omega_n)} \leqq$$

$$\leqq \sqrt{\frac{8}{a}} \sqrt{\sum_n' n\mathsf{P}(\Omega_n)} \sqrt{\sum_n' \mathsf{P}(\Omega_n) \mathsf{E}(f(\xi) - f(\mathsf{E}(\xi|\mathscr{F}_0))|\Omega_n)} \leqq \tag{3.18}$$

$$\leqq \sqrt{\frac{8}{a}} \sqrt{1 + \frac{2a}{b} \mathsf{E}(\mathsf{E}(|\xi||\mathscr{F}_0))} \sqrt{\mathsf{E}(f(\xi) - f(\mathsf{E}(\xi|\mathscr{F}_0)))} =$$

$$= \sqrt{\frac{8}{a} + \frac{16}{b} \mathsf{E}|\xi|} \sqrt{\mathsf{E}(f(\xi) - f(\mathsf{E}(\xi|\mathscr{F}_0)))}.$$

Thus Lemma 3. 2 is proved.

REMARK. If $f''(u) \geqq a > 0$ $(0 < u < +\infty)$, the distance of ξ and $\mathsf{E}(\xi|\mathscr{F}_0)$ can be easily estimated even in the sense of L_2-norm; in fact, in this case $\mathsf{E}(\xi - \mathsf{E}(\xi|\mathscr{F}_0))^2 \leqq \frac{2}{a} \mathsf{E}(f(\xi) - f(\mathsf{E}(\xi|\mathscr{F}_0)))$, see [2]. The condition $f''(u) \geqq a > 0$ $(0 < u < +\infty)$ is, however, very restrictive[10]; the importance of Lemma 3. 2 is that it provides an estimate of the L_1-distance of ξ and $\mathsf{E}(\xi|\mathscr{F}_0)$ under a relatively weak restriction on $f(u)$.

[10] Though it holds e. g. for $f(u) = (u-1)^2$; this means that for the χ^2-divergence (1. 11) a better estimate holds than the one given in theorem 3. 3.

The following estimate is an immediate consequence of Lemma 3. 2:

THEOREM 3. 3. *If there exist positive constants a and b such that* $f''(u) \geqq \min\left(a, \frac{b}{u}\right)$ *for all* $u>0$ *then* (f, ε)*-sufficiency of* $\mathscr{X}_0$ *with respect to the pair* (μ_1, μ_2) *implies*

$$|\mu_1 - \mu_{12}| \leqq 4 \sqrt{\frac{a+2b}{2ab}} \sqrt{\varepsilon}. \tag{3.19}$$

PROOF. We may assume $\mu_1 \ll \mu_2$, since otherwise, by the assumption on $f(u)$, $\mathscr{I}_f(\mu_1, \mu_2)$ cannot be finite.

In the case $\mu_1 \ll \mu_2$, however, Lemma 3. 2 applies with $(\Omega, \mathscr{F}, \mathsf{P}) = (X, \mathscr{X}, \mu_2)$, $\mathscr{F}_0 = \mathscr{X}_0$, $\xi = \frac{p_1(x)}{p_2(x)}$, $\mathsf{E}\xi = 1$, $\mathsf{E}(\xi|\mathscr{F}_0) = \frac{\bar{p}_1(x)}{\bar{p}_2(x)}$, $\mu_2(dx) = p_2(x)\lambda(dx)$ and (3. 16) yields (3. 19).

Of course, our results remain valid also for indirect observations of type (ii) or (iii), since these, too, can be reduced to type (i). The transliteration of theorems 3. 1, 3. 2 and 3. 3 to the case of indirect observations of type (iii), i.e. to noisy channels of observation is summarized in

THEOREM 3. 4. *If the observation channel* (X, ν, Y) *is* (f, ε)*-sufficient with respect to the pair of distributions* (μ_1, μ_2), *the following estimates hold (with the notations introduced in § 2):*

a) *If* $f(u)$ *is strictly convex on a compact set* $E \subset (0, +\infty)$ *and* $B = \left\{y: \frac{\bar{p}_1(x)}{\bar{p}_2(x)} \in E\right\}$ *then*

$$|\mu_1^* - \mu_{12}^*|(X \times B) \leqq \varphi(\varepsilon) \tag{3.20}$$

for an appropriate function φ *(depending only on f and E) such that* $\varphi(0) = \lim_{\varepsilon \to +0} \varphi(\varepsilon) = 0$.

b) *If in the* r_0*-neighbourhood of each point of a Borel set* $E \subset [0, +\infty)$ $f(u)$ *is twice differentiable and* $f''(u) \geqq a > 0$ *then for* $\varepsilon \leqq \frac{1}{2} a r_0^2$ (3. 20) *holds with* $\varphi(\varepsilon) = \sqrt{\frac{8}{a}} \sqrt{\varepsilon}$.

c) *If there exist positive constants a and b such that* $f''(u) \geqq \min\left(a, \frac{b}{u}\right)$ *for all* $u>0$, *then*

$$|\mu_1^* - \mu_{12}^*| \leqq 4 \sqrt{\frac{a+2b}{2ab}} \sqrt{\varepsilon}. \tag{3.21}$$

REMARK. Theorem 3. 4 is a sharpening of theorems 2 and 2′ of [2]. It is worth while to recall that — according to (2. 13) — on the set $X \times B$ (and in the case c) almost everywhere on $X \times Y$) the density of μ_{12}^* equals $p_{12}^* = \frac{\bar{p}_1(y)}{\bar{p}_2(y)} p_2(x)$. Moreover, in the latter case one may choose $\lambda = \mu_2$, i.e. $p_2(x) = \bar{p}_2(y) = 1$, $\bar{p}_i(y) = \frac{\bar{\mu}_1(dy)}{\bar{\mu}_2(dy)}$.

§ 4. Corollaries and Comments

In this section some corollaries of the results of § 3 are presented.

We obtain nontrivial inequalities already when choosing for $\mathcal{X}_0$ the trivial σ-algebra $\{\emptyset, X\}$. In this case $\mu_{12}=\mu_2$ and $\frac{\bar{p}_1(x)}{\bar{p}_2(x)}=1$, thus $\mathcal{I}_f(\bar{\mu}_1, \bar{\mu}_2)=1$ and one may choose $E=\{1\}$. Theorems 3. 1 and 3. 2 imply (cf. also the remark to Theorem 3. 1)

$$|\mu_1-\mu_2| \leqq \varphi(\mathcal{I}_f(\mu_1, \mu_2)-f(1)) \tag{4. 1}$$

where, if $f(u)$ is twice differentiable in the r_0-neighbourhood of $u_0=1$, and there $f''(u)\geqq a>0$, one may take $\varphi(v)=\sqrt{\frac{8}{a}}\sqrt{v}$ for $v\leqq\frac{1}{2}ar_0^2$. Of course, this inequality can be proved also directly, as it has been done in [3] (in [3], for the sake of simplicity, only the case $\mu_1 \ll \mu_2$ was treated).

In particular, for $f(u)=u\log u$ we obtain

$$|\mu_1-\mu_2| \leqq \sqrt{\frac{8}{1-r_0}}\sqrt{I(\mu_1\|\mu_2)} \quad \text{if} \quad I(\mu_1\|\mu_2) \leqq \frac{1}{2}\frac{r_0^2}{1-r_0}. \tag{4. 2}$$

An estimate of form

$$|\mu_1-\mu_2| \leqq C\sqrt{I(\mu_1\|\mu_2)} \tag{4. 3}$$

was obtained first by PINSKER [8]; as we have shown, such estimates hold for a wide class of f-divergences and can be proved without any reference to the concrete form of the function $f(u)$. Now we show, that for the case of I-divergences the best constant is $C=\sqrt{2}$ (this result has been announced without proof in [3] [11]).

THEOREM 4. 1. *We have for any two probability distributions μ_1 and μ_2 on $(X, \mathcal{X})$*

$$|\mu_1-\mu_2| \leqq \sqrt{2I(\mu_1\|\mu_2)}. \tag{4. 4}$$

PROOF. The choice $A=\{x: p_1(x)\leqq p_2(x)\}$, $\mathcal{X}_0=\{\emptyset, A, X-A, X\}$ in which case $|\bar{\mu}_1-\bar{\mu}_2|=|\mu_1-\mu_2|=2|\mu_1(A)-\mu_2(A)|$ and, according to (3. 1),

$$I(\mu_1\|\mu_2) \geqq I(\bar{\mu}_1\|\bar{\mu}_2) = \mu_1(A)\log\frac{\mu_1(A)}{\mu_2(A)}+\mu_1(X-A)\log\frac{\mu_1(X-A)}{\mu_2(X-A)},$$

reduces the problem of finding the smallest C satisfying (4. 3) to that of finding the minimal C such that for all $0\leqq u\leqq v\leqq 1$

$$2|u-v| \leqq C\sqrt{u\log\frac{u}{v}+(1-u)\log\frac{1-u}{1-v}}. \tag{4. 5}$$

We set for $0<u\leqq v<1$

$$\psi_C(u, v) = u\log\frac{u}{v}+(1-u)\log\frac{1-u}{1-v}-\frac{4}{C^2}(u-v)^2. \tag{4. 6}$$

[11] Inequality (4.4) appears, even in a somewhat sharper form, also in S. KULLBACK: A lower bound for discrimination information in terms of variation, IEEE Transactions on Information Theory **13** (1967) 126—127. (Added in proof.)

Then

$$\frac{\partial \psi_C}{\partial v} = (u-v)\left(\frac{1}{-v(1-v)}+\frac{8}{C^2}\right), \tag{4.7}$$

thus $C \geqq \sqrt{2}$ implies $\frac{\partial \psi_C}{\partial v} \geqq 0$ for $v > u$, hence

$$\psi_C(u, v) \geqq \psi_C(u, u) = 0 \quad \text{for} \quad 0 < u \leqq v < 1,$$

and, on the other hand, $C < \sqrt{2}$ implies $\frac{\partial \psi_C}{\partial v} < 0$ and hence $\psi_C\left(\frac{1}{2}, v\right) < \psi_C\left(\frac{1}{2}, \frac{1}{2}\right) = 0$ for $u = \frac{1}{2} \leqq v \leqq \frac{1}{2} + \varepsilon$ (with $\varepsilon > 0$ small enough).

As by (4. 6), $\psi_C(u, v) \geqq 0$ is equivalent to (4. 5), Theorem 4. 1 is proved. Let us remark that we have also proved the impossibility of improving the constant $C_{\min} = \sqrt{2}$ by restriction to small values of $I(\mu_1 \| \mu_2)$.

The best constant can be determined, at least principially, in the same way as also for arbitrary $f(u)$ for which an estimate of type $|\mu_1 - \mu_2| \leqq C\sqrt{\mathscr{I}_f(\mu_1, \mu_2) - f(1)}$ holds. Unfortunately, in the general case the extremum problem cannot be solved explicitely.

Theorem 4. 1 gives rise to the following corollary:

THEOREM 4. 2. *For indirect observations of type* (i) *or*[12] (ii)

$$|\mu_1 - \mu_{12}| \leqq \sqrt{2(I(\mu_1 \| \mu_2) - I(\bar{\mu}_1 \| \bar{\mu}_2))} \tag{4.8}$$

and for indirect observations of type (iii)

$$|\mu_1^* - \mu_{12}^*| \leqq \sqrt{2(I(\mu_1 \| \mu_2) - I(\bar{\mu}_1 \| \bar{\mu}_2))}\,. \tag{4.9}$$

PROOF. (4. 8) is an immediate consequence of (4. 4) and (3. 3); since indirect observations of type (iii) can be reduced to those of type (i), (4. 9) holds by the same argument.

REMARK. A weaker estimate of this type (based on PINSKER's result, referred to above) has been obtained by PÉREZ [13]. Our sharpening consists of obtaining the best constant $C_{\min} = \sqrt{2}$. Let us emphasise, however, that while in [13] the specia properties of the function $f(u) = u \log u$ (as e.g. relation (3. 3)) were essentially used, here we used them only to determine the best constant. The estimate itself (with a worse constant) follows already from the convexity of $f(u)$ alone (under a rather weak condition on $f''(u)$), as it has been shown in Theorems 3. 3 and 3. 4.

We have seen that for strictly convex $f(u)$ the equality

$$\mathscr{I}_f(\mu_1, \mu_2) = \mathscr{I}_f(\bar{\mu}_1, \bar{\mu}_2) < +\infty \tag{4.10}$$

is equivalent to the sufficiency of the indirect observation with respect to the pair (μ_1, μ_2). Utilising Lemma 2. 6, we hence obtain the corollary that for an observation

[12] For indirect observations of type (ii) μ_{12} is defined by (2. 5) with $\mathscr{X}_0 = T^{-1}\mathscr{Y}$.

channel (X, v, Y) the relation (4. 10) implies the existence of a lossless code (i.e. a code with zero probability of error) for the channel of length equal to the power of a set $\left\{u: u = \frac{p_1(x)}{p_2(x)} \text{ for some } x \in X - X_0\right\}$ where $\mu_1(X_0) = \mu_2(X_0) = 1$. In particular, if there exists no lossless code for the channel (X, v, Y) of length greater than one, the equality (4. 10) implies $\mu_1 = \mu_2$. Such channels are e.g. those where the ratios $\frac{v(B|x_1)}{v(B|x_2)}$ are uniformly bounded, i.e.

$$(4.11) \qquad 0 < C_1 \leqq \frac{v(B|x_1)}{v(B|x_2)} \leqq C_2 = \frac{1}{C_1} < +\infty \qquad (B \in \mathscr{Y},\ x_1, x_2 \in X).$$

For channels satisfying (4. 11) we also have $C_1 \leqq \frac{\bar{p}_1(y)}{\bar{p}_2(y)} \leqq C_2$ and therefore, by Theorem 3. 4, (f, ε)-sufficiency implies $|\mu_1^* - \mu_{12}^*| \leqq \sqrt{\frac{8}{a}}\sqrt{\varepsilon}$ if $f''(u) \geqq a > 0$ in the interval $(C_1 - r_0,\ C_2 + r_0)$ and $\varepsilon \leqq \frac{1}{2} a r_0^2$.

Finally I should like to point on the applications of the above results when adopting the Bayesian approach (cf. [13], [14]).

Let $\{\mu_\theta\}_{\theta \in \Theta}$ be an arbitrary family of distributions on $(X, \mathscr{X})$ and assume that the parameter θ has some a priori distribution which is a probability measure π on a σ-algebra $\mathscr{T}$ of subsets of Θ, such that $\mu_\theta(A)$ is $\mathscr{T}$-measurable for all fixed $A \in \mathscr{X}$. Then (Θ, μ, X) can be interpreted as an observation channel for θ (cf. definition 2. 1) and one can speak of the joint distribution π^* on $(\Theta \times X, \mathscr{T} \times \mathscr{X})$ defined by

$$(4.12) \qquad \pi^*(C \times A) = \int_C \mu_\theta(A)\pi(d\theta) \qquad (C \in \mathscr{T},\ A \in \mathscr{X})$$

and its marginal distribution on $(X, \mathscr{X})$

$$(4.13) \qquad \bar{\pi}(A) = \int_\Theta \mu_\theta(A)\pi(d\theta) \qquad (A \in \mathscr{X}).$$

Now if (X, v, Y) is an observation channel for x, we can consider the composite channel $(\Theta, \bar{\mu}, Y)$ having the transition probability function

$$(4.14) \qquad \bar{\mu}_\theta(B) = \int v(B|x)\mu_\theta(dx) \qquad (B \in \mathscr{Y}),$$

the corresponding joint distribution $\bar{\pi}^*$ on $(\Theta \times Y, \mathscr{T} \times \mathscr{Y})$

$$(4.15) \qquad \bar{\pi}^*(C \times B) = \int_C \bar{\mu}_\theta(B)\pi(d\theta) = \int_C \int_X v(B|x)\mu_\theta(dx)\pi(d\theta) \qquad (B \in \mathscr{Y},\ C \in \mathscr{T})$$

and the marginal distribution $\bar{\bar{\pi}}$ on $(Y, \mathscr{Y})$

$$(4.16) \qquad \bar{\bar{\pi}}(B) = \int_\Theta \bar{\mu}_\theta(B)\pi(d\theta) = \int_X v(B|x)\bar{\pi}(dx) \quad (B \in \mathscr{Y}).$$

Of course, all the distributions π, $\bar{\pi}$, $\bar{\bar{\pi}}$, π^*, $\bar{\pi}^*$ are marginals of the joint distribution π^{**} on $(\Theta \times X \times Y, \mathcal{T} \times \mathcal{X} \times \mathcal{Y})$ defined by

$$\pi^{**}(C \times A \times B) = \int_C \int_A \nu(B|x)\mu_\theta(dx)\pi(d\theta) \qquad (A \in \mathcal{X},\ B \in \mathcal{Y},\ C \in \mathcal{T}). \tag{4.17}$$

Since $\mathcal{I}_f(\pi^*, \pi \times \bar{\pi})$ is a measure of dependence of x and θ (see Section 1), it can be considered also as a measure of how informative is the observation of x with respect to θ. From (3. 1) follows, applied to the observation channel $(\Theta \times X, \hat{\nu}, \Theta \times Y)$ where $\hat{\nu}$ stands for

$$\hat{\nu}(D|\theta, x) = \nu(D_\theta|x) \quad (D \in \mathcal{T} \times \mathcal{X};\ D_\theta = \{x: (\theta, x) \in D\}) \tag{4.18}$$

that

$$\mathcal{I}_f(\bar{\pi}^*, \pi \times \bar{\bar{\pi}}) \leqq \mathcal{I}_f(\pi^*, \pi \times \bar{\pi}). \tag{4.19}$$

The difference $\mathcal{I}_f(\pi^*, \pi \times \bar{\pi}) - \mathcal{I}_f(\bar{\pi}^*, \pi \times \bar{\bar{\pi}})$ is a measure of the "loss of information" by the indirect observation (X, ν, Y).

If the "loss of information" $\mathcal{I}_f(\pi^*, \pi \times \bar{\pi}) - \mathcal{I}_f(\bar{\pi}^*, \pi \times \bar{\bar{\pi}})$ is not greater than ε, one may call the channel (X, ν, Y) (f, ε)-sufficient with respect to the family $\{\mu_\theta\}_{\theta \in \Theta}$ with the a priori distribution π. Observe that this (f, ε)-sufficiency is not directly related to definition 3. 1, not even in the case $\Theta = \{1, 2\}$; in terms of definition 3. 1 it would be more adequate — but less intuitive — to speak of (f, ε)-sufficiency of the channel $(\Theta \times X, \hat{\nu}, \Theta \times Y)$ as defined by (4. 18) with respect to the pair $(\pi^*, \pi \times \bar{\pi})$.

If the channel (X, ν, Y) is $(f, 0)$-sufficient with respect to $\{\mu_\theta\}_{\theta \in \Theta}$ in the sense described above, i.e. if in (4. 19) the equality holds and $\mathcal{I}_f(\pi^*, \pi \times \bar{\pi}) < +\infty$, we may assert (by the result of [2], quoted and proved in connection with Theorem 3. 1) that the channel defined by (4. 18) is sufficient with respect to the pair $(\pi^*, \pi \times \bar{\pi})$, provided that $f(u)$ is strictly convex. Hence, it is easy to deduce, utilising Lemma 2. 6, that the channel (X, ν, Y) must be sufficient with respect to the family $\{\mu_\theta\}_{\theta \in \Theta}$, if we neglect a subset of Θ of π measure zero.

Furthermore, applying Theorem 3. 4 to the channel defined by (4. 18) — with π^*, $\pi \times \bar{\pi}$, π^{**} and $\pi \times \pi^*$ playing the role of μ_1, μ_2, μ_1^* and μ_2^* respectively — we obtain (cf. also the remark to Theorem 3. 4)

THEOREM 4. 3. *In the case* $f''(u) \geqq \min\left(a, \frac{b}{u}\right)$ *(f, ε)-sufficiency of (X, ν, Y) with respect to $\{\mu_\theta\}_{\theta \in \Theta}$ (with a priori distribution π) implies*

$$\int |(p(x, \theta) - \bar{p}(y, \theta)|(\pi \times \bar{\pi}^*)(d\theta, dx, dy) \leqq 4\sqrt{\frac{a+2b}{2ab}}\sqrt{\varepsilon}\,, \tag{4.20}$$

where $p(x, \theta)$ and $\bar{p}(y, \theta)$ stand for the densities $\dfrac{\pi^*(d\theta, dx)}{(\pi \times \bar{\pi})(d\theta, dx)}$ *and* $\dfrac{\bar{\pi}^*(d\theta, dy)}{(\pi \times \bar{\bar{\pi}})(d\theta, dy)}$ *respectively.*

In the case $f(u) = u \log u$ the constant $4\sqrt{\dfrac{a+2b}{2ab}}$ can be replaced by the smaller constant $\sqrt{2}$, utilising Theorem 4. 2. In the latter case in the term "loss of information" the parentheses can be deleted, as then one really has to do with a difference of information quantities, namely the difference of the information in x and in y with respect to θ (cf. (1. 15)).

REFERENCES

[1] CSISZÁR I.: *Eloszlások eltérésének információ-típusú mértékszámai* (Information-type measures of difference of distributions, in Hungarian) Thesis submitted to the Scientific Qualifying Committee of the Hungarian Academy of Sciences in January, 1966.

[2] CSISZÁR, I.: Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **8** (1963) 85—108.

[3] CSISZÁR, I.: A note on Jensen's inequality, *Studia Sci. Math. Hungar.* **1** (1966) 227—230.

[4] KULLBACK, S. and LEIBLER, R.: On information and sufficiency, *Ann. Math. Statist.* **22** (1951) 79—86.

[5] KULLBACK, S.: *Information theory and statistics*, Wiley, New York, 1959.

[6] PÉREZ, A.: Notions généralisées d'incertitude, d'entropie et d'information du point de vue de la théorie de martingales, *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Prague* (1957), 183—208.

[7] PÉREZ, A.: Sur la théorie de l'information dans le cas d'un alphabet abstrait, *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes,* Prague (1957), 209—243.

[8] ПИНСКЕР, М. С.: *Информация и информационная устойчивость случайных величин и процессов,* Проблемы передачи информации (Вып. 7) Изд. Акад. Наук СССР, Москва, 1960.

[9] RÉNYI, A.: On measures of entropy and information. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability,* I., Berkeley (1960) 547—561.

[10] RÉNYI, A.: On measures of dependence, *Acta Math. Acad. Sci. Hungar.* **10** (1959) 441—451.

[11] BLACKWELL, D.: Equivalent comparisons of experiments, *Ann. Math. Statist.* **24** (1953) 265—272.

[12] BARNDORFF—NIELSEN, O.: Subfields and loss of information, *Z. Wahrscheinlichkeitstheorie Verw. Gebiete.* **2** (1964) 369—379.

[13] PÉREZ, A.: Information, ε-sufficiency and data reduction problems, *Kybernetika (Prague)* **1** (1965) 297—322.

[14] RÉNYI, A.: On the amount of information concerning an unknown parameter in a sequence of observations, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **9** (1964) 617—624.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received November 2, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 319—327.

ON FINITE AND INFINITE SEQUENCES OF EXCHANGEABLE EVENTS

by

D. G. KENDALL

Summary. Finite and infinite sequences of exchangeable events are reviewed from a unified point of view, and versions of a Poisson limit theorem are proved for both the finite and the infinite case. The paper concludes with two open problems.

1. On a probability space $(\Omega, \mathscr{A}, \mathrm{pr})$ the members of a (perhaps terminating) sequence $A_1, A_2, \ldots$ of elements of $\mathscr{A}$ are called *exchangeable events*[1] when the probabilities

$$\alpha_k \equiv \mathrm{pr}\,(A_{r_1} \cap A_{r_2} \cap \ldots \cap A_{r_k}) \tag{1}$$

(for $r_1, r_2, \ldots, r_k$ all different) depend on k alone, for all relevant values of the r's. (It is useful to set $\alpha_0 = 1$.) Finite and infinite sequences of exchangeable events were introduced by DE FINETTI in his celebrated memoir [1]. For the infinite situation he proved a strong law of large numbers; if $N(n)$ is the number of the events $A_1, A_2, \ldots, A_n$ which happen, then

$$\xi = \lim_{n\to\infty} \frac{N(n)}{n} \tag{2}$$

exists almost surely, and if π (a probability measure on $I = [0, 1]$) is the distribution of ξ, then he further showed that

$$\alpha_k = \int_I x^k \pi(dx) \qquad (k = 0, 1, 2, \ldots). \tag{3}$$

Conversely, any set of α's expressible in the form (3) can obviously be associated with an infinite sequence of exchangeable events defined on a π-mixture of direct-product spaces.

DE FINETTI also obtained the appropriate analogue to (3) for finite systems of exchangeable events. Far-reaching extensions of (3) were later found by HEWITT and SAVAGE [2]. It is natural now to see in (3) a typical example of CHOQUET's theorem [4]. But CHOQUET's work was not published until 1956, a year later than the appearance of [2]; HEWITT and SAVAGE, however, made essential use of extreme-point methods in their work, and especially of the KREIN—MIL'MAN theorem.

Recently RÉNYI and RÉVÉSZ [6] have proved a deeper result underlying (3): there exists a random variable λ such that

$$\mathrm{pr}\,(A_{r_1} \cap A_{r_2} \cap \ldots \cap A_{r_k} | \lambda) = \lambda^k \text{ a. s.} \tag{4}$$

[1] We prefer this to the older expression "equivalent events" because of the ambiguity of the latter in some contexts.

Their conditioning random variable λ is defined by them up to sets of measure zero as the Radon—Nikodym derivative associated with a certain stable sequence. We shall here give a simple martingale proof of the RÉNYI—RÉVÉSZ theorem, and then follow up some further topics suggested by it.

2. The following proof of the strong law of large numbers for an infinite sequence of exchangeable events is well known (LOÈVE [3], p. 400). Let I_r be the indicator of A_r and let $\mathscr{B}(n)$ denote the σ-algebra

$$\mathscr{B}(n) = \mathscr{A}\big(N(n),\ N(n+1),\ \ldots\big)$$

$$= \mathscr{A}\big(N(n),\ I_{n+1},\ I_{n+2},\ \ldots\big). \tag{5}$$

Then

$$\frac{N(n)}{n} = \frac{1}{n} E^{\mathscr{B}(n)}\big(N(n)\big) = \frac{1}{n} \sum_{j=1}^{n} E^{\mathscr{B}(n)}(I_j)$$

$$= E^{\mathscr{B}(n)}(I_1), \tag{6}$$

by exchangeability. But by martingale theory this last conditional expectation converges almost surely, as $n \to \infty$, to a finite random variable. If we set $\xi = \limsup N(n)/n$, then we shall have $0 \leqq \xi(\omega) \leqq 1$ for all $\omega \in \Omega$, ξ will be measurable with respect to the tail σ-algebra

$$\mathscr{C} = \bigcap_{m \geqq 1} \mathscr{A}\,(I_m,\ I_{m+1},\ \ldots), \tag{7}$$

and we shall have

$$\xi = \lim_{n \to \infty} \frac{N(n)}{n} \quad \text{a. s.} \tag{8}$$

3. Now if $n \geqq r_2 > r_1 \geqq 1$, we shall have

$$\begin{aligned} \operatorname{pr}\big(A_{r_1} \cap A_{r_2} | \mathscr{B}(n)\big) &= \operatorname{pr}\big(A_1 \cap A_n | \mathscr{B}(n)\big) \\ &= E^{\mathscr{B}(n)}\big(I_n E^{\mathscr{B}(n-1)}(I_1)\big) \\ &= E^{\mathscr{B}(n)}\left(I_n \frac{N(n-1)}{n-1}\right) \\ &= E^{\mathscr{B}(n)}\left(I_n \frac{N(n)-1}{n-1}\right) \\ &= \frac{N(n)-1}{n-1} \frac{N(n)}{n}, \end{aligned}$$

on using (6) twice and noting that $N(n-1) = N(n) - 1$ unless $I_n = 0$. An obvious inductive extension of this argument gives

$$\operatorname{pr}\big(A_{r_1} \cap A_{r_2} \cap \ldots \cap A_{r_k} | \mathscr{B}(n)\big) = \prod_{j=0}^{k-1} \frac{N(n)-j}{n-j}, \tag{9}$$

whenever $1 \leqq r_1 < r_2 < \ldots < r_k \leqq n$, the case $k = 1$ of this identity having been proved already in the form (6).

If we now let $n \to \infty$, we find that

$$\text{pr}\,(A_{r_1} \cap A_{r_2} \cap \ldots \cap A_{r_k} | \mathscr{B}) = \xi^k, \tag{10}$$

where

$$\mathscr{B} \equiv \bigcap_{n \geq 1} \mathscr{B}(n). \tag{11}$$

To relate (10) and (11) to the RÉNYI—RÉVÉSZ result (4), note that from (4) we must have

$$E^{\mathscr{A}(\lambda)}\left(\frac{N(n)}{n}\right) = \frac{1}{n}\sum_{j=1}^{n} E^{\mathscr{A}(\lambda)}(I_j)$$

$$= \text{pr}\,(A_1|\lambda) = \lambda \qquad \text{a. s.},$$

and so on letting $n \to \infty$, it follows that

$$E^{\mathscr{A}(\lambda)}(\xi) = \lambda. \tag{12}$$

Similarly

$$E^{\mathscr{A}(\lambda)}\left(\frac{N(n)}{n}\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E^{\mathscr{A}(\lambda)}(I_i I_j)$$

$$= \frac{1}{n}\,\text{pr}\,(A_1|\lambda) + \frac{n-1}{n}\,\text{pr}\,(A_1 \cap A_2|\lambda),$$

and so

$$E^{\mathscr{A}(\lambda)}(\xi^2) = \lambda^2. \tag{13}$$

From (12) and (13) we find that

$$E^{\mathscr{A}(\lambda)}((\xi - \lambda)^2) = \lambda^2 - 2\lambda^2 + \lambda^2 = 0,$$

and so

$$\xi = \lambda \quad \text{a.s.} \tag{14}$$

From our present point of view (10) is slightly preferable to (4); ξ has better measurability properties than λ and on the other hand the conditional smoothing involved in (10) is the less drastic. The DE FINETTI representation (3) of course follows immediately from (10), just as it did from (4).

4. The analogous results for a finite set of exchangeable events are easily obtained by consideration of the difference tableau

$$\begin{array}{lll}
\alpha_0 & & \\
 & \delta\alpha_0 & \\
\alpha_1 & & \delta^2\alpha_0 \ \ldots \\
 & \delta\alpha_1 & \\
\alpha_2 & & \\
\ldots & & \delta^n\alpha_0 \\
\alpha_{n-2} & & \\
 & \delta\alpha_{n-2} & \\
\alpha_{n-1} & & \delta^2\alpha_{n-2} \ \ldots \\
 & \delta\alpha_{n-1} & \\
\alpha_n & &
\end{array} \tag{15}$$

Here $\delta = 1 - E$, where $E\alpha_j = \alpha_{j+1}$. Clearly

$$\delta^r \alpha_s = \text{pr}\,(A_{m_1} \cap \ldots \cap A_{m_s} \cap A^*_{m_{s+1}} \cap \ldots \cap A^*_{m_{s+r}}) \tag{16}$$

when all the m's are different, and $A^* = \Omega \backslash A$. Thus

$$\delta^r \alpha_{n-r} \geqq 0 \qquad (r = 0, 1, \ldots, n) \tag{17}$$

and (because $\alpha_0 = 1$)

$$\sum_{r=0}^{n} \binom{n}{r} \delta^r \alpha_{n-r} = 1. \tag{18}$$

Conversely, (17) and (18) ensure that $(\alpha_0, \alpha_1, \ldots, \alpha_n)$ can be associated as at (1) with a set of n exchangeable events. We have only to take a probability space Ω of 2^n points labelled $\varepsilon_1 \varepsilon_2 \ldots \varepsilon_n$ (all ε's being 0 or 1) and let A_j be the set of points for which $\varepsilon_j = 1$. We then obtain the desired construction by giving $\varepsilon_1 \varepsilon_2 \ldots \varepsilon_n$ a mass $\delta^r \alpha_{n-r}$, where r is the number of ε's which are equal to zero.

From the usual formulae of the calculus of differences,

$$\alpha_m = \sum_{s=0}^{n-m} \binom{n-m}{s} \delta^s \alpha_{n-s};$$

let us write

$$\omega_j \equiv \binom{n}{j} \delta^{n-j} \alpha_j \qquad (j = 0, 1, \ldots, n), \tag{19}$$

so that ω is a probability distribution in virtue of (17) and (18). Then

$$\alpha_m = \sum_{s=m}^{n} \frac{s(s-1)\ldots(s-m+1)}{n(n-1)\ldots(n-m+1)} \omega_s \qquad (m = 0, 1, \ldots, n), \tag{20}$$

and this is the correct analogue of (3) in the finite case. A formula equivalent to (20) can be obtained from (1) of [1] on on setting $k = m$.

Now if exactly N of the n events occur, we shall have

$$\text{pr}\,(N = j) = \binom{n}{j} \delta^{n-j} \alpha_j = \omega_j, \tag{21}$$

and thus

$$\text{pr}\,(A_{r_1} \cap A_{r_2} \cap \ldots \cap A_{r_k} | N) = \binom{n-k}{N-k} \frac{\delta^{n-N} \alpha_N}{\omega_N} = \binom{n-k}{N-k} \Big/ \binom{n}{N},$$

so that

$$\text{pr}\,(A_{r_1} \cap A_{r_2} \cap \ldots \cap A_{r_k} | N) = \frac{N(N-1)\ldots(N-k+1)}{n(-1)\ldots(n-k+1)}; \tag{22}$$

this formula (22) corresponds to the RÉNYI—RÉVÉSZ theorem (4) in the finite situation. We may say that every system of n exchangeable events is equivalent to a random sampling scheme "without replacement", the number of items N in the sampling having an arbitrary distribution ω.

5. RÉNYI [5] (cf. also [2]) has remarked that for an infinite sequence of exchangeable events the simple condition

$$\alpha_2 = \alpha_1^2 \tag{23}$$

is necessary and sufficient for independence, because it implies that ξ (or λ) is a.s. constant and so that $\alpha_k = \alpha_1^k$ for all k. In general $\alpha_2 \geqq \alpha_1^2$ for an infinite sequence.

But when we are concerned with a finite set of equivalent events this result ceases to apply. For example, let Ω consist of $(m-1)^2$ equiprobable points arranged as an $m \times (m-2)$ array plus one "extra" point, and let A_j consist of the jth row of the array plus the extra point, for $j = 1, 2, \ldots, m$. These events are obviously exchangeable, and

$$\alpha_1 = \frac{1}{m-1}, \quad \alpha_2 = \frac{1}{(m-1)^2}, \quad \alpha_j = \frac{1}{(m-1)^2} \qquad (j \geqq 3).$$

Here $\alpha_2 = \alpha_1^2$, but the events are not independent for $m \geqq 3$ because $\alpha_3 \neq \alpha_1^3$.

There does exist a correct parallel to RÉNYI' sresult, however. The SCHWARZ inequality applied to (20) shows that

$$n(\alpha_1^2 - \alpha_2) \leqq \alpha_1 - \alpha_2,$$

with equality if and only if $\omega_s = 1$ for some $s = 0, 1, \ldots, n$. Thus

$$n(\alpha_1^2 - \alpha_2) = \alpha_1 - \alpha_2 \tag{24}$$

is the necessary and sufficient condition for the n equivalent events to be those associated with a random sampling scheme "without replacement" where the sample size is fixed (and is equal to $n\alpha_1$). In all cases $\alpha_1 \geqq \alpha_2$, but $\alpha_1^2 - \alpha_2$ can have either sign or may vanish. If $\alpha_1^2 > \alpha_2$ then we cannot associate $(\alpha_0, \alpha_1, \alpha_2)$ with a subset of a system containing more than

$$\left[\frac{\alpha_1 - \alpha_2}{\alpha_1^2 - \alpha_2}\right]$$

exchangeable events.

DE FINETTI in [1] formulated a conjecture which it is convenient to settle here. Let $\mathscr{J}_n$ denote the class of sequences $(\alpha_0, \alpha_1, \ldots)$ such that the n-ad $(\alpha_0, \alpha_1, \ldots, \alpha_n)$ can be associated with a set of n exchangeable events, and let $\mathscr{J}_\infty$ be defined similarly. Then it is clear that

$$\mathscr{J}_\infty \subseteq \mathscr{J}_{n+1} \subseteq \mathscr{J}_n$$

for all n and de Finetti conjectured that

$$\mathscr{J}_\infty = \bigcap_{n \geqq 1} \mathscr{J}_n. \tag{25}$$

We shall now prove that this conjecture is correct.

Suppose that $(\alpha_0, \alpha_1, \ldots) \in \mathscr{J}_n$ for all n. Then

$$\alpha_j = \int_I f_{j,n}(x)\, \pi_n(dx) \qquad (0 \leqq j \leqq n < \infty),$$

where π_n is an atomic probability measure defined by

$$\pi_n\left(x = \frac{s}{n}\right) = \omega_s^{(n)} \qquad (0 \leqq s \leqq n),$$

and where $f_{0,n}(x) \equiv 1$, and

$$f_{j,n}(x) \equiv \prod_{h=0}^{j-1} \left(\frac{nx-h}{n-h}\right)^+ \qquad (j = 1, 2, \ldots, n).$$

Now for $0 \leqq x \leqq 1$,

$$0 \leqq x^j - f_{j,n}(x) \leqq \sum_{h=0}^{j-1} \frac{h}{n} = \frac{j(j-1)}{2n},$$

and so

$$\left| \alpha_j - \int_I x^j \pi_n(dx) \right| \leqq \frac{j(j-1)}{2n}. \tag{26}$$

If we now choose a subsequence of $n = 1, 2, \ldots$ for which the corresponding subsequence of the sequence $(\pi_1, \pi_2, \pi_3, \ldots)$ of probability measures on I converges completely to a probability measure π on I, then from (26) we obtain

$$\alpha_j = \int_I x^j \pi(dx) \qquad (j = 0, 1, \ldots,),$$

and so (in virtue of the remark following (3)) $(\alpha_0, \alpha_1, \ldots) \in \mathscr{J}_\infty$, as required.

6. The following Poisson limit theorem appears to be new. Professor RÉNYI tells me that a related but different result has been found independently by A. BENCZUR. BENCZUR postulated the existence of *all* the limits $\lim_{\nu \to \infty} \nu^j \alpha_j^{(\nu)} = \mu_j$ $(j = 1, 2, \ldots)$, and obtained a mixture of Poissons as the resulting limit-law. Viewed against the background provided by BENCZUR's theorem, the essential content of our Theorem I is not so much that the limit-law will be Poisson when $\mu_2 = \mu_1^2$, as that in that case the existence of the limits μ_j for $j > 2$ need not be assumed.

THEOREM I. *Let* $(\Omega_\nu, \mathscr{A}_\nu, \mathrm{pr}_\nu)$ *(for each* $\nu = 1, 2, \ldots$*) be a probability space carrying an infinite sequence* $(A_r^{(\nu)})$ *of exchangeable events with de Finetti constants* $(\alpha_j^{(\nu)})$*, and let* X_ν *be the number of the events*

$$A_1^{(\nu)}, A_2^{(\nu)}, \ldots, A_\nu^{(\nu)}$$

which occur. Suppose that

$$\nu \alpha_1^{(\nu)} \to \mu \quad \text{and} \quad \nu^2 \alpha_2^{(\nu)} \to \mu^2, \tag{27}$$

as $\nu \to \infty$*; then*

$$\lim_{\nu \to \infty} \mathrm{pr}_\nu(X_\nu = s) = e^{-\mu} \frac{\mu^s}{s!} \qquad (s = 0, 1, \ldots).$$

PROOF. From (21) and (3) we see that

$$\mathrm{pr}_\nu(X_\nu = s) = \binom{\nu}{s} \int_I (1-x)^{\nu-s} x^s \pi_\nu(dx),$$

where π_ν is a probability measure on I, and so

$$E_\nu(z^{X_\nu}) = \int_{[0,\nu]} \left\{1 - \frac{(1-z)y}{\nu}\right\}^\nu \Pi_\nu(dy) \tag{28}$$

for $|z| \leqq 1$, where $y = vx$ and Π_v is the measure on the y-axis corresponding to π_v on the x-axis. Now ([8], p. 242) we know that

$$0 \leqq e^{-u} - \left(1 - \frac{u}{v}\right)^v \leqq \frac{u^2}{v} e^{-u} \leqq \frac{4e^{-2}}{v}$$

if $0 \leqq u \leqq v$, and so if we replace the integrand in (28) by

$$\exp\{-(1-z)y\},$$

then the resulting error in the integral will be at most $(4e^{-2})/v$, provided that z is real and that $0 \leqq z \leqq 1$.

Next we observe that if y has the distribution Π_v then

$$E_v y = v\alpha_1^{(v)} \quad \text{and} \quad \mathrm{var}_v(y) = v^2\{\alpha_2^{(v)} - \alpha_1^{(v)2}\},$$

and so from (27), using a Čebyševian argument, we find that for $0 \leqq z \leqq 1$ we must have

$$\lim_{v \to \infty} \int_{[0, v]} e^{-(1-z)y} \Pi_v(dy) = e^{-(1-z)\mu}.$$

But now it follows that $E_v(z^{X_v})$ must have this same limit as $v \to \infty$, if $0 \leqq z \leqq 1$, and the continuity theorem for probability-generating functions then gives the desired result.

It is worth noticing that (27), though sufficient, is not a necessary condition. To see this, let π_v assign masses $1 - \frac{1}{v}$ to $x = \frac{\mu}{v}$ and $\frac{1}{v}$ to $x = 1$. Then

$$\alpha_1^{(v)} = \frac{\mu}{v} - \frac{\mu}{v^2} + \frac{1}{v}$$

and

$$\alpha_2^{(v)} = \frac{\mu^2}{v^2} - \frac{\mu^2}{v^3} + \frac{1}{v},$$

so that

$$v\alpha_1^{(v)} \to \mu + 1 \text{ and } v^2\alpha_2^{(v)} \to \infty,$$

and (27) does not hold; nevertheless

$$\mathrm{pr}_v(X_v = s) = \binom{v}{s}\left(1 - \frac{\mu}{v}\right)^{v-s}\left(\frac{\mu}{v}\right)^s\left(1 - \frac{1}{v}\right) + \binom{v}{s}\frac{1}{v}\delta_{sv} \to e^{-\mu}\frac{\mu^s}{s!}.$$

In connexion with a problem in epidemic theory[2] the need has arisen for an analogue to Theorem I which will be valid when on $(\Omega_v, \mathscr{A}_v, \mathrm{pr}_v)$ we have v exchangeable events *and perhaps no more*. In these circumstance the condition (27) is *too weak*; for example, let $(\Omega_v, \mathscr{A}_v, \mathrm{pr}_v)$ and $A_1^{(v)}, A_2^{(v)}, \ldots, A_v^{(v)}$ be constructed as in the second

[2] See a forthcoming paper by Miss V. R. Cane (*J. Roy. Statist. Soc.* B, **28** (1966) 487—490).

paragraph of § 5, with $m = v$. Then we shall have

$$v\alpha_1^{(v)} = \frac{v}{v-1} \to 1$$

and

$$v^2\alpha_2^{(v)} = \left(\frac{v}{v-1}\right)^2 \to 1,$$

so that (27) is satisfied with $\mu = 1$, and yet

$$\mathrm{pr}_v(X_v = 1) = 1 - \frac{1}{(v-1)^2} \to 1,$$

and the limit-law is *not Poissonian*. A correct analogue to Theorem I in this situation is

THEOREM II. *Let* $(\Omega_v, \mathscr{A}_v, \mathrm{pr}_v)$ *for each* $v = 1, 2, \ldots$ *carry* v *exchangeable events* $A_1^{(v)}, A_2^{(v)}, \ldots, A_v^{(v)}$ *with de Finetti constants* $\alpha_0^{(v)}, \alpha_1^{(v)}, \ldots, \alpha_v^{(v)}$, *and let* X_v *be the number of these events which happen. Suppose that, for each* $j \geqq 1$,

$$v^j\alpha_j^{(v)} \to \mu^j \quad \text{as} \quad v \to \infty \qquad (0 \leqq j \leqq v < \infty). \tag{29}$$

Then

$$\lim_{v\to\infty} \mathrm{pr}_v(X_v = s) = e^{-\mu}\frac{\mu^s}{s!} \qquad (s = 0, 1, \ldots).$$

PROOF. From (20) we know that

$$v(v-1)\ldots(v-j+1)\alpha_j^{(v)} = \sum_{s=0}^{v} s(s-1)\ldots(s-j+1)\omega_s^{(v)},$$

and in view of (29) the left-hand side of this identity converges (when $v \to \infty$) to the limit

$$\mu^j = \sum_{s=0}^{\infty} s(s-1)\ldots(s-j+1)e^{-\mu}\frac{\mu^s}{s!}.$$

Now the jth moment of a distribution is a linear combination of the factorial moments of orders not exceeding j, and so it follows that

$$\sum_{s=0}^{v} s^j\omega_s^{(v)} \to \sum_{s=0}^{\infty} s^j e^{-\mu}\frac{\mu^s}{s!}$$

for each $j \geqq 0$. The conclusion of Theorem II now follows from the moments convergence theorem ([3], p. 185) and the fact that the Poisson distribution is uniquely determined by its moments.

Theorem II is a straightforward generalisation of a result of VON MISES concerning the limiting values of certain occupancy probabilities [7].

7. Theorems I and II, and the counter-example used to show that there is no naive generalisation of Theorem I, together suggest that attention might now profitably be directed to the following *open problems*.

For each $\nu = 1, 2, \ldots$ let $(\Omega_\nu, \mathscr{A}_\nu, \mathrm{pr}_\nu)$ be a probability space carrying a finite number M_ν of exchangeable events $\{A_r^{(\nu)} : r = 1, 2, \ldots, M_\nu\}$, where $M_\nu \geqq \nu$ and $M_\nu \uparrow \infty$ as $\nu \to \infty$. Let X_ν be the number of the first ν events $A_1^{(\nu)}, A_2^{(\nu)}, \ldots, A_\nu^{(\nu)}$ which happen. What conditions on the rate-of-growth of the sequence $M_1, M_2, \ldots$ make it possible to assert the desired conclusion,

$$\lim_{\nu \to \infty} \mathrm{pr}_\nu(X_\nu = s) = e^{-\mu} \frac{\mu^s}{s!} \qquad (s = 0, 1, \ldots), \tag{30}$$

when (29) of Theorem II is replaced by

$$\nu^j \alpha_j^{(\nu)} \to \mu^j \qquad (\nu \to \infty)$$

for $j = 1, 2, \ldots, J$? (The case $J = 2$ is of especial interest.)

On can also ask, when $M_\nu = \nu$ and $J = 2$, whether (30) follows if $\nu\alpha_1^{(\nu)}$ and $\nu^2\alpha_2^{(\nu)}$ are required to converge to their limits μ and μ^2 sufficiently rapidly, as $\nu \to \infty$.

REFERENCES

[1] De Finetti, B.: Funzione caratteristica di un fenomeno aleatorio, *Atti Accad. Naz. Lincei. Rend. Cl. Sci. Fis. Mat. Nat.* (6) **4** (1930) 86—133.

[2] Hewitt, E. and Savage, L. J.: Symmetric measures on cartesian products, *Trans. Amer. Math. Soc.* **80** (1955) 470—501.

[3] Loève, M.: *Probability Theory,* (3rh ed.) Princeton, 1963.

[4] Phelps, R. R.: *Lectures on Choquet's Theorem,* Princeton, 1966.

[5] Rényi, A.: On stable sequence of events, *Sankhya Ser. A.* **25** (1963) 293—302.

[6] Rényi, A. and Révész, P.: A study of sequences of equivalent events as special stable sequences, *Publ. Math. Debrecen.* **10** (1963) 319—325.

[7] Von Mises, R.: Über Aufteilungs- und Besetzungswahrscheinlichkeiten, *Rev. Fac. Sci. Univ. d'Istanbul* **4** (1939) 145—163.

[8] Whittaker, E. T. and Watson, G. N.: *Modern Analysis,* (4th ed.) Cambridge, 1927.

STATISTICAL LABORATORY, UNIVERSITY OF CAMBRIDGE

and

DEPARTMENT OF STATISTICS, JOHN HOPKINS UNIVERSITY

(Received November 9, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 329—339.

ON TOPOLOGICAL PROPERTIES OF f-DIVERGENCES

by

I. CSISZÁR

The concept of f-divergence of probability distributions has been introduced in [1] as a generalisation of KULLBACK's I-divergence („information for discrimination", [2]) and of RÉNYI's I-divergence („information gain") of order α [3]. The interpretation of the f-divergence as a measure of how different two probability distributions are, suggests the question, what kind of a topological structure it gives rise to in the space of probability distributions. This problem, for the case of the I-divergence (of any positive order $\alpha > 0$), has been considered in [4] and [5]; in this paper we extend this investigation to arbitrary f-divergences. The paper contains so far unpublished results of my thesis [6], with one sharpening (Theorem 4). Other results of the same thesis — concerning the decrease of f-divergence in indirect observations — are published in [8].

Let $f(u)$ be an arbitrary convex function defined on the real half-axis $(0, +\infty)$. The f-divergence of any two probability measures μ_1 and μ_2 on a measurable space $(X, \mathcal{X})$ is defined (cf. [1], [7], [8]) as

$$\mathcal{I}_f(\mu_1, \mu_2) = \int p_2(x) f\left(\frac{p_1(x)}{p_2(x)}\right) \lambda(dx) \tag{1}$$

where λ is an arbitrary (σ-finite) dominating measure of μ_1 and μ_2 and $p_i(x)$ stands for the density (Radon—Nikodym-derivative)

$$p_i(x) = \frac{\mu_i(dx)}{\lambda(dx)} \qquad (i = 1, 2). \tag{2}$$

In (1) we understand

$$f(0) = \lim_{u \to +0} f(u)$$

$$0 \cdot f\left(\frac{0}{0}\right) = 0 \tag{3}$$

$$0 \cdot f\left(\frac{a}{0}\right) = \lim_{\varepsilon \to +0} \varepsilon f\left(\frac{a}{\varepsilon}\right) = a \lim_{u \to +\infty} \frac{f(u)}{u} \qquad (a > 0).$$

Then[1] the integral (1) is always well-defined, its value does not depend on the choice of the dominating measure λ and we have $\mathcal{I}_f(\mu_1, \mu_2) \geqq f(1)$ with equality if and only if $\mu_1 = \mu_2$, provided that $f(u)$ is strictly convex at $u_0 = 1$.

[1] All these simple properties have been exhibited in [1]; the last one is also an immediate corollary of theorem 1 below.

The f-divergence $\mathcal{I}_f(\mu_1, \mu_2)$ or, more precisely, the difference $\mathcal{I}_f(\mu_1, \mu_2) - f(1)$ can be interpreted as a measure of how different is μ_1 from μ_2. This suggest the following

DEFINITION 1. Let $\mathcal{M}$ be a set of probability distributions (=probability measures) on a measurable space $(X, \mathcal{X})$. The f-neighbourhood of radius ε of a distribution $\mu_0 \in \mathcal{M}$ is the set of distributions

$$U_f(\mu_0, \varepsilon) = \{\mu: \mathcal{I}_f(\mu, \mu_0) - f(1) < \varepsilon, \mu \in \mathcal{M}\} \qquad (\varepsilon > 0) \tag{4}$$

REMARKS. 1. The choice

$$f(u) = u \log u = f_1(u) \quad \text{or} \quad f(u) = u^\alpha \operatorname{sgn}(\alpha - 1) = f_\alpha(u) \qquad (\alpha > 0, \alpha \neq 1)$$

leads to the I-divergence $I_\alpha(\mu_1 \| \mu_2)$ of order α:

$$I(\mu_1 \| \mu_2) = I_1(\mu_1 \| \mu_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} \lambda(dx) = \mathcal{I}_{f_1}(\mu_1, \mu_2) \tag{5}$$

$$I_\alpha(\mu_1 \| \mu_2) = \frac{1}{\alpha - 1} \log \int p_1^\alpha(x) p_2^{1-\alpha}(x) dx = \frac{1}{\alpha - 1} \log |\mathcal{I}_{f_\alpha}(\mu_1, \mu_2)|. \tag{6}$$

Thus the "information-neighbourhoods"

$$V_\alpha(\mu_0, \varepsilon) = \{\mu: I_\alpha(\mu \| \mu_0) < \varepsilon, \mu \in \mathcal{M}\} \tag{7}$$

introduced in [4], are special f-neighbourhoods.

2. Our definition of f-neighbourhoods is by no means the only one possible. E. g. one could substitute in (4) the term

$$\mathcal{I}_f(\mu, \mu_0) \quad \text{by} \quad \mathcal{I}_f(\mu_0, \mu) \quad \text{or} \quad \mathcal{I}_f(\mu, \mu_0) + \mathcal{I}_f(\mu_0, \mu)$$

(for $f(u) = u \log u$ the latter quantity is the so called J-divergence: $J(\mu_1, \mu_2) = I(\mu_1 \| \mu_2) + I(\mu_2 \| \mu_1)$). This substitution, however, leads again to f-neighbourhoods in the sense of definition 1 itself, only the original $f(u)$ has to be replaced by another convex function $f^*(u)$ or $\tilde{f}(u)$, respectively:

$$f^*(u) = uf\left(\frac{1}{u}\right), \tag{8}$$

$$\tilde{f}(u) = f(u) + f^*(u). \tag{9}$$

In fact, we have — according to (1) — for any two distributions μ_1 and μ_2

$$\mathcal{I}_f(\mu_2, \mu_1) = \mathcal{I}_{f^*}(\mu_1, \mu_2) \tag{10}$$

$$\mathcal{I}_f(\mu_1, \mu_2) + \mathcal{I}_f(\mu_2, \mu_1) = \mathcal{I}_{\tilde{f}}(\mu_1, \mu_2). \tag{11}$$

Definition 1 makes the set of distributions $\mathcal{M}$ to a so called FRÉCHET (V)-space (see [9]). In such spaces many topological concepts can be defined (open and closed sets, convergence, etc.), though they need not be topological spaces in the usual sense. In particular, we may say, that a sequence $\{\mu_n\}$ or, more generally, a net $\{\mu_d\}$ (d ranging over some directed set D) of distributions from $\mathcal{M}$ f-converges to $\mu_0 \in \mathcal{M}$ iff for any $\varepsilon > 0$ there exists an n_0 (or $d_0 \in D$) such that $n \geqq n_0$ implies

$\mu_n \in U_f(\mu_0, \varepsilon)$ (or $d > d_0$ implies $\mu_d \in U_f(\mu_0, \varepsilon)$, respectively.) For $f(u) = f_\alpha(u)$ (cf. remark 1 to definition 1) this f-convergence reduces to the „total" information-convergence of order α, considered in [4].

It follows directly from [7] (see also [8]) that the topological structure defined by the f-neighbourhoods — with $f(u)$ strictly convex at $u_0 = 1$ — is finer than the metric topology defined by the variation distance

$$\varrho(\mu_1, \mu_2) = |\mu_1 - \mu_2| = \int |p_1(x) - p_2(x)| \lambda(dx) \tag{12}$$

(Eventually, $\varrho(\mu_1, \mu_2)$ also belongs to the class of f-divergences: $\varrho(\mu_1, \mu_2) = \mathscr{I}_{|u-1|}(\mu_1, \mu_2)$).

For the reader's convenience, I reproduce here the simple proof.

THEOREM 1. *If $f(u)$ is strictly convex at $u_0 = 1$, for small values of $\mathscr{I}_f(\mu_1, \mu_2) - f(1)$ also $\varrho(\mu_1, \mu_2)$ is small; if, in particular, $f(u)$ is twice differentiable at $u_0 = 1$ and $f''(1) > a > 0$, then $\mathscr{I}_f(\mu_1, \mu_2) - f(1) < \varepsilon$ implies in case ε small enough*

$$\varrho(\mu_1, \mu_2) < \sqrt{\frac{8\varepsilon}{a}}. \tag{13}$$

PROOF. We have by the convexity of $f(u)$

$$f(u) \geqq f(1) + b(u-1) \tag{14}$$

where $b = f'(1)$ if $f'(1)$ exists, and b equals e.g. the arithmetic mean of the left and right derivatives of $f(u)$ at $u_0 = 1$ if $f'(1)$ does not exist. If $f(u)$ is strictly convex at $u_0 = 1$, for $|u-1| \geqq \varepsilon_0 > 0$ even the stronger inequality

$$f(u) \geqq f(1) + b(u-1) + \varepsilon_1 |u-1| \tag{15}$$

holds, where

$$\varepsilon_1 = \frac{1}{\varepsilon_0} \min \{f(1+\varepsilon_0) + b\varepsilon_0 - f(1), f(1-\varepsilon_0) - b\varepsilon_0 - f(1)\} > 0. \tag{16}$$

(14) and (15) can be rewritten as

$$f(u) \geqq f(1) + b(u-1) + \varepsilon_1 |u-1| (1 - \chi_{\varepsilon_0}(u)) \qquad (u \geqq 0) \tag{17}$$

where $\chi_{\varepsilon_0}(u) = 1$ for $|u-1| < \varepsilon_0$ and 0 for $|u-1| \geqq \varepsilon_0$. If we substitute in (17) $u = \frac{p_1(x)}{p_2(x)}$ and multiply formally by $p_2(x)$, the obtained inequality remains valid even for $p_2(x) = 0$, in sense of the conventions (3). Integrating with respect to λ we get, according to (1) and (12),

$$\mathscr{I}_f(\mu_1, \mu_2) \geqq f(1) + \varepsilon_1 \varrho(\mu_1, \mu_2) - \varepsilon_1 \int\limits_{\left|\frac{p_1(x)}{p_2(x)} - 1\right| < \varepsilon_0} |p_1(x) - p_2(x)| \lambda(dx) \tag{18}$$

whence

$$\varrho(\mu_1, \mu_2) \leqq \frac{1}{\varepsilon_1} (\mathscr{I}_f(\mu_1, \mu_2) - f(1)) + \varepsilon_0. \tag{19}$$

As ε_0 can be chosen arbitrarily, (19) proves the first statement of the theorem;

(13) follows by the choice $\varepsilon_0 = \sqrt{\frac{2\varepsilon}{a}}$, using that (16) implies $\varepsilon_1 \leqq \frac{2}{a}\varepsilon_0$, if $f''(1) > a$ and ε_0 is small enough.

Though of little importance, let us remark, that the constant $\sqrt{\frac{8}{a}}$ in (13) is smaller than the corresponding constant obtained in [7]. For the case $f(u) = u \log u$, the best possible constant has been determined in [8], where the estimate

$$\varrho(\mu_1, \mu_2) \leqq \sqrt{2I(\mu_1 \| \mu_2)} \tag{20}$$

was proved.

Theorem 1 means, in particular, that the „f-convergence" of distributions necessarily implies uniform convergence, i.e. convergence in the metric $\varrho(\mu_1, \mu_2) = |\mu_1 - \mu_2|$. In the opposite direction only the following holds:

THEOREM 2. *If both* $f(0)$ *and* $f^*(0) = \lim_{u \to +\infty} \frac{f(u)}{u}$ *are finite, then*

$$\mathscr{I}_f(\mu_1, \mu_2) \leqq f(1) + C\sqrt{\varrho(\mu_1, \mu_2)}, \tag{21}$$

where $C > 0$ *is a constant, depending only on* f.

PROOF. By the conditions of the theorem, in the interval [0, 1] both $f(u)$ and $f^*(u) = uf\left(\frac{1}{u}\right)$ are bounded, thus there exist constants K_1 and K_2 such that

$$|f(u) - f(1)| < K_1, \qquad \left| uf\left(\frac{1}{u}\right) - uf(1) \right| < K_2 \qquad (0 \leqq u \leqq 1). \tag{22}$$

We set $(0 < \varepsilon < 1)$

$$\begin{aligned} E_\varepsilon^1 &= \{x : (1-\varepsilon)p_1(x) \geqq p_2(x), p_1(x) > 0\} \\ E_\varepsilon^2 &= \{x : (1-\varepsilon)p_2(x) \geqq p_1(x)\} \end{aligned} \tag{23}$$

$$E_\varepsilon = X - (E_\varepsilon^1 \cup E_\varepsilon^2) = \left\{x : 1 - \varepsilon < \frac{p_1(x)}{p_2(x)} < \frac{1}{1-\varepsilon}\right\}.$$

Then for $x \in E_\varepsilon$

$$\left| f\left(\frac{p_1(x)}{p_2(x)}\right) - f(1) \right| \leqq K_\varepsilon \cdot \varepsilon \tag{24}$$

with

$$K_\varepsilon = \max\left\{\frac{1}{\varepsilon}|f(1) - f(1-\varepsilon)|, \frac{1}{\frac{1}{1-\varepsilon} - 1}\left| f\left(\frac{1}{1-\varepsilon}\right) - f(1)\right|\right\},$$

where, by the convexity of $f(u)$, K_ε is a monotone increasing function of ε. Furthermore, by (22) and (23).

$$\begin{aligned} \int_{E_\varepsilon^1} \left| f\left(\frac{p_1(x)}{p_2(x)}\right) - f(1) \right| \frac{p_2(x)}{p_1(x)} p_1(x)\lambda(dx) &\leqq K_2 \mu_1(E_\varepsilon^1) \\ \int_{E_\varepsilon^2} \left| f\left(\frac{p_1(x)}{p_2(x)}\right) - f(1) \right| p_2(x)\lambda(dx) &\leqq K_1 \mu_2(E_\varepsilon^2) \end{aligned} \tag{25}$$

and by (23) and (12)

$$\varrho(\mu_1, \mu_2) \leqq \int_{E_\varepsilon^i} |p_1(x) - p_2(x)| \lambda(dx) \leqq \varepsilon \mu_i(E_\varepsilon^i) \qquad (i = 1, 2) \tag{26}$$

Finally, the inequalities (24), (25) and (26) give rise to

$$\begin{aligned} \mathscr{I}_f(\mu_1, \mu_2) - f(1) &\leqq \int \left| f\left(\frac{p_1(x)}{p_2(x)}\right) p_2(x) - f(1) p_2(x) \right| \lambda(dx) = \\ &= \int_{E_\varepsilon} + \int_{E_\varepsilon^1} + \int_{E_\varepsilon^2} \leqq K_\varepsilon \varepsilon + \frac{K_1 + K_2}{\varepsilon} \varrho(\mu_1, \mu_2) \end{aligned} \tag{27}$$

whence, e.g. by the choice $\varepsilon = \frac{1}{2}\sqrt{\varrho(\mu_1, \mu_2)}$ (implying $\varepsilon \leqq \frac{\sqrt{2}}{2}$ and thus $K_\varepsilon \leqq K_{\frac{\sqrt{2}}{2}}$) we obtain (21).

According to theorems 1 and 2, if $f(u)$ satisfies the conditions of theorem 2 and it is strictly convex at $u_0 = 1$, one may assert that the f-neighbourhoods generate a topology, namely the metric topology associated with $\varrho(\mu_1, \mu_2)$. For the functions $f_\varepsilon(u)$ (remark 1 to definition 1) these conditions hold for $0 < \alpha < 1$; thus for such an α, the „information-neighbourhoods" (7) generate the topology of the uniform convergence of distributions, as it has been shown already in [4].

Now we are going to show that if $f(0)$ or $f^*(0) = \lim_{u \to \infty} \frac{f(u)}{u}$ is infinite, the f-neighbourhoods (4) do not generate a topology for $\mathscr{M}$, except for very special sets of distributions $\mathscr{M}$. To get a counterexample, one may choose even $X = \{x_1, x_2, \ldots\}$ i.e. a denumerable set, $\mathscr{X}$ consisting of all subsets of X. Then any probability distribution μ on $(X, \mathscr{X})$ is uniquely determined by the sequence $P = \{p_1, p_2, \ldots\}$, $p_k = \mu(\{x_k\})$. We choose for $\mathscr{M}$ the set of all distributions on $(X, \mathscr{X})$ for which each p_k is positive. Observe, that this is a rather well-behaved set of distributions, e.g. all the distributions in $\mathscr{M}$ are absolutely continuous with respect to each other.

THEOREM 3. *If either of* $f(0)$ *and* $\lim_{u \to +\infty} \frac{f(u)}{u}$ *is infinite, the* (V)*-space defined by the f-neighbourhoods is no topological space in general; this is the case already if* X *is countable and* $\mathscr{M}$ *is the set of distributions described above.*

REMARK. The corresponding theorem for I-divergences of order $\alpha > 0$ has been proved in [4]; the idea of the proof is similar also in the general case, but the lack of knowledge of the concrete form of $f(u)$ causes some difficulties.

In order that a FRÉCHET (V)-space E be a topological space (in the sense that the given neighbourhood systems of the points of E are bases for the neighbourhood systems of the points of E in some topology on E) the following property is clearly necessary[2]:

[2] This property is well known to be sufficient, too, provided that each point of E possesses at least one neighbuorhood and is contained in all of its neighbourhoods.

(B) *For all points $e \in E$, each neighbourhood U of e contains another neighbourhood U' of e such that each point $e' \in U'$ possesses a neighbourhood which is a subset of U.*

We shall show that under the conditions of theorem 3 the neighbourhoods (4) do not possess property (B). We shall need two lemmas on infinite series; the first one is trivial and we formulate it only for the sake of later reference while the second one seems to be of some independent interest, too.

LEMMA 1. *If $\sum_{k=1}^{\infty} a_k$ is a convergent series with positive terms then for any two sequences of positive numbers $\varepsilon_k \to 0$ and $\gamma_k \to +\infty$ there exists a sequence $\beta_k \to +\infty$ such that $\sum_{k=1}^{\infty} \varepsilon_k a_k \beta_k < +\infty$, $\sum_{k=1}^{\infty} a_k \beta_k = +\infty$ and each β_k is equal to some γ_l.*

We omit the simple proof.

LEMMA 2. *Let $\sum_{k=1}^{\infty} a_k$ be an arbitrary convergent series with positive terms $a_k > 0$ $(k = 1, 2, \ldots)$ and $\psi(u)$ an arbitrary positive valued function defined for $u > 0$ and tending to infinity as $u \to +\infty$. Then there exists a sequence $b_k \to +\infty$ such that*

$$\sum_{k=1}^{\infty} a_k b_k < +\infty; \qquad \sum_{k=1}^{\infty} a_k b_k \psi(b_k) = +\infty. \tag{28}$$

PROOF. Let $\psi_1(u)$ be a monotone and continuous function such that $0 \leqq \psi_1(u) \leqq \psi(u)$ $(0 < u < +\infty)$ and $\lim_{u \to +\infty} \psi_1(u) = +\infty$. Let $\{n_l\}$ be a sequence of positive integers growing so rapidly that the positive numbers c_l defined by

$$\sum_{k=n_l}^{\infty} a_k = \frac{1}{c_l \psi_1(c_l)} \tag{29}$$

satisfy

$$\sum_{l=1}^{\infty} \frac{1}{\psi_1(c_l)} < +\infty. \tag{30}$$

Then, setting $b_k = \sum_{n_l \leqq k} c_l$ (of course, $b_k \to +\infty$) the conditions (28) will be satisfied. In fact, using (29) and (30), we obtain

$$\sum_{k=1}^{\infty} a_k b_k = \sum_{l=1}^{\infty} c_l \sum_{k=n_l}^{\infty} a_k = \sum_{l=1}^{\infty} \frac{1}{\psi_1(c_l)} < +\infty$$

and, since by $\psi_1(u) \leqq \psi(u)$ and the monotonicity of $\psi_1(u)$

$$b_k \psi(b_k) \geqq b_k \psi_1(b_k) \geqq \sum_{n_l \leqq k} c_l \psi_1(c_l),$$

also

$$\sum_{k=1}^{\infty} a_k b_k \psi(b_k) \geqq \sum_{l=1}^{\infty} c_l \psi_1(c_l) \sum_{k=n_l}^{\infty} a_k = \sum_{l=1}^{\infty} 1 = +\infty.$$

PROOF of Theorem 3. The distributions we are dealing with are of form $P=\{p_1, p_2, \ldots\}$, $Q=\{q_1, q_2, \ldots\}$ $(p_i>0, q_i>0, i=1, 2, \ldots)$. The f-divergence of two such (discrete) distributions equals, according to (1),

$$\mathcal{I}_f(P, Q) = \sum_{i=1}^{\infty} q_i f\left(\frac{p_i}{q_i}\right). \tag{31}$$

We have to prove that if $\lim_{u\to+0} f(u) + \lim_{u\to+\infty} \frac{f(u)}{u} = +\infty$, the property (B) does not hold for $\mathcal{M}$ and the f-neighbourhoods defined by (4). We shall prove even a bit more, namely that for each $P=\{p_1, p_2, \ldots\}\in\mathcal{M}$ and $\varepsilon>0$ there exists $Q\in U_f(P, \varepsilon)$ such that for any $\varepsilon'>0$ there exists $R\in U_f(Q, \varepsilon')$ for which $\mathcal{I}_f(R, P) = +\infty$ (i.e. $R\notin U_f(P, c)$, for any $c \leqq +\infty$). If $\sum_{k=1}^{\infty} \hat{q}_k$ and $\sum_{k=1}^{\infty} \hat{r}_k$ are two convergent series with positive terms such that

$$\sum_{k=1}^{\infty} p_k f\left(\frac{\hat{q}_k}{p_k}\right) < +\infty \tag{32}$$

$$\sum_{k=1}^{\infty} \hat{q}_k f\left(\frac{\hat{r}_k}{\hat{q}_k}\right) < +\infty, \qquad \sum_{k=1}^{\infty} p_k f\left(\frac{\hat{r}_k}{p_k}\right) = +\infty, \tag{33}$$

we may set

$$q_k = \begin{cases} cp_k & \text{if} \quad k < N \\ \hat{q}_k & \text{if} \quad k \geqq N \end{cases} \qquad c = \frac{1 - \sum_{k=N}^{\infty} \hat{q}_k}{\sum_{k=1}^{N-1} p_k} \tag{34}$$

and

$$r_k = \begin{cases} c'q_k & \text{if} \quad k < N' \\ \hat{r}_k & \text{if} \quad k \geqq N' \end{cases} \qquad c' = \frac{1 - \sum_{k=N'}^{\infty} \hat{r}_k}{\sum_{k=1}^{N'-1} q_k} \tag{35}$$

Then, if N is large enough, we have by (31) and (32)

$$\mathcal{I}_f(Q, P) = f\left(\frac{1}{c}\right) \sum_{k=1}^{N-1} p_k + \sum_{k=N}^{\infty} p_k f\left(\frac{\hat{q}_k}{p_k}\right) < f(1) + \varepsilon \tag{36}$$

and, similarly, if $N'>N$ is large enough,

$$\mathcal{I}_f(R, Q) < f(1) + \varepsilon' \tag{37}$$

while, as the series $\sum_{k=1}^{\infty} p_k f\left(\frac{\hat{r}_k}{p_k}\right)$ diverges,

$$\mathcal{I}_f(R, P) = +\infty. \tag{38}$$

(36), (37) and (38) mean just what we wanted to prove, thus our only task left is

to show that if either $f(0) = \lim_{u \to +0} f(u)$ or $\lim_{u \to +\infty} \frac{f(u)}{u}$ is infinite, there do exist convergent series $\sum_{k=1}^{\infty} \hat{q}_k$ and $\sum_{k=1}^{\infty} \hat{r}_k$ with the properties (32), (33). We distinguish two (not mutually exclusive) cases:

Case a): $f(0) = \lim_{u \to +0} f(u) = +\infty$. Let us choose a sequence $\alpha_k \to +\infty$ such that $\sum_{k=1}^{\infty} \alpha_k p_k < +\infty$ and set $\hat{q}_k = p_k g_1(\alpha_k)$, where $g_1(\alpha)$ stands for the smallest value u for which $f(u) = \alpha$ or an arbitrary fixed positive number if $f(u) > \alpha$ for all $0 < u < +\infty$. Then $\frac{\hat{q}_k}{p_k} = g_1(\alpha_k) \to 0$, thus $\sum_{k=1}^{\infty} \hat{q}_k < +\infty$ and (32) is valid. Let further $\beta_k \to +\infty$ a sequence such that $\sum_{k=1}^{\infty} \beta_k \hat{q}_k < +\infty$, $\sum_{k=1}^{\infty} \beta_k p_k = +\infty$; since $\frac{\hat{q}_k}{p_k} \to 0$ such sequences exist by lemma 1. Then, setting $\hat{r}_k = \hat{q}_k g_1(\beta_k)$, $\sum_{k=1}^{\infty} \hat{r}_k < +\infty$ and (33) holds, as $\frac{\hat{q}_k}{p_k} \to 0$ implies, for k large enough, $f\left(\frac{\hat{r}_k}{p_k}\right) > f\left(\frac{\hat{r}_k}{\hat{q}_k}\right) = \beta_k$.

Case b): $\lim_{u \to +\infty} \frac{f(u)}{u} = +\infty$. In this case a more subtle reasoning is needed. Let us apply lemma 2 to the series $\sum_{k=1}^{\infty} p_k$ and the function $\psi(u) = g_2\left(\frac{u}{g_2(u)}\right)$ where $g_2(\alpha)$ stands for the largest value u for which $f(u) = \alpha$ or an arbitrary fixed positive number if $f(u) > \alpha$ for all $0 < u < +\infty$. The validity of the condition $\lim_{u \to +\infty} \psi(u) = +\infty$ of lemma 2 follows from the assumption $\lim_{u \to +\infty} \frac{f(u)}{u} = +\infty$. Thus there exists a sequence $b_k \to +\infty$ satisfying (28) (with $a_k = p_k$). We set

$$\hat{q}_k = p_k g_2(b_k); \quad \hat{r}_k = \hat{q}_k g_2\left(\frac{b_k}{g_2(b_k)}\right) = \hat{q}_k \psi(b_k). \tag{39}$$

Then $\frac{\hat{q}_k}{p_k} \to +\infty$ $(k \to +\infty)$ and $f\left(\frac{\hat{q}_k}{p_k}\right) = b_k$ if k is large enough; thus the first half of (28) gives just (32). Furthermore, from the second half of (39) follows $f\left(\frac{\hat{r}_k}{\hat{q}_k}\right) = \frac{b_k}{g_2(b_k)} = b_k \frac{p_k}{q_k}$ (for k large enough) and thus the first half of (28) yields also the first half of (33). Finally, as $f(u)$ is convex and $\lim_{u \to +\infty} \frac{f(u)}{u} = +\infty$, for u_0 large enough necessarily $f(u) > \frac{f(u_0)}{u_0} u$ $(u > u_0)$; substituting here $u_0 = \frac{\hat{q}_k}{p_k}$ and $u = \frac{\hat{r}_k}{q_k}$, we obtain — using (39) — for k large enough

$$f\left(\frac{\hat{r}_k}{p_k}\right) > \frac{f\left(\frac{\hat{q}_k}{p_k}\right)}{\frac{\hat{q}_k}{p_k}} \cdot \frac{\hat{r}_k}{p_k} = f\left(\frac{\hat{q}_k}{p_k}\right) \frac{\hat{r}_k}{\hat{q}_k} = b_k \psi(b_k).$$

Thus the second half of (28) gives rise to the second half of (33). The proof of theorem 3 is complete.

It follows directly from theorem 3 that in the case $f(0)+\lim_{u\to+\infty}\frac{f(u)}{u}=+\infty$ there exists no strictly monotone function $h(u)$ continuous and vanishing at $u_0=f(1)$ for which $h(\mathcal{I}_f(\mu_1, \mu_2))$ would satisfy the triangle inequality. This negative result can be extended to functions of more general type, too, in the same way as it has been done for I-divergences of order $\alpha\geqq 1$ in [5]. To this end we need the following sharpening of Theorem 3:

THEOREM 4. *If either of* $f(0)$ *and* $\lim_{u\to+\infty}\frac{f(u)}{u}$ *is infinite, then for each* $P\in\mathcal{M}$ *and* $\varepsilon>0$ *there exists* $Q\in\mathcal{M}$ *such that* $\mathcal{I}_f(Q, P)+\mathcal{I}_f(P, Q)<\varepsilon$ *and that for any* $\varepsilon'>0$ *there exists* $R\in\mathcal{M}$ *with* $\mathcal{I}_f(R, Q)+\mathcal{I}_f(Q, R)<\varepsilon'$ *and* $\mathcal{I}_f(R, P)=\mathcal{I}_f(P, R)=+\infty$. *Here* $\mathcal{M}$ *stands for the set of all discrete distributions* $P=\{p_1, p_2, \ldots\}$ *such that* $p_i>0$ $(i=1, 2, \ldots)$.

PROOF. We proceed as in the proof of theorem 3, with some modifications. The construction there should be applied to the convex function $\tilde{f}(u)=f(u)+f^*(u)=$ $=f(u)+uf\left(\frac{1}{u}\right)$ (cf. remark 2 to definition 1; observe, that the condition of the theorem implies both $\tilde{f}(0)=+\infty$ and $\lim_{u\to+\infty}\frac{\tilde{f}(u)}{u}=+\infty$), and, instead of P, to the auxiliary distributions $P_i=\{p_1^i, p_2^i, \ldots\}$ $(i=1, 2)$ defined by

$$p_k^1=c_1p_{2k-1},\quad p_k^2=c_2p_{2k}\qquad (k=1, 2, \ldots) \tag{40}$$

where c_1 and c_2 are appropriate norming constants.

Let $\varepsilon>0$ be given and set $\varepsilon_i=\frac{\varepsilon}{2c_i}$, further, for arbitrary $\varepsilon'>0$ set $\varepsilon_i'=\frac{\varepsilon'}{2c_i}$. Then, as in the proof of theorem 3, we can construct $Q_i\in U_{\tilde{f}}\left(P_i, \frac{\varepsilon}{2c_i}\right)$ $(i=1, 2)$ such that for any $\varepsilon'>0$ there exist $R_i\in U_{\tilde{f}}\left(Q_i, \frac{\varepsilon'}{2c_i}\right)$ with $\mathcal{I}_{\tilde{f}}(R_i, P_i)=+\infty$ $(i=1, 2)$. We show, that this can be done in such a way that $\mathcal{I}_f(R_1, P_1)=+\infty$ and $\mathcal{I}_f(P_2, R_2)=\mathcal{I}_{f^*}(R_2, P_2)=+\infty$. In fact, let us first assume $f(0)=+\infty$, $f^*(0)=\lim_{u\to+\infty}\frac{f(u)}{u}<+\infty$. Then we have

$$f(u)\geqq f^*(u)\qquad \text{for } u \text{ small enough,} \tag{41}$$

$$f^*(u)\geqq f(u)\qquad \text{for } u \text{ large enough,} \tag{42}$$

and therefore, using for $i=1$ the construction given for case a) and for $i=2$ the one for case b) we see at once that $\mathcal{I}_{\tilde{f}}(R_i, P_i)=\mathcal{I}_f(R_i, P_i)+\mathcal{I}_{f^*}(R_i, P_i)=+\infty$ $(i=1, 2)$ implies $\mathcal{I}_f(R_1, P_1)=\mathcal{I}_{f^*}(R_2, P_2)=+\infty$. The same argument applies also when $f(0)$ is finite and $f^*(0)=\lim_{u\to+\infty}\frac{f(u)}{u}=+\infty$ with the only difference that

for $i=1$ the construction of case b) and for $i=2$ the one of case a) should be used. At last, if both $f(0)$ and $f^*(u)=\lim_{u\to+\infty}\frac{f(u)}{u}$ are infinite, it is still possible that (41) — and then, necessarily, also (42) — holds true or that the reversed inequalities are valid, in which cases the above argument applies. On the other hand, if there exist sequences of positive numbers $\delta_k^1\to 0$ and $\delta_k^2\to 0$ such that

$$f(\delta_k^1)\geqq f^*(\delta_k^1), \qquad f^*(\delta_k^2)\geqq f(\delta_k^2) \qquad (k=1, 2, \ldots), \tag{43}$$

then, according to lemma 1, the construction for case a) in the proof of theorem 3 can be applied both for $i=1$ and $i=2$ in such a way that each β_k^i be equal to some of the numbers $\tilde{f}(\delta_l^i)$ $(i=1, 2)$. Then (43) implies (41) for $u=g_1(\beta_k^1)$ and the reversed inequality for $u=g_1(\beta_k^2)$, and hence, by the proof of theorem 3, we get again

$$\mathcal{I}_f(R_1, P_1)=\mathcal{I}_{f^*}(R_2, P_2)=+\infty.$$

The proof of theorem 4 is now easily completed; all we have to do is to build from $Q_i=\{q_1^i, q_2^i, \ldots\}$, $R_i=\{r_1^i, r_2^i, \ldots\}$ $(i=1, 2)$ the distributions $Q=\{q_1, q_2, \ldots\}$ and $R=\{r_1, r_2, \ldots\}$ in the analogy of (40):

$$q_{2k-1}=\frac{1}{c_1}q_k^1, \quad q_{2k}=\frac{1}{c_2}q_k^2 \qquad (k=1, 2, \ldots)$$

$$r_{2k-1}=\frac{1}{c_1}r_k^1, \quad r_{2k}=\frac{1}{c_2}r_k^2 \qquad (k=1, 2, \ldots).$$

COROLLARY. *If either of* $f(0)$ *and* $f^*(0)=\lim_{u\to+\infty}\frac{f(u)}{u}$ *is infinite, there does not exist such a nonnegative function* $h(u, v)$ *defined for* $f(1)\leqq u\leqq+\infty$, $f(1)\leqq v\leqq+\infty$ *which is continuous at* $(f(1), f(1))$ *and vanishes there but does not vanish at* $(+\infty, +\infty)$, *that the function* $d_{h,f}(\mu_1, \mu_2)=h(\mathcal{I}_f(\mu_1, \mu_2), \mathcal{I}_f(\mu_2, \mu_1))$ *defined on* $\mathcal{M}$ *would satisfy the triangle inequality.*

REMARKS. For the case of I-divergences of order α $(\alpha\geqq 1)$ the statement of theorem 4 and its corollary has been proved in [5]. The knowledge of the explicite form of $f(u)$, however, considerably facilitated the proof. In the cases when both $f(0)$ and $f^*(0)=\lim_{u\to+\infty}\frac{f(u)}{u}$ are finite, one may look for functions of $\mathcal{I}_f(\mu_1, \mu_2)$ — or of $\mathcal{I}_f(\mu_1, \mu_2)$ and $\mathcal{I}_f(\mu_2, \mu_1)$ — which are non-trivial quasi-metrics (asymmetric „metrics") or even metrics in $\mathcal{M}$ and which, preferably, generate the same topology in $\mathcal{M}$ as the f-neighbourhoods. For such investigations, for the case of I-divergences (of order $0<\alpha<1$) see [10].

Finally let us remark, that it is easy to exhibit on $\mathcal{M}$ such a metric which makes it a complete metric space with a finer topological structure than the one of the f-neighbourhoods, for any convex $f(u)$. Such a metric is e.g.

$$\varrho'(\mu_1, \mu_2)=\min\{C, \log(\max\{q(\mu_1, \mu_2), q(\mu_2, \mu_1)\})\} \tag{44}$$

where $C>0$ is an arbitrary constant and

$$q(\mu_1,\mu_2)=\begin{cases}\text{vrai sup}\dfrac{\mu_1(dx)}{\mu_2(dx)} & \text{if} \quad \mu_1 \ll \mu_2\\ +\infty & \text{if} \quad \mu_1 \not\ll \mu_2.\end{cases}$$

The convergence $\mu_n \overset{\delta'}{\to} \mu$ is equivalent to the uniform convergence of $\dfrac{\mu_n(A)-\mu(A)}{\mu_n(A)+\mu(A)}$ ($A\in\mathcal{X}$) to zero, and this obviously implies the f-convergence for any convex $f(u)$. Of course, this implication can not be reversed, in general.

REFERENCES

[1] Csiszár, I.: Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **8** (1964) 85—108.

[2] Kullback, S.: *Information Theory and Statistics.* Wiley, New York, 1959.

[3] Rényi, A.: On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1, Berkeley, 1961; 541—561.

[4] Csiszár, I.: Informationstheoretische Konvergenzbegriffe im Raum der Wahrscheinlichkeitsverteilungen. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **7** (1962) 137—158.

[5] Csiszár, I.: Über topologische und metrische Eigenschaften der relativen Information der Ordnung α. *Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes,* Prague 1964, 63—73.

[6] Csiszár I.: *Eloszlások eltérésének információ-tipusu mértékszámai* (Information type measures of difference of distributions, in Hungarian). Thesis submitted to the Scientific Qualifying Committee of the Hungarian Academy of Sciences in January, 1966.

[7] Csiszár, I. A note on Jensen's inequality, *Studia Sci. Math. Hungar.* **1** (1966) 227—230.

[8] Csiszár, I.: On information-type measures of difference of probability distributions, *Studia Sci. Math. Hung.*

[9] Sierpiński, W.: *General Topology,* Univ. of Toronto Press, Toronto, 1966.

[10] Csiszár, I. und Fischer, J.: Informationsentfernungen im Raum der Wahrscheinlichkeitsverteilungen. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **7** (1962) 159—180.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received November 16, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 341—350.

ESTIMATION AFTER SELECTION[1]

by

K. SARKADI

1. Introduction

Let $\mu_1, \ldots, \mu_n$ be parameters characterising the populations $A_1, \ldots, A_n$ respectively. These parameters are unknown but we know their unbiassed estimators: $x_1, \ldots, x_n$:

$$\mathsf{E}(x_i) = \mu_i \qquad (i = 1, \ldots, n).$$

$x_1, \ldots, x_n$ are supposed to be independent.

According to some predetermined decision rule, one of the populations $A_1, \ldots, A_n$ will be selected. This decision is based on the actual values of the statistics $x_1, \ldots, x_n$.

This paper deals with the problem of estimating the parameter of the chosen population.

Such selection procedures often occur in practice. Usually they are performed without any theoretical statistical tools; for the theory of selection, see, e. g. [4], [2], where additional references are given.

In most cases the characterisation of the selected population remains interesting after the selection. The fact that, apart from a preliminary report [6] and an unpublished paper [7] of RUBINSTEIN, the above problem seems to have nowhere been mentioned in the literature, indicates that usually the effect of the preliminary selection is ignored in the estimation. Evidently this causes bias.

RUBINSTEIN's mentioned paper treats the estimation problem after selection according to a special sequential model.

Some related problems, however, are dealt with by several writers. The problem of estimation after preliminary tests of significance have been first treated by BANCROFT [1]. In 1963 the results of several authors on this field were reviewed and summarized by KITAGAWA [3]. In the binomial acceptance sampling model, KOLMOGOROV [4] considered the problem of the estimation of the average fraction defective in the accepted lots.

In the latter-mentioned problems the parameters to be estimated are unknown constants as far as in our model they are random variables.

As it is pointed out, regarding his model, by RUBINSTEIN, the post-selection estimation problem has particular importance at the design of equipments or systems having a large number of components. If both the number of components and the assortments are large the accumulated bias may cause quite a misleading result if the hazard rate or other characteristic quantity of the whole equipment or system is estimated.

[1] This paper has been presented to the 1st European Meeting of Statisticians held in London, 5—10 September 1966.

Sometimes more data are available in the time of estimation than were in the time of selection. Accordingly, we consider the following to models simultaneously:

Model A

With the above notations, let the selection rule be characterised by the partition of the n-dimensional Euclidian space E_n into the subsets $A_1^x, \ldots, A_n^x$ $(A_1^x + \ldots + A_n^x = E_n,\ A_i^x A_j^x = 0$ if $i \neq j,\ i, j = 1, \ldots, n)$ in such a way that if for the vector $x = \{x_1, \ldots, x_n\}$ $x \in A_i^x$ the population A_i will be selected. With these notations the task is the estimation of the random variable

$$m = \mu_i \quad \text{if} \quad x \in A_i^x \qquad (i = 1, \ldots, n). \tag{1. 1}$$

Model B

We denote by $y_1, \ldots, y_n$ the statistics available in the time of selection and $z_1, \ldots, z_n$ the additional statistics which are available in the time of the estimation. $y_1, \ldots, y_n, z_1, \ldots, z_n$ are supposed to be independent variables with

$$\mathsf{E}(y_i) = \mathsf{E}(z_i) = \mu_i \qquad (i = 1, \ldots, n).$$

We use the notation m_1 for the random variable

$$m_1 = \mu_i \quad \text{if} \quad y = \{y_1, \ldots, y_n\} \in A_i^y \qquad (i = 1, \ldots, n) \tag{1. 2}$$

where $A_1^y + \ldots + A_n^y = E_n$; $A_i^y A_j^y = 0$ for $i \neq j,\ i, j = 1, \ldots, n$. This partition of E_n is predetermined.

The variances $\mathsf{D}^2(y_i)$ and $\mathsf{D}^2(z_i)$ are supposed to exist.

In this case, for convenience, we use the notation x_i for the pooled statistics

$$x_i = \frac{\mathsf{D}^2(z_i) y_i + \mathsf{D}^2(y_i) z_i}{\mathsf{D}^2(y_i) + \mathsf{D}^2(z_i)} \qquad (i = 1, \ldots, n) \tag{1. 3}$$

On the other hand, the symbols y_i, z_i $(i = 1, \ldots, n)$, m_1 are used as auxiliary variables in Model A. Where they are used, they are supposed to fulfil the mentioned relations, i.e. $y_1, \ldots, y_n, z_1, \ldots, z_n$ are independent and m_1 is defined by (1. 2), in addition $\mathsf{D}^2(x_i), \mathsf{D}^2(y_i)$ exist and (1. 3) holds.

Section 2. enumerates some possible solutions of the problem of estimating m and m_1 in the general case. The cases of normal and Poisson distributions, $n = 2$, are dealt with in details in Sections 3. and 4., respectively.

2. The General Case

2. 1. The simplest unbiassed estimator of m_1 is the statistic

$$u = z_i \quad \text{if} \quad y \in A_i^y. \tag{2. 1}$$

Evidently u is unbiassed in the sense that $\mathsf{E}(u) = \mathsf{E}(m_1)$, moreover in this case u is conditionally unbiassed given A_i^y. The accuracy of the estimator may be characterised by the conditional variances

$$\mathsf{D}^2(u - m_1 | A_i^y) = \mathsf{D}^2(z_i) \qquad (i = 1, \ldots, n)$$

where A_i^y stands for the event $y \in A_i^y$, or the unconditional variance

$$\mathsf{D}^2(u-m_1) = \sum_{i=1}^{n} \mathsf{P}(A_i^y)\mathsf{D}^2(z_i)$$

2. 2. In general, u will be biassed for m. However, if the correlations between x_i and y_i are high, and the partitions $A_1^x, \ldots, A_n^x$ and $A_1^y, \ldots, A_n^y$ nearly agree, the probabilities $\mathsf{P}(x \in A_i^x, y \in A_j^y)$ for $i \neq j$ will be small, therefore, in general, $|\mathsf{E}(u-m)| = |\mathsf{E}(m_1)-\mathsf{E}(m)|$ will be small. We shall see in the particular cases that this bias can be brought below an arbitrary level and this can be expected to hold under rather mild conditions, in the general case of Model A.

(1. 3) shows that high correlation between x_i and y_i implies large value of $\mathsf{D}^2(z_i)$. This means, since

$$\mathsf{E}(u-m)^2 \geqq \mathsf{D}^2(u-m_1) = \sum_{i=1}^{n} \mathsf{P}(A_i^y)\mathsf{D}^2(z_i)$$

that decrease of bias will mean loss in efficiency.

In reliability applications, once we shall know the exact relationship between bias and variances the optimal estimator can be chosen from amongst the possible ones, taking the size and the structure of the whole system into account. Usually the limitation of the bias of the component estimator is the more important the larger the system is.

The exact evaluation of $\mathsf{D}^2(u-m)$ or of the conditional variances is difficult, even in the simplest particular cases.

If the selection is based on the values of $x_1, \ldots, x_n$, and m is estimated by u this method uses full information in the selection but partial information in the estimation procedure. It can be performed relatively easily, almost independently of the forms of the initial distributions. In general, the bias can be made as small as desired.

2. 3. The estimation procedure, given in 2. 2. can be ameliorated by replacing u by its conditional expected value given $x_1, \ldots, x_n$:

$$t = \mathsf{E}(u|x_1, \ldots, x_n).$$

In two concrete cases we shall give explicit formulae for t (see Eqs. (3. 4) and (4. 2)).

This modification does not change the bias but decreases the variance $\mathsf{D}^2(u-m)$. The evaluation of the variances seems to be difficult.

The estimators u and t will be investigated in the mentioned cases in Sections 3 and 4.

2. 4. Also, the problem of the existence of an unbiassed estimator for m is investigated, in the concrete cases. In the normal case, no unbiassed estimator having finite variance exists (Section 3. 3), in the Poisson case such an estimator is given, and turns out to be the limiting case of the estimator t.

2. 5. Another possible way is the "sometimes pooling" method. If, e.g. $n=2$ and

$$m = \begin{cases} \mu_1 & \text{if} \quad x_1 \leqq x_2 \\ \mu_2 & \text{if} \quad x_2 < x_1 \end{cases}$$

then — with a slight modification of (2. 2) of [3] — the following estimator can be constructed:

$$t^* = \begin{cases} x_1 & \text{if} \quad x_2 - x_1 > K \\ \dfrac{\mathsf{D}^2(x_2)x_1 + \mathsf{D}^2(x_1)x_2}{\mathsf{D}^2(x_1) + \mathsf{D}^2(x_2)} & \text{if} \quad |x_2 - x_1| \leqq K \\ x_2 & \text{if} \quad x_1 - x_2 > K \end{cases}$$

where K denotes the significance point of some predetermined level of the distribution of $|x_2 - x_1|$.

The computation of this estimator is relatively simple. It is clear that its bias can not be decreased in arbitrary degree, if $\mu_1 \neq \mu_2$. We do not deal with this type of estimator in this paper.

2. 6. A further possible way is the Bayes solution. This way avoids the paradoxical phenomenon that the estimations of m and μ_i are not equivalent problems; using Bayes' method they are. The assumption that the μ_i are random variables is a reasonable one since they are affected, in the course of the production process, by random sources; however, their a priori distributions are not known in general.

If n is large and the assumption of a common a priori distribution of a given type for the variables μ_i seems to be justified, the method is particularly appropriate. In this case the parameters of the a priori distribution may be estimated from $x_1, \ldots, x_n$.

3. Normal Distribution, $n = 2$

3. 1. Let us consider the case of the normal distribution with two variables, i.e. let us suppose x_1 and x_2 to be independent normally distributed variables with unknown expectations $\mathsf{E}(x_1) = \mu_1$ and $\mathsf{E}(x_2) = \mu_2$ and with known variances $\mathsf{D}^2(x_1) = \sigma_1^2$ and $\mathsf{D}^2(x_2) = \sigma_2^2$. Let the variable m be defined here as follows:

$$m = \begin{cases} \mu_1 & \text{if} \quad x_1 \leqq x_2 \\ \mu_2 & \text{if} \quad x_1 > x_2 . \end{cases} \tag{3. 1}$$

In the practical application this definition corresponds to the selection of the better population if less values of μ_i mean better quality.

Since

$$p = \mathsf{P}(m = \mu_2) = \mathsf{P}(x_2 - x_1 < 0) = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)$$

where $\Phi(x)$ denotes the standardised normal distribution function,

$$\mathsf{E}(m) = \mu_1 + (\mu_2 - \mu_1)p$$

and

$$\mathsf{D}^2(m) = (\mu_2 - \mu_1)^2 p(1-p).$$

The variables y_i and z_i (cf. (1. 3)) may be defined, in the case of Model A, by the following formulae:

$$y_i = x_i + v_i/c$$

$$z_i = x_i - cv_i \qquad (i = 1, 2)$$

where c is a positive number, v_1, v_2 are independent normally distributed random variables, with parameters 0, σ_1^2, and 0, σ_2^2, respectively; they are independent of x_1 and x_2. They may be generated by a table of random numbers or may be defined as functions of the original sample elements. The suitable choice of the constant c will be discussed later.

With this definition, y_1, y_2, z_1, z_2 are independent and (1. 3) is fulfilled; $\mathsf{D}^2(y_i)=(1+c^2)\sigma_i^2/c^2$, $\mathsf{D}^2(z_i)=(1+c^2)\sigma_i^2$ $(i=1, 2)$.

In accordance with Section 2, let us define m_1 and u by the formulae

$$m_1 = \begin{cases} \mu_1 & \text{if} \quad y_1 \leqq y_2 \\ \mu_2 & \text{if} \quad y_1 > y_2 \end{cases}$$

$$u = \begin{cases} z_1 & \text{if} \quad y_1 \leqq y_2 \\ z_2 & \text{if} \quad y_1 > y_2 . \end{cases}$$

Since

$$\mathsf{E}(u)=\mathsf{E}(m_1) = \mu_1+(\mu_2-\mu_1)p_c$$

where

$$p_c = \Phi\left(\frac{(\mu_1-\mu_2)c}{\sqrt{(\sigma_1^2+\sigma_2^2)(1+c^2)}}\right)$$

u is unbiassed for m_1 and biassed for m, the bias being

$$\mathsf{E}(u-m) = (\mu_2-\mu_1)(p_c-p). \tag{3. 2}$$

Evidently this bias is positive, except in the trivial case $\mu_1=\mu_2$ when the estimator is unbiassed. The bias tends to 0 if c or $|\mu_1-\mu_2|$ tends to infinity.

The variances of u and $u-m_1$ are

$$\mathsf{D}^2(u) = (1+c^2)[\sigma_1^2+(\sigma_2^2-\sigma_1^2)p_c]+(\mu_2-\mu_1)^2p_c(1-p_c)$$

$$\mathsf{D}^2(u-m_1) = (1+c^2)[\sigma_1^2+(\sigma_2^2-\sigma_1^2)p_c]$$

$$\mathsf{D}^2(u|A_i^y) = \mathsf{D}^2(u-m_1|A_i^y) = (1+c^2)\sigma_i^2 \qquad (i=1,2).$$

The variances of $u-m$ cannot be evaluated explicitely.

If u and m are positively correlated we have the inequalities

$$\mathsf{D}^2(u-m)\leqq\mathsf{D}^2(u)+\mathsf{D}^2(m)$$

i.e.

$$\mathsf{D}^2(u-m)\leqq(1+c^2)[\sigma_1^2+(\sigma_2^2-\sigma_1^2)p_c]+ \tag{3. 3}$$

$$+(\mu_2-\mu_1)^2[p_c(1-p_c)+p(1-p)].$$

In Lemma 3. 1 (Section 3. 2) a sufficient condition for being the correlation between u and m positive is given. The condition includes the case $\sigma_1=\sigma_2$.

A weaker upper bound, holding in any case, is given by the inequality

$$\mathsf{D}(u-m)<\mathsf{D}(u)+\mathsf{D}(m).$$

On the other hand we have the following lower bound:

$$\mathsf{D}(u-m)>|\mathsf{D}(u)-\mathsf{D}(m)|.$$

It follows that the variance of $u-m$ tends to infinity with increasing c, in contrast to the bias which is, as can be seen from (3. 2) a decreasing function of c.

Let us now consider the estimator of the type introduced in 2. 3.

It is evident that the estimator $t=\mathsf{E}(u|x_1, x_2)$ i.e.

$$t = x_1+(x_2-x_1)\Phi\left(\frac{c(x_1-x_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)+c\sqrt{\sigma_1^2+\sigma_2^2}\,\varphi\left(\frac{c(x_1-x_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) \tag{3. 4}$$

where $\varphi(x)=\Phi'(x)$, is a better estimator than u, since

$$\mathsf{E}(t)=\mathsf{E}(u) \tag{3. 5}$$

and

$$\mathsf{D}^2(u-m) = \mathsf{E}(\mathsf{D}^2(u-m|x_1, x_2))+\mathsf{D}^2(t-m) \tag{3. 6}$$

which implies that $\mathsf{D}^2(t-m)<\mathsf{D}^2(u-m)$.

Moreover it can be shown that for any value of x_1, x_2

$$\mathsf{D}^2(u-m|x_1, x_2) = \mathsf{D}^2(u|x_1, x_2) \geqq \frac{c^2}{\sigma_1^2+\sigma_2^2}\left[\sigma_1^2\sigma_2^2+\left(1-\frac{2}{\pi}\right)\min(\sigma_1^4, \sigma_2^4)\right] \tag{3. 7}$$

Since the direct evaluation of $\mathsf{D}^2(t-m)$ seems to be tedious, the limit we may obtain by using formulae (3. 3), (3. 6) and (3. 7) may be useful.

A lower limit can be obtained in the following way:

First a lower limit for $\mathsf{D}^2(t)$ can be obtained:

$$\mathsf{D}^2(t) \geqq \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}+$$

$$+(\sigma_1^2+\sigma_2^2)\left[p_c+\frac{c(\mu_2-\mu_1)}{\sqrt{(1+c^2)(\sigma_1^2+\sigma_2^2)}}\,\varphi\left(\frac{c(\mu_1-\mu_2)}{\sqrt{(1+c^2)(\sigma_1^2+\sigma_2^2)}}\right)-\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right]^2$$

This formula can be deduced from the CRAMÉR—RAO lower bound of the variable $t-\dfrac{\sigma_2^2x_1+\sigma_1^2x_2}{\sigma_1^2+\sigma_2^2}$, as estimator of $(\mu_2-\mu_1)\left(p_c-\dfrac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)$.

A lower limit for $\mathsf{D}^2(t-m)$ can be obtained using the inequality

$$\mathsf{D}(t-m)>\mathsf{D}(t)-\mathsf{D}(m).$$

Another lower limit may be obtained from the inequality

$$\mathsf{D}(t-m)>\mathsf{D}(t_2)-\mathsf{D}(t_1-m) \tag{3. 8}$$

where

$$t_1 = x_1+(x_2-x_1)\Phi\left(\frac{c(x_1-x_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$$

and

$$t_2 = c\sqrt{\sigma_1^2+\sigma_2^2}\,\varphi\left(\frac{c(x_1-x_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$$

Since t_1 is between x_1 and x_2, $\mathsf{D}(t_1 - m)$ has an upper bound not depending on c. On the other hand

$$\mathsf{D}^2(t_2) = \frac{c^2(\sigma_1^2+\sigma_2^2)}{2\pi}\left(\frac{1}{\sqrt{2c^2+1}}\,e^{-\frac{2c^2(\mu_1-\mu_2)^2}{(2c^2+1)(\sigma_1^2+\sigma_2^2)}} - \frac{1}{c^2+1}\,e^{-\frac{c^2(\mu_1-\mu_2)^2}{(c^2+1)(\sigma_1^2+\sigma_2^2)}}\right) \tag{3.9}$$

which means that $\mathsf{D}^2(t_2)$ and, by virtue of (3. 8), also $\mathsf{D}^2(t-m)$ tends to infinity with increasing c.

There are, in addition, two other reasons against large values for c. If c is large, the function $t(x_1, x_2)$ has the following properties in the neighbourhood of the line $x_1 = x_2$:

1. The estimator varies very rapidly with $x_1 - x_2$ and therefore rounding errors have a large effect.

2. The value of t surpasses $\max(x_1, x_2)$ by a large multiple of $\max(\sigma_1^2, \sigma_2^2)$.

In reliability estimation problems the appropriate choice of the value of c should depend on the number of components; we have to assure that the standard error and the bias of the final estimator for the reliability characteristic of the equipment or system should be of approximately the same magnitude.

3. 2. The following Lemma gives a sufficient condition for being the correlation between u and m non-negative, and thus (3. 3) in effect.

LEMMA 3. 1. *If* $(\mu_1 - \mu_2)(\sigma_2 - \sigma_1) \geqq 0$, $\operatorname{cov}(u-m) \geqq 0$.

PROOF. Excluding trivial cases, suppose without loss of generality

$$\mu_1 - \mu_2 = \mu > 0, \quad \sigma_2 \geqq \sigma_1, \quad \sigma_2 > 0. \tag{3.10}$$

Let us introduce the notation

$$u_0 = u - \frac{\sigma_2^2 z_1 + \sigma_1^2 z_2}{\sigma_1^2 + \sigma_2^2} \tag{3.11}$$

Both u_0 and m are independent of $\dfrac{\sigma_2^2 z_1 + \sigma_1^2 z_2}{\sigma_1^2 + \sigma_2^2}$ (they are functions of $z_1 - z_2$, y_1, y_2) and therefore it suffices to prove that

$$\operatorname{cov}(u_0, m) \geqq 0 \tag{3.12}$$

(3. 10) implies that

$$\mathsf{E}(u_0) = \mu\left(\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} - p_c\right) \leqq 0 \tag{3.13}$$

and that

$$\mathsf{P}(m = \mu_1)\mathsf{E}(u_0 | m = \mu_1) =$$

$$= B_1\sigma_1^2 \int_0^\infty \int_0^\eta \zeta\left[\varphi(B_2(\zeta-\mu)) - \varphi(B_2(\zeta+\mu))\right]\varphi(B_3(\eta+\mu))\,d\zeta\,d\eta + \tag{3.14}$$

$$+ B_1 \int_0^\infty \int_\eta^\infty \zeta\varphi(B_2(\zeta+\mu))\left[\sigma_2^2\varphi(B_3(\eta-\mu)) - \sigma_1^2\varphi(B_3(\eta+\mu))\right]d\zeta\,d\eta > 0$$

(Here B_1, B_2, B_3 are positive constants depending on $\sigma_1^2+\sigma_2^2$ and c).

Since m takes on two values only, (3. 10), (3. 13) and (3. 14) prove (3. 12) and hence the lemma.

On the other hand $\operatorname{cov}(u, m)$ can be negative, as it may be seen on the following counter example.

Let be

$$\mu = \mu_1 - \mu_2 > 0, \quad \sigma_2^2 = 0, \quad \sigma_1^2 > 0$$

we have (cf. 3. 13) $\mathsf{E}(u_0) > 0$ and (cf. 3. 14) $\lim_{u=+0} \mathsf{E}(u_0 | m = \mu_1) < 0$ which shows that if μ is a sufficiently small positive number $\operatorname{cov}(u_0, m) < 0$.

3. 3. In 3. 1 we gave biassed estimators for m, and the variance of those estimators tended to infinity as the bias tended to 0. Now we prove that, in fact, there is no unbiassed estimator, having finite variance, for m.

THEOREM 3. 1. *For the variable m defined by* (3. 1) *no unbiassed estimator, having finite variance, exists.*

PROOF. Suppose that such estimator exists, let us denote it by $\tau = \tau(x, x_2)$ and suppose, accordingly, that

$$\mathsf{E}(\tau) = \mu_1 + (\mu_2 - \mu_1)p, \qquad \mathsf{D}^2(\tau) < \infty.$$

Let us introduce the notations

$$\mu = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \qquad x = \frac{x_2 - x_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \qquad \tau_1 = \tau_1(x_1, x_2) = \frac{\tau - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

then

$$\mathsf{E}(\tau_1) = \mu \Phi(\mu)$$

and $\mathsf{D}^2(\tau_1)$ is finite.

$\mathsf{E}(\tau_1)$ depends on μ_1 and μ_2 through μ only therefore x is a sufficient statistic for $\mathsf{E}(\tau_1)$. If

$$\mathsf{E}(\tau_1 | x) = \tau_2(x) = \tau_2$$

then

$$\mathsf{D}^2(\tau_2) \leqq \mathsf{D}^2(\tau_1)$$

therefore $\mathsf{D}^2(\tau_2)$ is finite, we denote it by $\mathsf{D}^2(\tau_2(x)) = V(\mu)$.

Let Z be a normally distributed random variable independent to x with

$$\mathsf{E}(Z) = \frac{\mu}{c^2}$$

$$\mathsf{D}^2(Z) = \frac{1}{c^2} \qquad (c > 0).$$

Let us introduce the notations:

$$X = x + Z, \quad \mathsf{E}(\tau_2(x) | X) = \tau_3(X) = \tau_3.$$

τ_3 does not depend explicitely on μ since x is a sufficient statistic for μ.

On the other hand let us introduce the variable (cf. 3. 4)

$$\tau_4 = \tau_4(X) = \frac{c^2}{1 + c^2} X \Phi\left(\frac{c^2}{\sqrt{1 + c^2}} X\right) - \frac{c^2}{\sqrt{1 + c^2}} \varphi\left(\frac{c^2}{\sqrt{1 + c^2}} X\right)$$

τ_3 and τ_4 are both functions of the random variable X only and they have a common expectation. It follows then from the completeness of X that therefore they agree in their variance as well: $\mathsf{D}^2(\tau_4)=\mathsf{D}^2(\tau_3)\leqq V(\mu)$.

But (cf. 3. 9)

$$\mathsf{D}^2(\tau_4)>A(\mu)c-B(\mu)$$

with $A(\mu)$, $B(\mu)$ not depending on c, $A(\mu)>0$, and therefore

$$A(\mu)c-B(\mu)<V(\mu) \tag{3. 15}$$

c may take on the value of any positive number, i.e. inequality (3. 15) holds for any positive value of c. This is clearly a contradiction, $V(\mu)$ hence $\mathsf{D}(\tau)$ can not be finite.

4. Poisson Distribution, $n=2$

4. 1. Let us suppose x_1 and x_2 have Poisson distributions with expectations μ_1 and μ_2, respectively. Here we define, with a slight modification of (1. 1), m as follows:

$$m=\begin{cases}\mu_1 & \text{if} \quad x_1<x_2\\ \dfrac{\mu_1+\mu_2}{2} & \text{if} \quad x_1=x_2\\ \mu_2 & \text{if} \quad x_1>x_2\end{cases}$$

By equating the coefficients in the expressions of the expectations we obtain the following unbiassed estimator for m:

$$t_0=t_0(x_1,x_2)=\begin{cases}x_1 & \text{if} \quad x_1<x_2-1\\ \dfrac{3x_2}{2}-1 & \text{if} \quad x_1=x_2-1\\ 2x_1 & \text{if} \quad x_1=x_2\\ t_0(x_2,x_1) & \text{if} \quad x_2>x_1\end{cases} \tag{4. 1}$$

$$(x_1, x_2=0, 1, \ldots).$$

For large μ_1 and μ_2 this statistic has the disadvantage that it has high jumps nearly to the line $x_1=x_2$.

This fact hints that in this case it is not advantageous to insist on unbiassedness.

In the mentioned case the normal approximation can be used but the same method as in 3. 1 can be applied exactly for the Poisson case as well. The formulae we give below are, however, more complicated than in the former case.

4. 2. Let y_i and z_i be defined by the following formula (cf. (1. 3))

$$y_i=\alpha^{-1}v_i$$

$$z_i=\beta^{-1}(x_i-v_i) \qquad (i=1, 2)$$

where v_i is generated by a random experiment so as to have binomial distribution with parameters x_i and α; $0<\alpha<1$, $\beta=1-\alpha$.

It is known that the unconditional distributions of αy_i and βz_i, so defined, are Poisson distributions with expectations $\alpha\mu_i$ and $\beta\mu_i$, respectively, and they are independent.

Now we define $u=u(x_1, x_2, v_1, v_2)$ by the following formula (cf. 2. 1))

$$u=\begin{cases} z_1 & \text{if} \quad y_1 < y_2 \\ \dfrac{z_1+z_2}{2} & \text{if} \quad y_1 = y_2 \\ z_2 & \text{if} \quad y_1 > y_2 \end{cases}$$

Let be again

$$t=t(x_1, x_2)=\mathsf{E}(u|x_1, x_2)$$

we obtain, after some calculation, the following expression for t:

$$\begin{aligned} t = & \sum_{i=0}^{x_1}\sum_{j=0}^{i-1}\binom{x_1}{i}\binom{x_2}{j}(x_2-j)\alpha^{i+j}\beta^{x_1+x_2-i-j-1}+ \\ & +\frac{1}{2}\sum_{i=0}^{x_1}\binom{x_1}{i}\binom{x_2}{i}(x_1+x_2-2i)\alpha^{2i}\beta^{x_1+x_2-2i-1}+ \\ & +\sum_{i=0}^{x_1}\sum_{j=i+1}^{x_2}\binom{x_1}{i}\binom{x_2}{j}(x_1-1)\alpha^{i+j}\beta^{x_1+x_2-i-j-1} \end{aligned} \tag{4. 2}$$

Putting in (4. 2) $\alpha=1$, $\beta=0$ we obtain (4. 1).

REFERENCES

[1] BANCROFT, T. A.: On biases in estimation due to the use of preliminary tests of significance, *Ann. Math. Statist.* **15** (1944) 190—204.

[2] EATON, M. L.: Some optimum properties of ranking procedures, *Ann. Math. Statist.* **38** (1967) 124—137.

[3] KITAGAWA, T.: Estimation after preliminary test of significance, *University of California, Publications in Statistics* **3** (1963) 147—186.

[4] КОЛМОГОРОВ, А. Н.: Несмещенные оценки, *Изв. АН СССР, Сер. Мат.* **14** (1950) 303—326.

[5] LEHMANN, E. L.: On a theorem of Bahadur and Goodman, *Ann. Math. Statist.* **37** (1966) 1—6.

[6] RUBINSTEIN, D.: Estimation of failure rates of systems in development. *Ann. Math. Statist.* **32** (1964) 924.

[7] RUBINSTEIN, D.: On the estimation of a random parameter, To be published (1967).

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received December 8, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 351—354.

FIXING SYSTEM FOR CONVEX BODIES

by

B. BOLLOBÁS

Let T be an open convex body in E^n (n dimensional Euclidean space). Let the points P_i ($i \in I$, where I is an index set) lie on the boundary of T. The system of points P_i ($i \in I$) is called a *fixing system of* T, if there is an $\varepsilon > 0$, such that by translating T in any direction through a distance less than ε, at least one of the points P_i ($i \in I$) will get into T. This means that no matter how short a distance we want to translate to solid body in any direction, at least one of the points P_i will prevent this, if these points must not enter the interior of the body. A fixing system is said to be *primitive* if for any proper subset J of I, the points P_i ($i \in J$) do not fix T any more. FEJES TÓTH [2] suggested the problem of finding the maximal number of points which can form a primitive fixing system of an n dimensional convex body. More precisely, we are looking for the supremum $p(n)$ of the powers of all primitive fixing systems of all n dimensional convex bodies, i.e. we want determine $p(n) = \sup \{|I|^1 :$ there is an open convex n dimensional body T and a primitive fixing system P_i $(i \in I)\}$.

TOMOR [5] proved that $p(2) = 6$, and that the only extremal configurations are the convex hexagons whose opposite sides are parallel and which are fixed by the vertices. FEJES TÓTH [2] pointed out that for the rhombic dodecahedron the vertices form a primitive fixing system and consequently $p(3) \geqq 14$. DANZER [1] and HAJÓS (see [3]) showed simple n dimensional bodies having $2(2^n - 1)$ points as primitive fixing systems. It seemed likely (see [1], [4] and [5]), that there is no fixing system with more points.

The purpose of this paper is to prove the following theorem.

THEOREM. $p(n) = \mathfrak{a}$ *if* $n \geqq 3$, *but every primitive fixing system contains only a finite number of points.*

Take a convex body in E^n and a fixing system of the body. The directions of the translations can be represented by the points of S^{n-1}, the unit sphere of E^n. (If P is a point of S^{n-1}, P corresponds to the direction OP, where O is the centre of S^{n-1}.)

Every point of the fixing system excludes a certain set of translations, and the corresponding subset of S^{n-1} is easily seen to be an *open subset*. A system of points is a primitive fixing system if and only if these open subsets cover S^{n-1} but by omitting any of them the remaining sets do not cover S^{n-1}. As S^{n-1} is *compact*, a finite number of these open subsets cover S^{n-1}, and so according to the previous

[1] $|I|$ denotes the power of the index set I. In most cases I has only finitely many elements, so $|I|$ is the number of elements, but this is not necessarily so.

statement, these are *all the subsets* corresponding to the fixing points. Consequently the primitive fixing system contains only a finite number of points.

Now it will be proved that for an arbitrary number n there is a convex body in E^n having a primitive fixing system of more than n points.

Let $0<\alpha<\beta<\frac{\pi}{2}$, and let F_1 be a frustum of a right circular cone with an angle at the apex equal to 2α. Erect a right circular cone with the angle 2β, C_1 onto the smaller base of F_1. By reflecting the bodies F_1 and C_1 in the plane of the larger base of F_1, the bodies F_2 and C_2 are obtained. Denote by T the interior of the union of the closed bodies F_1, C_1, F_2, C_2 (Fig. 1). Since $\alpha<\beta$, T is a convex body.

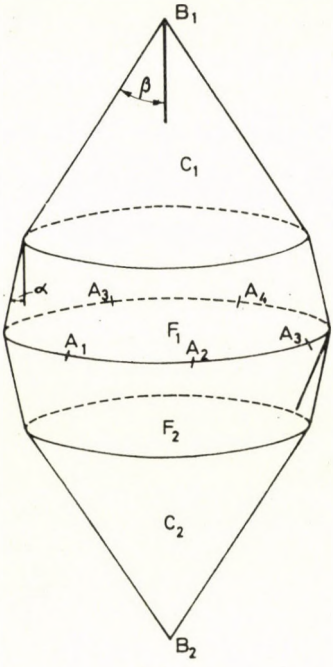


Fig. 1

Let $A_1A_2 \ldots A_m$ ($m\geqq 5$) be a regular m-gon inscribed in the common base of F_1 and F_2, and denote by B_1 and B_2 the vertices of the cones C_1 and C_2. It will be shown that for some choice of α and β the points $A_1, A_2, \ldots, A_m, B_1, B_2$ form a primitive fixing system of T.

The directions of the translations will be represented again by the points of S^2. For the sake of simplicity, the great circle parallel to the plane $A_1A_2 \ldots A_m$ will be called the equator, the point N corresponding to the direction B_2B_1 will be called the north pole and the diametrically opposite point, S, the south pole.

Take the tangent planes to F_1 and F_2 at the points A_i. These planes form four angles. Denote by α_i that angle which contains T. (This angle equals $\pi-2\alpha$.) The point A_i excludes those directions v, for which the endpoint of the vector $-v$, starting from A_i, is in the interior of α_i. Consequently A_i excludes the interior points of a spherical digon, of angle $\pi-2\alpha$.

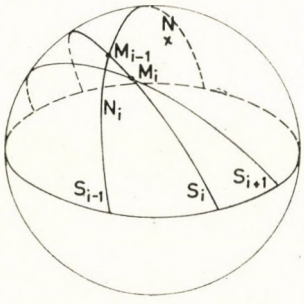


Fig. 2

Denote this open digon by S_i, and put $S_0\equiv S_m$, $S_{m+1}\equiv S_1$. S_{i+1} can be obtained from S_i by rotation around NS by $2\pi/m$ ($i=1, 2, \ldots, m$). It is easily seen that the boundaries of S_i, S_{i+1} and S_{i-1}, S_{i+1} have two common points, respectively, situated symmetrically with respect to the equator. Denote by M_i and N_i, respectively, the common points in the norther hemisphere. Then obviously $\alpha<NM_i<NN_i<\frac{\pi}{2}$ and the spherical distances NM_i, NN_i do not depend on i (see Fig. 2).

It is immediately clear that the interior points of the circle of centre N and radius β are excluded by B_1, and the corresponding domain of B_2 is the open circle of centre S and radius β.

Choose an arbitrary β in the interval $NM_i<\beta<$

$< NN_i$. For this β the domains S_i and the circles of radius β with centres N and S civer the sphere S^2, i.e. the points $A_1, A_2, \ldots, A_m, B_1, B_2$ fix T. Omitting B_1 (resp. B_2) the remaining system does not fix T, for e.g. N (resp. S) is not covered. Omitting the point A_i $(i=1, 2, \ldots, m)$ the system does not fix T, since the point N_i is contained only by S_i.

Consequently the above system of $m+2$ points is a primitive fixing system for T. The number m can be chosen arbitrary large, so this proves $p(3)=\mathfrak{a}$.

Let K be an $n-1$ dimensional convex body, having a "primitive fixing system" $P_1, P_2, \ldots, P_l$. Place K in E^n at the plane $x_1=0$ in such a way that the origin is in the interior of K. Denote by C the convex cylinder of height 2, having K as mid-cut, i.e. C contains the points $(x_1, x_2, \ldots, x_n)$, for which $|x_1| \leqq 1$ and $(0, x_2, x_3, \ldots, x_n)$ is in K. It is obvious that the points $(1, 0, 0, \ldots, 0)$, $(-1, 0, 0, \ldots, 0)$, $P_1, P_2, \ldots, P_l$ form a primitive fixing system of C. Consequently $p(n) \geqq p(n-1)$ and so $p(n)=\mathfrak{a}$ if $n \geqq 3$.

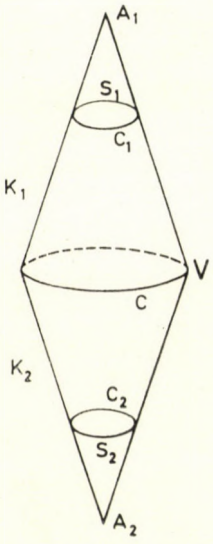


Fig. 3

Now we give a slightly different definition for the fixing systems.

Let V be an arbitrary convex body in E^n. The system of points P_i $(i \in I)$ is called a *weakly fixing system* of V, if $P_i \in V$ $(i \in I)$, but there is an $\varepsilon > 0$, such that by translating V in any direction trough at most ε, there will be a position of V in which V contains a point P_{i_0} $(i_0 \in I)$. A system of points is a *primitive weakly fixing system* of V if it weakly fixes V but no proper subset of the points fixes V any more. Let $q(n)$ denote the supremum of the powers of all primitive weakly fixing systems of all n dimensional convex bodies.

For the first sight it is a little astonishing that

$$q(2)=6 \quad \text{and} \quad q(n)=\mathfrak{c}, \quad \text{if} \quad n \geqq 3.$$

$q(2)=6$ is the result of TOMOR [5]. To prove $q(3)=\mathfrak{c}$ we construct a convex body and a primitive weakly fixing system of power $\mathfrak{c}$. Let K_1 and K_2 be two closed symmetrical right circular cones of the same base and with apices A_1, A_2. Denote by c the circumference of the base and let c_1, c_2 be circles parallel to the base on the superficies of K_1 and K_2, respectively (see Fig. 3). S_1, S_2 denote the boundary points between c_1 and A_1, and between c_2 and A_2, respectively. Finally let V be the union of K_1 and K_2, without the points of c, S_1 and S_2. V is a convex body, and it is easily checked that the points of c, together with A_1 and A_2, form a primitive weakly fixing system of V, and the power of this system is $\mathfrak{c}$. As it is immediate, that $q(n) \geqq q(n-1)$, this proves the assertion.

Similarly to the problems discussed above, one can ask the question, how "efficiently" is it possible to fix a convex body in $E.^n$ This problem was solved by GRÜNBAUM [4]. He proved that any convex body in E^n can be fixed by $2n$ points. The cube shows that fewer points are not always sufficient.

REFERENCES

[1] DANZER, L. W.: *Math. Reviews,* **26** (1963) 269—570.
[2] FEJES TÓTH, L.: On primitive polyhedra, *Acta. Math. Acad. Sci. Hungar.* **13** (1962).
[3] FEJES TÓTH, L.: New results in discrete geometry (in Hungarian) *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.,* **13** (1963) 341—354.
[4] GRÜNBAUM, B.: Fixing systems and inner illumination, *Acta. Math.* **15** (1964) 161—163.
[5] TOMOR, B.: Fixing problem for convex figures (in Hungarian), *Mat. Lapok* **14** (1963) 120—123.

EÖTVÖS L. UNIVERSITY, BUDAPEST

(Received December 9, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 355—361.

SOME EXAMPLES IN MEASURE-THEORETIC REPRESENTATION OF RANDOM VARIABLES

by

P. R. SATYAMURTY and S. S. SENGUPTA

Introduction

It is well known in the theory of probability that a random variable is a measurable function on a sample space. The purpose of this paper is to illustrate a special case of what seems to be an effective method of constructing the probability measures of random variables starting from the sample space itself. The method is one of synthesizing a random variable i.e., expressing a random variable as a finite linear combination of a family of statistically independent *elementary* random variables defined on an appropriate sample space. *Elementary* random variables with known Lebesgue measures are chosen from which are then obtained the probability measure of the random variable of interest. The advantage of such an approach follows from the fact that the relationship between the probability measure of the random variable and its structural properties can be seen clearly. The idea behind our approach is due to KAC (1959) who has obtained the probability measures of binomial and normal random variables. By a generalization of KAC's method, we derive the probability measures of multinomial and bivariate binomial random variables. Approxiate passage to limit from the bivariate binomial measure then gives the probability measure of the bivariate normal random variables. In all the cases, the method of derivation is based on the definition of probability as the mathematical expectation of the indicator function.

1. Probability Measure of a Multinomial Random Variable

Let $\Omega=\{\omega|0\leqq\omega\leqq 1\}$ be a sample space on which is defined a sequence of *elementary* random variables $\{\varepsilon_k(\omega);\ k=1, 2, \ldots, N\}$ as follows:

$$\varepsilon_1(\omega)=\begin{cases} 0 & (0 \leqq \omega \leqq \alpha_1)\\ 1 & (\alpha_1 < \omega \leqq \alpha_2)\\ 2 & (\alpha_2 < \omega \leqq \alpha_3)\\ \vdots & \vdots\ \vdots\ \vdots\\ m-1 & (\alpha_{m-1} < \omega \leqq 1)\end{cases}$$

$$\varepsilon_k(\omega)=\varepsilon_{k-1}(T_\alpha(\omega)) \qquad k=2, 3, \ldots, N \tag{1.1}$$

$$T_\alpha(\omega)=T_{\alpha_0,\alpha_1,\ldots,\alpha_m}(\omega)=$$

$$=\frac{\omega-\alpha_j}{\alpha_{j+1}-\alpha_j} \qquad (\alpha_j<\omega\leqq\alpha_{j+1})\ (j=0, 1, 2, \ldots, m-1;\ \alpha_0=0;\ \alpha_m=1)$$

LEMMA 1. *The sequence of ε's thus defined is a sequence of mutually independent random variables.*

To PROVE the lemma, it is sufficient to show

$$\Pr\{\varepsilon_2(\omega)=j|\varepsilon_1(\omega)=i\}=\Pr\{\varepsilon_2(\omega)=j\} \qquad (i,j=0,1,2,\ldots,m-1) \tag{1.2}$$

The mutual independence of the ε's would follow from it by induction. We have,

$$\Pr\{\varepsilon_2(\omega)=j|\varepsilon_1(\omega)=i\}=\frac{\Pr\{\varepsilon_2(\omega)=j;\,\varepsilon_1(\omega)=i\}}{\Pr\{\varepsilon_1(\omega)=i\}}=$$

$$=\frac{\Pr\left\{\alpha_i<\dfrac{\omega-\alpha_j}{\alpha_{j+1}-\alpha_j}\leqq\alpha_{i+1}\right\}}{\Pr\{\varepsilon_1(\omega)=i\}}=$$

$$=\frac{(\alpha_{j+1}-\alpha_j)(\alpha_{i+1}-\alpha_i)}{(\alpha_{i+1}-\alpha_i)}=$$

$$=\Pr\{\varepsilon_2(\omega)=j\}$$

which proves the lemma.

Let us write $X_N(\omega)=\sum_{k=1}^{N}\varepsilon_k(\omega)$. We are interested in the probability measure

$$\mu_N=\mu_N(x)=\mu\{\omega|X_N(\omega)=x\} \tag{1.3}$$

A representation of the indicator function of the set whose measure we wish to find is,

$$I_x=(2\pi)^{-1}\int_0^{2\pi}e^{i\xi[X_N(\omega)-x]}d\xi= \tag{1.4}$$

$$=\begin{cases}1 & (X_N(\omega)=x)\\ 0 & (\text{Otherwise}).\end{cases}$$

We have,

$$\mu_N=EI_x=(2\pi)^{-1}\int_0^1 d\omega\int_0^{2\pi}e^{i\xi[X_N(\omega)-x]}d\xi$$

After a change of the order of integration (justified by Fubini's theorem) and integrating we get,

$$\mu_N=(2\pi)^{-1}\int_0^{2\pi}e^{-i\xi x}\left[\sum_0^{m-1}p_j e^{i\xi j}\right]^N d\xi= \tag{1.5}$$

$$=\sum\binom{N}{x_0,x_1,\ldots,x_{m-1}}p_0^{x_0}p_1^{x_1}\ldots p_{m-1}^{x_{m-1}}$$

$$\sum_0^{m-1}x_j=N$$

$$\sum_0^{m-1}jx_j=x$$

where we have written $p_0=\alpha_1$, $p_j=\alpha_{j+1}-\alpha_j$ $(j=0, 1, 2, \ldots m-2)$ and $p_{m-1}=$ $=1-\alpha_{m-1}$. Now we write $X_N(\omega)=\sum_0^{m-1} jX^{(j)}(\omega)$ and interpret $X^{(j)}(\omega)$ as the number of *units* in the jth class and thus deduce the multinomial probability from (1. 5)

$$v_N = \mu\left\{\omega | X^{(j)}(\omega) = x_j; \sum_0^{m-1} x_j = N\right\} = \tag{1.6}$$

$$= \binom{N}{x_0, x_1, \ldots, x_{m-1}} p_0^{x_0} p_1^{x_1} \ldots p_{m-1}^{x_{m-1}}$$

As it is obvious, $m=2$ gives the binomial probability as a special case.

2. Bivariate Binomial Probability Measure

We shall now introduce the concept of dependence between two random variables by defining two sequences of *elementary* random variables which are independent within each sequence, but are dependent between sequences.

Let Ω be the sample space as defined in section 1. Let

$$\Phi(\omega) = \begin{matrix} 1 & (0 \leqq \omega \leqq \alpha) \\ 0 & (\alpha < \omega \leqq 1) \end{matrix}$$

$$\psi(\omega) = \begin{matrix} 1 & (0 \leqq \omega \leqq \beta) \\ 0 & (\beta < \omega \leqq 1) \end{matrix} \tag{2.1}$$

where β is a function of α determined uniquely by the following probability measures:

$$\mu_{rs} = \mu\{\omega: \psi(\omega)=s | \Phi(\omega)=r\} \qquad (r, s=0, 1). \tag{2.2}$$

The relationship between α and β is expressed conveniently by

$$(1-\beta, \beta) = (1-\alpha, \alpha)\begin{pmatrix} \mu_{00} & \mu_{01} \\ \mu_{10} & \mu_{11} \end{pmatrix} \tag{2.3}$$

We now define two sequences of random variables as follows:

$$\Phi_1(\omega) = \varphi(\omega)$$

$$\Phi_k(\omega) = \varphi_{k-1}(T_\alpha(\omega)) \qquad k=2, 3, \ldots, N$$

where,

$$T_\alpha(\omega) = \begin{cases} \dfrac{\omega}{\alpha} & (0 \leqq \omega \leqq \alpha) \\ \dfrac{\omega-\alpha}{1-\alpha} & (\alpha < \omega \leqq 1) \end{cases}$$

$$\psi_1(\omega) = \psi(\omega) \tag{2.4}$$

$$\psi_k(\omega) = \psi_{k-1}(T_\beta(\omega)) \qquad k = 2, 3, \ldots, N$$

where,

$$T_\beta(\omega) = \begin{cases} \dfrac{\omega}{\beta} & (0 \leqq \omega \leqq \beta) \\ \dfrac{\omega-\beta}{1-\beta} & (\beta < \omega \leqq 1) \end{cases}$$

LEMMA 2. *Let* $\Gamma_k(\omega)=(\Phi_k(\omega), \psi_k(\omega))$ $k=1, 2, \ldots, N$. *The sequence of the vector random variables* $\{\Gamma_k(\omega);\ k=1, 2, \ldots, N\}$ *are mutually independent.*

PROOF follows on similar lines as in Lemma 1. We have, for instance,

$$\Pr\{\Gamma_2(\omega)=(0, 0)|\Gamma_1(\omega)=(0, 0)\}=(1-\alpha)\mu_{00}=\Pr\{\Gamma_2(\omega)=(0, 0)\}.$$

We now define the sums $X_N(\omega)=\sum_{k=1}^{N}\Phi_k(\omega)$ and $Y_N(\omega)=\sum_{k=1}^{N}\psi_k(\omega)$ and wish to determine the probability measure

$$\mu_N(x, y)=\mu\{\omega|X_N(\omega)=x;\ Y_N(\omega)=y\}. \tag{2. 5}$$

With the following representation of the indicator function of the set

$$I_{x\cap y} = (2\pi)^{-2}\int_0^{2\pi}\int_0^{2\pi} e^{i\xi[X_N(\omega)-x]+i\eta[Y_N(\omega)-y]}d\xi\, d\eta = \tag{2. 6}$$

$$= \begin{cases} 1 & (X_N(\omega)=x;\ Y_N(\omega)=y) \\ 0 & (\text{Otherwise}) \end{cases}$$

We have,

$$\mu_N(x, y) = EI_{x\cap y} =$$

$$= (2\pi)^{-2}\int_0^1 d\omega \int_0^{2\pi}\int_0^{2\pi} e^{i\xi[X_N(\omega)-x]+i\eta[Y_N(\omega)-y]}d\xi\, d\eta$$

Integrating, after a change of the order of integration (justified by similar arguments as in section 1) because of the mutual independence of the sequence $\{\Gamma_k(\omega);\ k=1, 2, \ldots, N\}$ one obtains,

$$\mu_N(x, y) = (2\pi)^{-2}\int_0^{2\pi}\int_0^{2\pi} e^{-i[\xi x+\eta y]}\, G^{(N)}(\xi, \eta)d\xi\, d\eta \tag{2. 7}$$

where,

$$G^{(N)}(\xi, \eta) = [(1-\alpha)(\mu_{00}+\mu_{01}e^{i\eta})+\alpha e^{i\xi}(\mu_{10}+\mu_{11}e^{i\eta})]^N$$

whence,

$$\mu_N(x, y) = \binom{N}{x}\alpha^x(1-\alpha)^{N-x}\sum_{k=0}^{\min(x, y)}\binom{x}{k}\mu_{10}^{x-k}\mu_{11}^{k}\binom{N-x}{y-k}\mu_{00}^{N-x-y+k}\mu_{01}^{y-k} \tag{2. 8}$$

This is what we have called the bivariate binomial probability. There are several problems in applied probability where the bivariate binomial model can be used to describe the underlying process. We shall describe one such application which has been pointed out to us by Professor A. RÉNYI. We consider a source which emits independent signals with the associated values 1 and 0 which are

assumed with probabilities α and $1-\alpha$ respectively. These signals are then transmitted through a channel one by one with the transition probabilities μ_{rs} $(r, s=0, 1)$ i.e., μ_{rs} is the conditional probability that the output has the value s given that the input is r. If $X_N(\omega)$ denotes the number of input signals which take the value 1 and $Y_N(\omega)$ the corresponding number of output signals which are equal the value 1, then our result (2. 8) gives the joint distribution of $X_N(\omega)$, $Y_N(\omega)$.

It is easily verified that the marginal distributions are binomial. We shall show in the next section that the bivariate binomial probability tends to the bivariate normal.

3. Bivariate Normal Probability

We first consider the two sequences of *elementary* random variables defined in (2. 4) and define the following sequences of *elementary* random variables.

$$\nu_k(\omega) = \frac{\Phi_k(\omega)-\alpha}{\sqrt{N\alpha(1-\alpha)}}$$

$$\delta_k(\omega) = \frac{\psi_k(\omega)-\beta}{\sqrt{N\beta(1-\beta)}} \tag{3. 1}$$

so that,

$$\nu_k(\omega) = \begin{cases} \sqrt{\dfrac{1-\alpha}{N\alpha}} & (0 \leqq \omega \leqq \alpha) \\ -\sqrt{\dfrac{\alpha}{N(1-\alpha)}} & (\alpha < \omega \leqq 1) \end{cases}$$

$$\delta_k(\omega) = \begin{cases} \sqrt{\dfrac{1-\beta}{N\beta}} & (0 \leqq \omega \leqq \beta) \\ -\sqrt{\dfrac{\beta}{N(1-\beta)}} & (\beta < \omega \leqq 1) \end{cases} \tag{3. 2}$$

Let us write,

$$U_N(\omega) = \sum_{k=1}^{N} \nu_k(\omega)$$

$$V_N(\omega) = \sum_{k=1}^{N} \delta_k(\omega)$$

We shall determine

$$\mu_N = \mu\{\omega: x_1 < U_N(\omega) < x_2\,;\, y_1 < V_N(\omega) < y_2\}$$

and derive the required probability by an appropriate passage to limit

$$\mu = \lim_{N\to\infty} \mu_N$$

Let

$$g(x, y) = \begin{cases} 1 & (x_1 < U_N(\omega) < x_2\,;\, y_1 < V_N(\omega) < y_2) \\ 0 & \text{(Otherwise).} \end{cases}$$

Then using Fourier's formula [see e.g., SNEDDON (1951)] we can write,

$$g(x,y) = (2\pi)^{-2} \int_{-\infty}^{\infty}\int\int\int g(u,v)\, e^{i\xi[u-x]+i\eta[v-y]}\, du\, dv\, d\xi\, d\eta \tag{3.3}$$

Thus,

$$\mu_N = \int_0^1 g\left(\sum_1^N v_k(\omega),\ \sum_1^N \delta_k(\omega)\right) d\omega$$

$$= (2\pi)^{-2} \int_0^1 d\omega \int_{-\infty}^{\infty}\int\int\int g(u,v)\, e^{i\xi\left[u-\sum_1^N v_k(\omega)\right]+i\eta\left[v-\sum_1^N \delta_k(\omega)\right]}\, du\, dv\, d\xi\, d\eta \tag{3.4}$$

Changing the order of integration and integrating with respect to first we have,

$$\mu_N = (2\pi)^{-2} \int_{-\infty}^{\infty}\int\int\int e^{i\xi u+i\eta v}\, H^{(N)}(\xi,\eta)\, du\, dv\, d\xi\, d\eta \tag{3.5}$$

where

$$H^{(N)}(\xi,\eta) = \left[(1-\alpha)e^{i\xi/\theta_1\sqrt{N}}\left\{\mu_{00}e^{i\eta/\theta_2\sqrt{N}} + \mu_{01}e^{-i\eta\theta_2/\sqrt{N}}\right\} + \right.$$
$$\left. + \alpha e^{-i\xi\theta_1/\sqrt{N}}\left\{\mu_{10}e^{i\eta/\theta_2\sqrt{N}} + \mu_{11}e^{-i\eta\theta_2/\sqrt{N}}\right\}\right]^N$$

where we have written $\sqrt{\left(\frac{1-\alpha}{\alpha}\right)} = \theta_1$ and $\sqrt{\left(\frac{1-\beta}{\beta}\right)} = \theta_2$. After simplification, we can write,

$$H^{(N)}(\xi,\eta) = \left[1 - \frac{1}{2N}(\xi^2+\eta^2+2\xi\eta\varrho) + o\left(\frac{1}{N}\right)\right]^N \tag{3.6}$$

where $\varrho = \frac{1}{\theta_1\theta_2}[\mu_{00}(1-\alpha) - \theta_1^2\alpha\mu_{10} - (1-\alpha)\theta_2^2\mu_{01} + \alpha\mu_{11}\theta_1^2\theta_2^2]$. It follows that,

$$H(\xi,\eta) = \lim_{N\to\infty} H^{(N)}(\xi,\eta) = e^{-\frac{1}{2}(\xi^2+\eta^2+2\xi\eta\varrho)} \tag{3.7}$$

We now have,

$$\mu = \lim_{N\to\infty} \mu_N = (2\pi)^{-2} \int_{-\infty}^{\infty}\int\int\int g(u,v)\, e^{i\xi u+i\eta v-\frac{1}{2}[\xi^2+\eta^2+2\xi\eta\varrho]}\, du\, dv\, d\eta\, d\xi \tag{3.8}$$

In (3. 5) we have changed the order of integration. In (3. 8) we have taken the limit under the integral sign. Since the limits of integration are $-\infty$ and ∞, the integrand is not absolutely integrable. However, by argument similar to that in KAC [see p. 38—39], the integrand can be made absolutely integrable and hence the operations we have carried out are valid. Thus, we obtain, finally

$$\mu = \mu\{\omega: x_1 < U_N(\omega) < x_2;\ y_1 < V_N(\omega) < y_2\}$$
$$= \frac{1}{2\pi(1-\varrho^2)^{\frac{1}{2}}} \int_{x_1}^{x_2}\int_{y_1}^{y_2} e^{-\frac{1}{2}\left[\frac{u^2+v^2-2uv\varrho}{1-\varrho^2}\right]} du\, dv \tag{3.9}$$

which is the bivariate normal distribution with the correlation coefficient ϱ.

REFERENCES

[1] KAC, M.: *Statistical Independence in Probability, Analysis and Number Theory,* published by the Mathematical Society of America, 1959.
[2] SNEDDON, I. A.: *Fourier Transforms,* McGraw-Hill, 1951; p. 16.

OPERATIONS RESEARCH DEPARTMENT, CORNING GLASS WORKS, CORNING, NEW YORK
DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH, UNIVERSITY OF PENNSYLVANIA, PHILADELPHIA

(Received December 9, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 363—367.

ON THE NUMBER OF EQUAL DISCS THAT CAN TOUCH ANOTHER OF THE SAME KIND

by

L. FEJES TÓTH

Balls, translates of a body, higher neighbours. On a table we can put at most six nickels around a nickel each touching the middle one. To start with, we mention some problems and results which are closely related to this simple fact.

COXETER [1] asked about the maximal number N_n of equal rigid „material" balls in Euclidean n-space which can be brought into contact with a ball of the same size. The case when $n=3$ has an interesting story which started with cosmogonic problems discussed by NEWTON and DAVID GREGORY and ended with various proofs of the fact that $N_3=12$. For $n>3$ the value of N_n is not known. But special constructions show that $N_4 \geqq 24$, $N_5 \geqq 40$, $N_6 \geqq 72$, $N_7 \geqq 126$ and $N_8 \geqq 240$. On the other hand, a result of BÖRÖCZKY [2] implies $N_4 \leqq 26$, and we have good reason for conjecturing that $N_5 \leqq 48$, $N_6 \leqq 85$, $N_7 \leqq 146$ and $N_8 \leqq 244$.

The following nice result is an immediate consequence of a theorem due to HADWIGER and DEBRUNNER [3] and GRÜNBAUM [4]: In n-space a convex body cannot be touched by more than 3^n-1 non-overlapping translates of the body. The number 3^n-1 is attained only by the parallelotope.

A body b touching the body a is said to be a neighbour, or first neighbour, of a. A body other than a touching b is said to be second neighbour of a, etc. In a set of nickels, let T_n denote the maximum of the total number of the first, second, ... and n-th neighbours of one nickel. It may be conjectured that for „small" values of n, say for $n \leqq 10$, we have $T_n=3n(n+1)$. But this rule must soon break down, since $\lim_{n\to\infty} T_n/n^2 = 2\pi/\sqrt{3}$. We intend to return to this and to some analogous problems in another paper.

Convex discs. In the present paper we shall deal with another variant of the "nickel problem". In the Euclidean plane we consider an arbitrary convex disc. We want to give an upper bound for the number of congruent (not necessarily translated) replicas of the disc that can touch the original disc, in terms of some simple data of the disc. We choose two data which can be measured by a slide-gauge: the maximum and the minimum of the breadth of the disc in various directions, i.e. the diameter and the width. Our main result is contained in the following

THEOREM. *A convex disc with diameter d and width w cannot be touched by more than*

$$\left[(4+2\pi)\frac{d}{w}+2+\frac{w}{d}\right] \tag{1}$$

non-overlapping replicas of the disc.

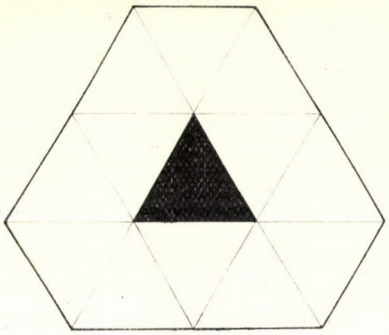


Fig. 1

For a regular triangle (1) equals 14. This bound favorably compares with the number 12 of equal regular triangles which can be put around a regular triangle of the same size (Fig. 1). Even more interesting is the case when the quotient w/d is small. If $w/d = \sin \pi/19$, the bound (1) equals 64. This bound is exact. Fig. 2 shows a set of congruent isosceles triangles with the above value of the quotient w/d in which the central triangle has exactly 64 neighbours. To construct this set, we observe that the quotient of the leg-length and base-length of one triangle is $1/2 \sin \pi/38 > 19/\pi > 6$. Thus we can put $2 \cdot 6$ triangles with their bases on the two legs, $2 \cdot 5$ triangles between the above triangles touching the legs with their apices, 19 triangles around the apex, $2 \cdot 11$ around the remaining vertices and one triangle with its base on the base.

We will prove the above theorem in a slightly sharper form, showing that in (1) w can be replaced by the breadth b of the disc in the direction perpendicular to a diameter. Since $b \geqq w$ and for $q \geqq 1$

$$(4+2\pi)q+2+1/q$$

is an increasing function of q, the new bound is, in fact, sharper than the original bound (1).

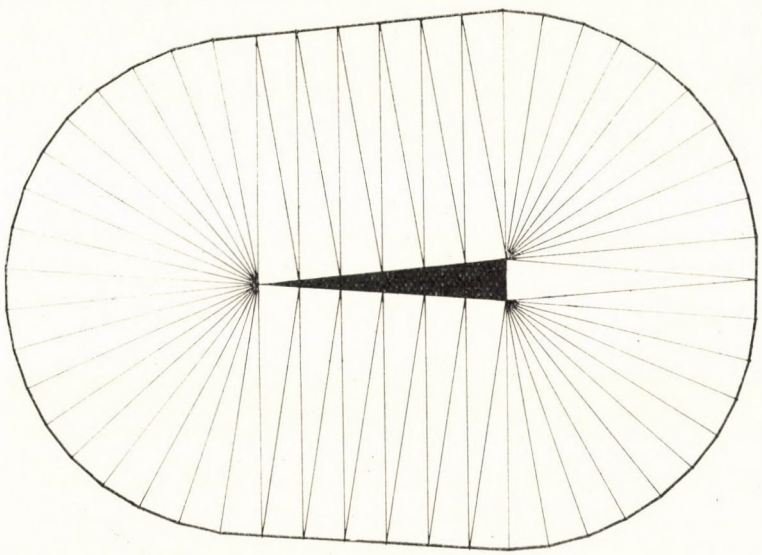


Fig. 2

We denote the area of a disc by a, its perimeter by p and the number of its neighbours by n. Since the neighbours are contained in the parallel domain of the disc at distance d, we have

$$na \leqq pd + \pi d^2. \tag{2}$$

We claim that

$$\frac{p}{a} < \frac{d + b + \sqrt{d^2 + b^2}}{\frac{1}{2}bd}. \tag{3}$$

To show this, we observe that the disc is inscribed into a rectangle with side-lengths d and b in such a way that it has a pair of points on the sides of length b equally distant from one of the sides of length d. The inequality (3) will be proved by showing that under these conditions the quotient p/a attains its maximum for a triangle with perpendicular sides of length d and b.

We suppose the shorter side of the rectangle to be in a vertical position. Furthermore, we may suppose, without loss of generality, that the disc is a polygon Π. We translate each horizontal chord of Π to the right until it reaches the right side of the rectangle (Fig. 3). The translated chords form a new polygon Π'. Obviously,

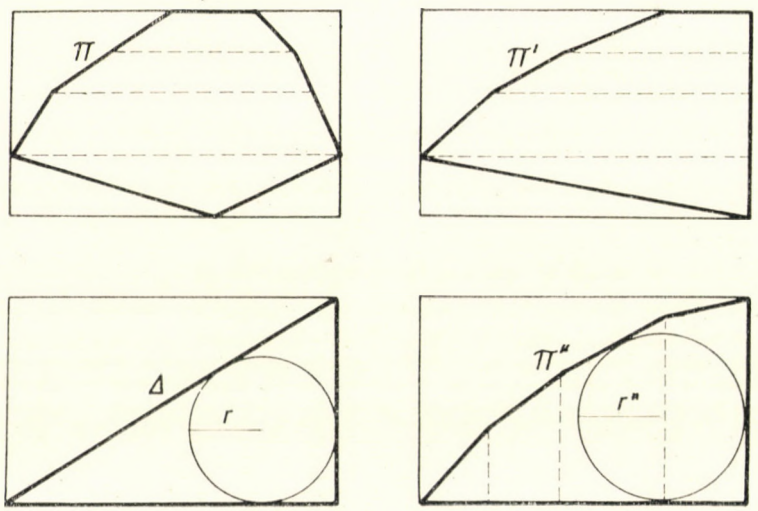


Fig. 3

this operation preserves the convexity, as well as the area. On the other hand, the perimeter will be increased by this operation. This can be easily seen by comparing the perimeters of the trapezoids and triangles with horizontal sides, into which Π and Π' can be decomposed.

We apply the same operation to Π', translating its vertical chords downwards until they hit the lower side of the rectangle. The new polygon Π'' will have two sides which coincide with the right and lower sides of the rectangle. Since Π'' contains

the right triangle Δ spanned by these two sides, we have, in accordance with (3),

$$\frac{p}{a} < \frac{p''}{a} \leqq \frac{2}{r''} \leqq \frac{2}{r} = \frac{d+b+\sqrt{d^2+b^2}}{\frac{1}{2}bd},$$

where p'' is the perimeter of Π'', r'' its inradius and r the inradius of Δ.

Combining (2) and (3), we obtain

$$n < 2\frac{d+b+\sqrt{d^2+b^2}}{b} + \pi\frac{d^2}{a},$$

whence, in view of $a \geqq \frac{1}{2}bd$,

$$n < 2\left(q+1+q\sqrt{1+q^{-2}}\right)+2\pi q,$$

where $q=d/b$. Since

$$\sqrt{1+q^{-2}} < 1+\frac{1}{2}q^{-2},$$

we have

$$n<(4+2\pi)q+2+1/q.$$

This completes the proof of our theorem.

The same method shows that in a set of congruent discs the total number of the first, second, ... and k-th neighbours of a disc is less than

$$(4k+2\pi k^2)q+2k+k/q.$$

Discs of constand breadth. For values of d/w close to 1 the bound (1) is rough. To finish, we give a better estimate in the case when $d/w=1$, i.e. for discs of constant

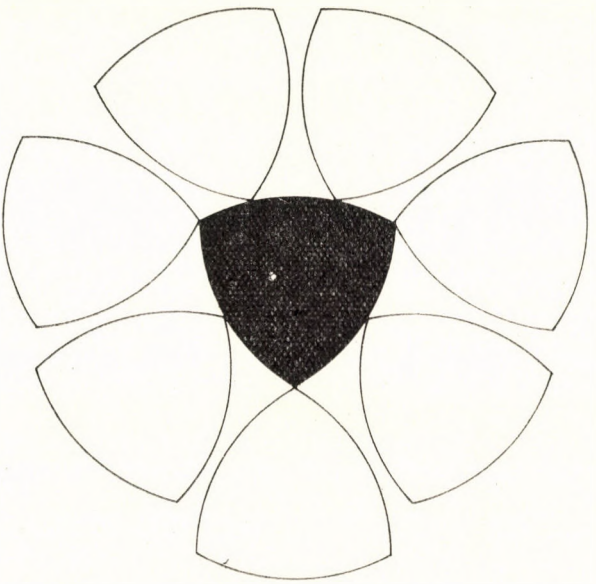


Fig. 4

breadth. Now we have

$$p = \pi d \tag{4}$$

and

$$a \geqq \frac{\pi - \sqrt{3}}{2} d^2. \tag{5}$$

The well-known equality (4) is due to BARBIER. The inequality (5) expresses the fact, due to Lebesgue, that of the domains of prescribed constant breadth the REULEAUX-triangle has the least area.

Combining (2), (4) and (5), we obtain the inequality

$$n \leqq \frac{4\pi}{\pi - \sqrt{3}} = 8.915\ldots.$$

Thus *a disc of constant breadth cannot have more than* 8 *congruent neighbours*. Can the number 8 be replaced by 7? It is very probable that the answer is: Yes. Fig. 4 exhibits a REULEAUX-triangle having 7 congruent neighbours.

REFERENCES

[1] COXETER, H. S. M.: An upper bound for the number of equal nonoverlapping spheres that can touch another of the same size, *Proceedings of Symposia in Pure Mathematics*, Volume VII, Convexity, Amer. Math. Soc. 1963, pp. 53—71.

[2] BÖRÖCZKY, K. and FLORIAN, A.: Über die dichteste Kugelpackung im hyperbolischen Raum, *Acta Math. Acad. Sci. Hungar.* **15** (1964) 237—245.

[3] HADWIGER, H. DEBRUNNER, H. and KLEE, V.,: *Combinatorial Geometry in the Plane*, New York, 1964; Theorem 43, p. 18.

[4] GRÜNBAUM, B.: On a conjecture of Hadwiger, *Pacific J. Math.* **11** (1961) 215—219.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received December 10, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 369—383.

SOME NEW RESULTS IN THE THEORY OF STABILITY

(Proof of a conjecture of A. M. Aizerman)

by

T. FREY

§ 1. Introduction

Modern control-technique has given great importance to the investigation of the LIAPUNOFF-stability of solutions of systems of differential equations. From the point of view of practical applications AIZERMAN's researches seem to be particularly important. He was the first to deal with the stability of control circuits containing also relays or switching elements with diods. The problem has decisive importance from the point of view of the general theory of control circuits containing also nonlinear elements since the characteristics of the nonlinear elements can be well approximated by a polygonal path and the latter can be realized by elements consisting of switching elements and elements of linear characteristics.

The control circuits of the simplest structure — containing a single switching element or an element of nonlinear characteristics resp. — are described by AIZERMAN by the system

$$\frac{dx_1}{dt} = \sum_{j=1}^{n} a_{1j}x_j + f(x_k) \tag{1}$$

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} a_{ij}x_j. \qquad (i = 2, 3, \ldots, n)$$

AIZERMAN supposed that $f(x_k)$ can be limited by two linear functions, i.e. $f(x_k)$ satisfies the inequality

$$a_1 x_k \leqq f(x_k) \leqq a_2 x_k. \tag{2}$$

Besides, the unicity of solution is provided by the Lipschitz-condition valid for $f(x_k)$. Finally AIZERMAN supposed that the solution $\mathbf{x} \equiv \mathbf{o}$ of the system

$$\dot{x}_1 = \sum_{j=1}^{n} a_{1j}x_j + ax_k; \qquad \dot{x}_i = \sum_{j=1}^{n} a_{ij}x_j \tag{3}$$

is asymptotically stable if $t \to \infty$ when $a \in (A_1, A_2)$ and $[a_1, a_2] \subset (A_1, A_2)$, (see e.g. [1]).

AIZERMAN suspected that — on the above assumption — the solution $\mathbf{x} \equiv \mathbf{o}$ of system (1) is also stable asymptotically, but he could not prove this conjecture but for $n = 2$.

Below we shall give general results in the stability-theory by means of which we can make a conclusion regarding the stability of control systems containing a great number of switching elements or nonlinear subsystems as well. We get the proof of the theorem suspected by AIZERMAN — in a form more general than the original — by means of the theorems we shall prove later.

§ 2. deals with the stability of the perturbed linear systems. On the basis of the results obtained in this way we shall develop further a theorem of CESARI (see [2]), and its generalization by BIHARI as well.

§ 3. deals with linear systems having piecewise constant coefficient matrices and perform several corollaries of the theorem we obtained.

Finally in § 4 we shall prove AIZERMAN's conjecture in an essentially generalized form.

§ 2. Stability of Perturbed Linear Systems and Perturbed almost Linear Systems respectively

There are many known results about asymptotic behaviours of the solutions of the linear system

$$\dot{\mathbf{x}} = [\mathbf{A}(t) + \mathbf{B}(t)]\mathbf{x} \tag{4}$$

in cases when the solutions of the linear system

$$\dot{\mathbf{y}} = \mathbf{A}(t) \cdot \mathbf{y} \tag{5}$$

are known and they may either be majorized as well as minorized by some exponential function or $\lim_{t\to\infty} \int_0^t \operatorname{trace} \mathbf{A}\, d\tau > -\infty$ is valid.

Below we shall prove a theorem in which, when comparing asymptotically the solutions of systems (4) and (5), we shall only assume that $\mathbf{A}$ and $\mathbf{B}$ are integrable on every finite interval.

THEOREM 1. *Let* $\mathbf{Y}(t)$ *and* $\mathbf{X}(t)$ *be fundamental matrices of* (5) *and* (4) *respectively. Then the relations*

$$\|\mathbf{X}^{-1}(t)\mathbf{Y}(t) - \mathbf{E}\| \leqq \int_{t_0}^{t} \|\mathbf{B}(\tau)\|\, d\tau \exp\left\{\int_{t_0}^{t} \|\mathbf{B}(\tau)\|\, d\tau\right\} \tag{6}$$

and

$$\|\mathbf{Y}^{-1}(t)\mathbf{X}(t) - \mathbf{E}\| \leqq \int_{t_0}^{t} \|\mathbf{B}(\tau)\|\, d\tau \exp\left\{\int_{t_0}^{t} \|\mathbf{B}(\tau)\|\, d\tau\right\} \tag{7}$$

which characterize the asymptotic behaviours of fundamental matrices $\mathbf{Y}(t)$ *and* $\mathbf{X}(t)$ *resp. are valid provided* $\mathbf{X}(t_0) = \mathbf{Y}(t_0)$. *Here* $\|,\|$ *denotes an arbitrary matrix norm which is invariant with respect to similarity transformations. If* $\int_{t_0}^{\infty} \|\mathbf{B}(\tau)\| d\tau < \infty$ *is fulfilled as well, then for every* $\mathbf{Y}(t)$ *we can find a fundamental matrix* $\mathbf{X}^0(t)$ *such that*

$$\|\mathbf{X}^0(t)^{-1}\mathbf{Y}(t) - \mathbf{E}\| \leqq 2\|\mathbf{E}\| \int_{t}^{\infty} \|\mathbf{B}(\tau)\| d\tau \tag{8}$$

and

$$\|\mathbf{Y}^{-1}(t)\mathbf{X}^0(t) - \mathbf{E}\| \leqq 2\|\mathbf{E}\| \int_{t}^{\infty} \|\mathbf{B}(\tau)\| d\tau \tag{9}$$

are fulfilled for any sufficiently large t *respectively.*

PROOF. Let us consider the matrix equation

$$\dot{\mathbf{X}} = (\mathbf{A}+\mathbf{B})\mathbf{X} \tag{10}$$

which corresponds to (4). Then (10) may be solved applying the well-known method of successive approximations (with the initial condition $\mathbf{X}(t_0)=\mathbf{Y}(t_0)$) by means of the iterative sequence

$$\mathbf{X}_{n+1}(t) = \mathbf{Y}(t)+\int_{t_0}^{t}\mathbf{Y}(t)\mathbf{Y}^{-1}(\xi)\mathbf{B}(\xi)\mathbf{X}_n(\xi)d\xi \tag{11}$$

Let us consider the following transformation:

$$\mathbf{T}(\mathbf{U}) = \mathbf{Y}(t)+\int_{t_0}^{t}\mathbf{Y}(t)\mathbf{Y}^{-1}(\xi)\mathbf{B}(\xi)\mathbf{U}(\xi)d\xi \tag{12}$$

and (in the space of n-dimensional quadratic matrices integrable on every finite interval) let us introduce the following pseudo-metric (see [5])

$$\varrho(\mathbf{U},\mathbf{V})=\|\mathbf{Y}^{-1}(t)\{\mathbf{U}(t)-\mathbf{V}(t)\}\|. \tag{13}$$

Here $\|,\|$ is an arbitrary matrix norm which is invariant with respect to the similarity transformations and satisfies the relation $\|\mathbf{CD}\|\leqq\|\mathbf{C}\|\|\mathbf{D}\|$. Obviously we assume $\varrho\in L$ and $\varrho_1\leqq\varrho_2$ means that almost everywhere $\varrho_1(t)\leqq\varrho_2(t)$.

Thus

$$\varrho(\mathbf{T}(\mathbf{U}),\mathbf{T}(\mathbf{V})) = \left\|\mathbf{Y}^{-1}(t)\int_{t_0}^{t}\mathbf{Y}(t)\mathbf{Y}^{-1}(\xi)\mathbf{B}(\xi)\{\mathbf{U}(\xi)-\mathbf{V}(\xi)\}d\xi\right\| =$$

$$\left\|\int_{t_0}^{t}\{\mathbf{Y}^{-1}(\xi)\mathbf{B}(\xi)\mathbf{Y}(\xi)\}\{\mathbf{Y}^{-1}(\xi)[\mathbf{U}(\xi)-\mathbf{V}(\xi)]\}d\xi\right\| \leqq \tag{15}$$

$$\leqq\int_{t_0}^{t}\|\mathbf{Y}^{-1}(\xi)\mathbf{B}(\xi)\mathbf{Y}(\xi)\|\,\varrho(\mathbf{U},\mathbf{V})d\xi=\int_{t_0}^{t}\|\mathbf{B}(\xi)\|\,\varrho(\mathbf{U},\mathbf{V})d\xi$$

(if $t\geqq t_0$).

Consequently we may apply theorems 1., 2., 3. of [5] choosing $P_n=0$,

$$Q\varrho=\int_{t_0}^{t}\|\mathbf{B}(\xi)\|\,\varrho(\xi)d\xi \tag{16}$$

and $v=w_0$, $\sigma_0=0$.

In this case

$$\sigma_1=\tau\leqq\varrho(\mathbf{U}_0,\mathbf{U}_1)=\left\|\mathbf{Y}^{-1}(t)\left\{\mathbf{Y}(t)+\int_{t_0}^{t}\mathbf{Y}(t)\mathbf{Y}^{-1}(\xi)\mathbf{B}(\xi)\mathbf{Y}(\xi)-\mathbf{Y}(t)\right\}\right\| =$$

$$=\left\|\int_{t_0}^{t}\mathbf{Y}^{-1}(\xi)\mathbf{B}(\xi)\mathbf{Y}(\xi)d\xi\right\|\leqq\int_{t_0}^{t}\|\mathbf{B}(\xi)\|d\xi \tag{17}$$

Moreover σ_∞ satisfies the equation

$$\sigma_\infty = \tau + \int_{t0}^{t} \|\mathbf{B}(\xi)\| \sigma_\infty(\xi) d\xi \tag{18}$$

and thus by the BELLMAN Lemma

$$\sigma_\infty - \sigma_0 = \sigma_\infty = \tau \exp\left\{\int_{t_0}^{t} \|\mathbf{B}(\xi)\| d\xi\right\} \leqq \left(\int_{t_0}^{t} \|\mathbf{B}(\xi)\| d\xi\right) \exp\left\{\int_{t_0}^{t} \|\mathbf{B}(\xi)\| d\xi\right\} \tag{19}$$

Thus by Theorem 2 of [5] $\mathbf{T}(\mathbf{U})$ has only one fixed point which satisfies the initial value problem by (12):

$$\dot{\mathbf{X}} = (\mathbf{A}+\mathbf{B})\mathbf{X}, \quad \mathbf{X}(t_0) = \mathbf{Y}(t_0)$$

and by theorem 3 of the same paper:

$$\varrho(\mathbf{X}, \mathbf{Y}) = \varrho(\mathbf{U}_\infty, \mathbf{U}) = \|\mathbf{Y}^{-1}(t)\{\mathbf{X}(t) - \mathbf{Y}(t)\}\| \leqq \sigma_\infty - \sigma_0. \tag{20}$$

(19) and (20) implies, that statement (7) of our theorem is really fulfilled.

Now if $\int_{t_0}^{t} \|\mathbf{B}(\xi)\| \, d\xi < \infty$ is also valid then — as we can see from (7) — for a sufficiently large $t_0 \lim_{t\to\infty} \|\mathbf{Y}^{-1}\mathbf{X} - \mathbf{E}\|$ is arbitrarily small; so if for $t_1 < t_0$ we do not start from the fundamental matrix corresponding to $\mathbf{X}(t_1) = \mathbf{Y}(t_1)$ but from an $\mathbf{X}^{(0)}(t)$ that — for this sufficiently large t_0 — fulfils the equation $\mathbf{X}^{(0)}(t_0) = \mathbf{Y}(t_0)$, then for this $\mathbf{X}^{(0)}(t)$ $\lim_{t\to\infty} \|\mathbf{Y}^{-1}\mathbf{X}^{(0)} - \mathbf{E}\|$ is arbitrarily small. Consequently for a suitable fundamental matrix $\mathbf{X}^{(0)}$ we can guarantee also the relation $\lim_{t\to\infty} \|\mathbf{Y}^{-1}\mathbf{X}^{(0)} - \mathbf{E}\| = 0$. The product of matrices $\mathbf{X}^0$ and $\mathbf{Y}^{-1}$ — because of the convergence of the improprius integral — satisfies the relation:

$$\begin{aligned} \mathbf{Y}^{-1}(t)\mathbf{X}^{(0)}(t) &= \mathbf{E} - \int_{t}^{\infty} \mathbf{Y}^{-1}(\xi)\mathbf{B}(\xi)\mathbf{X}^{(0)}(\xi)\, d\xi = \\ &= \mathbf{E} - \int_{t}^{\infty} \{\mathbf{Y}^{-1}(\xi)\mathbf{B}(\xi)\mathbf{Y}(\xi)\}\{\mathbf{Y}^{-1}(\xi)\mathbf{X}^{(0)}(\xi)\} d\xi \end{aligned} \tag{21}$$

and thus

$$\|\mathbf{Y}^{-1}(t)\mathbf{X}^{(0)}(t) - \mathbf{E}\| \leqq \int_{t}^{\infty} \|\mathbf{B}(\xi)\| \|\mathbf{Y}^{-1}(\xi)\mathbf{X}^{(0)}(\xi)\| d\xi < 2\|\mathbf{E}\| \int_{t}^{\infty} \|\mathbf{B}(\xi)\| d\xi \tag{22}$$

for a sufficiently large t. For in this case $\lim_{t\to\infty} \|\mathbf{Y}^{-1}\mathbf{X}^{(0)} - \mathbf{E}\| = 0$, and for a sufficiently large t $\|\mathbf{Y}^{-1}\mathbf{X}^{(0)}\| < 2\|\mathbf{E}\|$. By this we have proved statement (9) of our theorem, too.

Introducing the notations $\mathbf{A}+\mathbf{B}=\mathbf{C}$, $\mathbf{A}=\mathbf{C}-\mathbf{B}$ we get the two remaining relations immediately.

THEOREM 2. *Under the assumptions of the former theorem, let us consider the differential equation*

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\boldsymbol{\varphi}(\mathbf{x}) \tag{23}$$

where $\boldsymbol{\varphi}(\mathbf{x})$ *is a function continuous in* $\mathbf{x}$ *and satisfies for* $\|\mathbf{x}\|>0$ *a Lipschitz-condition and* $\boldsymbol{\varphi}(\mathbf{o})=0$. *Further let* $\boldsymbol{\varphi}(\mathbf{x})$ *satisfy the condition*

$$\omega(t) \geqq \max_{\|\mathbf{x}\| \leqq t} \|\boldsymbol{\varphi}(\mathbf{x})\| \tag{24}$$

where $\omega(t)$ *is a strictly monotone, continuous and concave function for which* $\lim_{\varepsilon \to 0} \int_{\varepsilon}^{1} \frac{1}{\omega(\xi)} d\xi$ *is infinite.*

In this case for the fundamental matrix $\mathbf{X}(t)$ *in* (23) *satisfying the condition* $\mathbf{X}(t_0)=\mathbf{Y}(t_0)$ *holds the relation*

$$\|\mathbf{Y}^{-1}(t)\mathbf{X}(t)-\mathbf{E}\| \leqq \Omega^{-1}\left\{\Omega\left(\int_{t_0}^{t}\|\mathbf{B}(\xi)\|\,d\xi\right)+\int_{t_0}^{t}\|\mathbf{B}(\xi)\|\,d\xi\right\} \tag{25}$$

where

$$\Omega(u)=\int_{a}^{u}\frac{d\xi}{\omega(\xi)}$$

and $\Omega^{-1}(v)$ *is the inverse of* $\Omega(u)$.

If $\int_{t_0}^{\infty}\|\mathbf{B}(\xi)\|\,d\xi<\infty$, *then there exists a fundamental matrix* $\mathbf{X}^{(0)}(t)$ *for which*

$$\|\mathbf{Y}^{-1}(t)\,\mathbf{X}^{(0)}(t)-\mathbf{E}\| \to 0 \quad \textit{provided } t \to \infty. \tag{27}$$

PROOF. The proof of this theorem is based upon the idea of the former theorem but we apply theorems 1**, 2**, 3** of [5] and BIHARI's generalization of the BELLMAN Lemma (see [6], [7]) respectively.

Applying the results just obtained we shall sharpen CESARI's asymptotic theorem and its generalization by BIHARI. Let us consider the differential equation

$$\dot{\mathbf{z}}=(\mathbf{A}+\mathbf{V}(t))\mathbf{z}+\mathbf{B}(t)\boldsymbol{\varphi}(\mathbf{z}) \tag{28}$$

where $\mathbf{V}(t)\to\mathbf{O}$ if $t\to\infty$, further $\operatorname{Var}_{(t_0\infty)}(\|\mathbf{V}(t)\|)<\infty$ and $\int_{t_0}^{\infty}\|\mathbf{B}(\xi)\|d\xi<\infty$ are fulfilled.

Let us denote the eigen-values of matrix $\mathbf{A}$ by λ_k and the corresponding eigen-values of matrix $\mathbf{A}+\mathbf{V}(t)$ by $\lambda_k(t)$ $(k=1, 2, \ldots n$; the eigen-values are not necessarily simple).

THEOREM 3. *Under the given conditions for every* k *we can find a solution* $\mathbf{z}_k$ *of* (28) *satisfying the relation*

$$\lim_{t\to\infty}\mathbf{z}_k(t)e^{-\int_{t_0}^{t}\lambda_k(\xi)d\xi}=\mathbf{s}_k \tag{29}$$

where $\mathbf{A}\mathbf{s}_k=\lambda_k\mathbf{s}_k$ *i.e.* $\mathbf{s}_k$ *is the eigen-vector of* $\mathbf{A}$ *belonging to* λ_k. *If* λ_k *is a multiple eigen-value then denote* $\mathbf{s}_k^{(2)}\mathbf{s}_k^{(3)}$ *... the corresponding main-vectors of second, third ... order.*

In this case for every i in question we can find a solution $\mathbf{z}_k^{(i)}$ so that

$$\lim_{t\to\infty}\frac{1}{t^{i-1}}\left\{\mathbf{z}_k^{(i)}\exp\left(-\int_{t_0}^{t}\lambda_k^{(i)}(\xi)d\xi\right)-\mathbf{s}_k^{(i)}\right\}=\mathbf{o} \tag{30}$$

is valid.

PROOF. We may follow the ideas in the proof of the Cesari theorem in [4], and of its generalization in [7]. The only essential difference is that we omit the assumptions concerning the simplicity of eigen-values of $\mathbf{A}$ (cf. Theorem 1) depending on t, that is why we have to show — in a way different from that in [4] — that the matrix $\mathbf{S}(t)$ in the transformation of $\mathbf{A}+\mathbf{V}(t)$ has bounded variation.

It follows from the assumption $\mathbf{V}(t)\to\mathbf{O}$ that for a sufficiently large t we may assign — in a one to one way — to each simple eigen-value of $\mathbf{A}$ a unique simple eigen-value of $\mathbf{A}+\mathbf{V}(t)$ that for $t\to\infty$ converge to the corresponding eigen-value of $\mathbf{A}$. We shall assign, however, to a multiple eigen-value of $\mathbf{A}$ only one eigen-vector and a suitable number of main-vectors even if the eigen-value in question is a simple root of the minimal-polynomial of $\mathbf{A}$, because in the minimal-polynomial of $\mathbf{A}+\mathbf{V}(t)$ the multiplicity of the corresponding eigen-value depends on t, moreover, the corresponding group of eigen-values can split and fuse resp. depending on t. For this reason we correspond the submatrix of transformation $\mathbf{T}^{(j)}$ to the l-multiple eigen-value λ_j of $\mathbf{A}$ so that the structure of the l-dimensional block $\mathbf{\Lambda}^{(j)}$ that corresponds to the matrix $\mathbf{T}^{(j)}\mathbf{A}\mathbf{T}^{(j)-1}$ be of the form $\begin{bmatrix}\lambda_j & 0 & 0 & 0 & \dots & 0\\ 1 & \lambda_j & 0 & 0 & \dots & 0\\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots\\ 0 & 0 & 0 & 0 & 1 & \lambda_j\end{bmatrix}$ independently of the multiplicity of λ_j in the minimal polynomial of $\mathbf{A}$.

We construct for a sufficiently large fixed t_0 the transform of $\mathbf{A}+\mathbf{V}(t_0)$ in the same way trough $\mathbf{\Lambda}^{(j)}(t)$ — the l-dimensional hyperblock of the corresponding $\mathbf{S}^{(j)}(t)[\mathbf{A}+\mathbf{V}(t)]\mathbf{S}^{(j)-1}(t)$ — whose structure takes the form

$$\begin{bmatrix}\lambda_j^{(1)}(t) & 0 & 0 & \dots & \\ 1 & \lambda_j^{(2)}(t) & 0 & \dots & \\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & 1 & \lambda_j^{(l)}(t)\end{bmatrix}$$

— may contain but different eigen-values. After all, by choosing a suitable $\mathbf{S}^{(j)}(t)$ we may guarantee that in all points of continuity of $\mathbf{V}(t)$ both $\mathbf{S}(t)^{(j)}$ and $\lambda_j^{(\nu)}(t)$ were continuous and because of $\mathbf{V}(t)\to\mathbf{O}$ both $\mathbf{S}^{(j)}(t)\to\mathbf{T}^{(j)}$ and $\lambda_j^{(\nu)}(t)\to\lambda_j$ were fulfilled. After the transformation introducing the variable $\mathbf{x}(t)=\mathbf{S}^{-1}(t)\mathbf{z}$ — our equation takes the form

$$\dot{\mathbf{x}}=\mathbf{\Lambda}(t)\mathbf{x}+\frac{d\mathbf{S}}{dt}\mathbf{S}^{-1}+\mathbf{S}\mathbf{B}\mathbf{S}^{-1}\boldsymbol{\varphi}(\mathbf{S}(t)\mathbf{x}) \tag{31}$$

In the following, first of all, we must deal with the question of how to define in a unique way $\mathbf{S}(t)$ in the points of discontinuity of $\mathbf{V}(t)$ and how to estimate $\left\|\frac{d\mathbf{S}}{dt}\right\|$ and $\operatorname{Var}(\|\mathbf{S}(t)\|)$ by $\left\|\frac{d\mathbf{V}}{dt}\right\|$ and $\operatorname{Var}(\|\mathbf{V}(t)\|)$, resp.

These questions may be answered on the basis of perturbation Theorems in [8]. Namely $\mathbf{V}(t) \to \mathbf{O}$ implies that we may find such a t_0 that we can apply theorem 1 and its generalization in [8] concerning the multiple eigen-values i.e. for $t_1 > t_0$, and $t_2 > t_0$ the eigen-values, the eigen-vectors and the main-vectors of $\mathbf{A}+\mathbf{V}(t_2)$ can be obtained both from $\mathbf{A}$ and $\mathbf{A}+\mathbf{V}(t_1)$ by means of perturbation. So the elements of $\mathbf{S}(t)$ — on the basis of the perturbation theorems just mentioned — may be extended to the points of discontinuity of $\mathbf{V}(t)$ as well. Besides, by theorems we mentioned the inequalities

$$\operatorname{Var}\left(\|\mathbf{S}(t)\|\right) \leqq \frac{1}{1-q} \operatorname{Var}\left(\|\mathbf{V}(t)\|\right) \tag{32}$$

and

$$\operatorname{Var}\left(|\lambda_i^{(j)}(t)|\right) \leqq \frac{1}{1-q_1} \operatorname{Var}\left(\|\mathbf{V}(t)\|\right) \tag{33}$$

are also fulfilled for every $t > t_0$, where — for a sufficiently large t_0 —

$$0 < q = q(t_0) = q\left(\sup_{t \geqq t_0} \|\mathbf{V}(t)\|\right) < 1 \tag{34}$$

and

$$0 < q_1 = q(t_0) = q_1\left(\sup_{t \geqq t_0} \|\mathbf{V}(t)\|\right) < 1$$

are valid.

Since for a sufficiently large t_0 $\|\mathbf{S}(t)\| \leqq 2\|\mathbf{T}\|$ and $\|\mathbf{S}^{-1}(t)\| \leqq 2\|\mathbf{T}^{-1}\|$ are also valid, the coefficient of $\mathbf{x}$ in the second term of the right side of (31) satisfies the assumptions of Theorem 1.

In a similar way the coefficient of φ in the third term of the right side of (31) also satisfies the assumption of Theorem 2. Theorem 2 can be applied in spite of having $\mathbf{S}(t)\mathbf{x}$ instead of $\mathbf{x}$ in the argument of φ — as it was verified by BIHARI in [7]. Consequently we may apply Theorems 1 and 2. Thus the relation (29) connected with the simple eigen-values of $\mathbf{A}$ is immediately obtained. Concerning the relation (30) we must consider that in the suitable solution of the equation

$$\dot{\mathbf{y}} = \mathbf{\Lambda}(t)\mathbf{y}$$

(we choose with regard to the application of Theorem 1) the linear combination of the terms of the form $\exp\left\{\int_{t_0}^{t} \lambda_k^{(r)}(\xi)\, d\xi\right\}$ $(r = 1, 2, \ldots i)$ or the products of such terms and powers of t not higher than $i-1$ figure according to the validity of

$$\int_{t_0}^{t} \lambda_k^{(v_1)}(\xi)\, d\xi \not\equiv \int_{t_0}^{t} \lambda_k^{(v_2)}(\xi)\, d\xi. \tag{33}$$

implies that the quotient of terms of the form $\exp\left\{\int_{t_0}^{t} \lambda_k^{(v)} d\xi\right\}$ may be restricted between two positive bounds for every $t > t_0$, and so the expression multiplied by $1/t^{i-1}$ in every possible case tends to zero. By this we have proved the Theorem.

Finally we should remark that relation (30), connected with the multiple eigenvalues, characterizes much less exactly the asymptotic behaviour of $\mathbf{z}_k^{(i)}$ than (29) does in case of $\mathbf{z}_k$.

By (30) the „order" behaviour of $\mathbf{z}_k^{(i)}$ remains, however, doubtful only if we are confronted with the case $\operatorname{Re}\{\lambda_k^{(i)}(t)\} \to 0$.

§ 3. Asymptotic Investigation of Differential Equations with Piecewise Constant Recurrent Coefficients

Let us consider the linear differential equation

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} \tag{35}$$

and assume that $\mathbf{A}(t)$ is equal to matrices $\mathbf{B}$ and $\mathbf{C}$ on successive intervals (t_i, t_{i+1}) $(i=0, 1, 2, \ldots)$ alternatively. It is well-known that in the case of (general) integrable matrix $\mathbf{A}(t)$ the asymptotic stability of the solution $\mathbf{x} \equiv \mathbf{o}$ of (35) is not guaranteed by the requirement that each eigen-value of $\mathbf{A}(t)$ for every t has a negative real part. (A simple counterexample is the system $\dot{x}_1 = -x_1, \dot{x}_2 = e^{2t}x_1 - 2x_2$).

We shall, however, prove

THEOREM 4. *If each eigen-value of* $\mathbf{B}$ *and* $\mathbf{C}$ *has a negative real part* $(\mathbf{A}(t) = \mathbf{B}$ *for* $t \in (t_{2i}, t_{2i+1})$ *and* $\mathbf{A}(t) = \mathbf{C}$ *for* $t \in (t_{2i+1}, t_{2i+2})$ $(i=0, 1, 2, \ldots))$ *then the solution* $\mathbf{x} \equiv \mathbf{o}$ *of* (35) *is asymptotically stable.*

PROOF. Let us assume that in the minimal polynomials of $\mathbf{B}$ and $\mathbf{C}$ every eigen-value is simple. Let us denote the eigen-values of $\mathbf{B}$ by β_j $(j=1, 2, \ldots n)$ and the right side eigen-vectors belonging to them by $\mathbf{b}_j$, the eigen-values and eigen-vectors of $\mathbf{C}$ by γ_j and $\mathbf{c}_j$ $(j=1, 2, \ldots n)$. On the interval (t_{2i}, t_{2i+1}) let us decompose $\mathbf{x}(t)$ to components parallel with $\mathbf{b}_j$ and on the interval (t_{2i+1}, t_{2i+2}) to components parallel with $\mathbf{c}_j$

$$\mathbf{x} = \sum_{j=1}^{n} \xi_j \mathbf{b}_j \quad \text{for} \quad t \in (t_{2i}, t_{2i+1})$$

(36)

and

$$\mathbf{x} = \sum_{j=1}^{n} \eta_j \mathbf{c}_j \quad \text{for} \quad t \in (t_{2i+1}, t_{2i+2})$$

The solution must be continuous on the boundary of intervals, i.e. for every k the relation

$$\mathbf{x}(t_k - 0) = \mathbf{x}(t_k + 0); \quad \sum_{j=1}^{n} \xi_j(t_k)\mathbf{b}_j = \sum_{j=1}^{n} \eta_j(t_k)\mathbf{c}_j \qquad (k = 0, 1, 2, \ldots) \tag{37}$$

must be valid. Since the system of vectors $\mathbf{b}_j$ and $\mathbf{c}_j$ is complete, the vectors $\xi_j(t_k)$ are linear combinations of $\eta_j(t_k)$ and vice versa

$$\boldsymbol{\xi}(t_k) = \mathbf{T}\boldsymbol{\eta}(t_k) \tag{38}$$

$$\boldsymbol{\eta}(t_k) = \mathbf{T}^{-1}\boldsymbol{\xi}(t_k)$$

(by means of a non-singular transformation-matrix $\mathbf{T}$).

As the vectors $\mathbf{b}_j$ and $\mathbf{c}_j$ are the eigen-vectors of $\mathbf{B}$ and $\mathbf{C}$ therefore on the corresponding intervals ξ and η satisfy the equations

$$\dot{\xi} = \Lambda(\mathbf{B})\xi \quad \text{and} \quad \dot{\eta} = \Lambda(\mathbf{C})\eta \tag{39}$$

where $\Lambda(\mathbf{B}) = \langle \beta_1, \ldots, \beta_n \rangle$; $\Lambda(\mathbf{C}) = \langle \gamma_1, \ldots, \gamma_n \rangle$.

Thus if $\mathbf{x}(t_0+0) = \xi_{10}\mathbf{b}_1 + , \ldots, \xi_{n0}\mathbf{b}_n$ i.e. if $\mathbf{x}(t_0)$ is determined by the vector $\xi_0 = \xi(t_0)$, then for $t \in (t_0+0, t_1-0)$

$$\xi(t) = \exp\{(t-t_0)\Lambda(\mathbf{B})\}\xi_0 \tag{40}$$

Besides (38) implies

$$\eta(t_1+0) = \mathbf{T}^{-1}\xi(t_1-0) = \mathbf{T}^{-1}\exp\{(t_1-t_0)\Lambda(\mathbf{B})\}\xi_0 \tag{41}$$

Consequently for $t \in (t_1+0, t_2-0)$

$$\eta(t) = \exp\{(t-t_1)\Lambda(\mathbf{C})\}\exp\{(t_1-t_0)\Lambda(\mathbf{B})\}\xi_0 \tag{42}$$

Generally also for $t \in (t_{2i}+0, t_{2i+1}-0)$

$$\begin{aligned}\xi(t) = \exp\{(t-t_{2i})\Lambda(\mathbf{B})\}\mathbf{T}\exp\{(t_{2i}-t_{2i-1})\Lambda(\mathbf{C})\}\cdot \\ \cdot \mathbf{T}^{-1}\exp\{(t_{2i-1}-t_{2i-2})\Lambda(\mathbf{B})\}\mathbf{T}\ldots\mathbf{T}^{-1}\exp\{(t_1-t_0)\Lambda(\mathbf{B})\}\xi_0\end{aligned} \tag{43}$$

similarly for $t \in (t_{2i+1}+0, t_{2i+2}-0)$

$$\eta(t) = \exp\{(t-t_{2i+1})\Lambda(\mathbf{C})\}\mathbf{T}^{-1}\ldots\mathbf{T}^{-1}\exp\{(t_1-t_0)\Lambda(\mathbf{B})\}\xi_0 \tag{44}$$

Let us denote the square-root of the non-singular matrix $\mathbf{T}$ by $\mathbf{T}^{\frac{1}{2}}$ and its inverse by $\mathbf{T}^{-\frac{1}{2}}$.

We introduce the notations

$$\mathbf{E}_{2i}(\mathbf{B}) = \mathbf{T}^{-\frac{1}{2}}\exp\{(t_{2i+1}-t_{2i})\Lambda(\mathbf{B}\}\mathbf{T}^{\frac{1}{2}} \tag{45}$$

and

$$\mathbf{E}_{2i}(\mathbf{C}) = \mathbf{T}^{\frac{1}{2}}\exp\{(t_{2i}-t_{2i-1})\Lambda(\mathbf{C})\}\mathbf{T}^{-\frac{1}{2}} \tag{46}$$

Together with them (43) and (44) take the form

$$\xi(t) = \exp\{(t-t_{2i})\Lambda(\mathbf{B})\}\mathbf{T}^{\frac{1}{2}}\mathbf{E}_{2i}(\mathbf{C})\mathbf{E}_{2i-2}(\mathbf{B})\mathbf{E}_{2i-2}(\mathbf{C})\ldots\mathbf{E}_0(\mathbf{B})\mathbf{T}^{-\frac{1}{2}}\xi_0 \tag{47}$$

and

$$\eta(t) = \exp\{(t-t_{2i+1})\Lambda(\mathbf{C})\}\mathbf{T}^{-\frac{1}{2}}\mathbf{E}_{2i}(\mathbf{B})\mathbf{E}_{2i}(\mathbf{C})\ldots\mathbf{E}_0(\mathbf{B})\mathbf{T}^{-\frac{1}{2}}\xi_0 \tag{48}$$

Let us consider now a matrix-norm of the same property as in § 2 and a vector-norm compatible with it.

Then

$$\|\mathbf{E}_{2i}(\mathbf{B})\| = \|\exp\{(t_{2i+1}-t_{2i})\Lambda(\mathbf{B})\}\| \tag{49}$$

and

$$\|\mathbf{E}_{2i}(\mathbf{C})\| = \|\exp\{(t_{2i}-t_{2i-1})\Lambda(\mathbf{C})\|\} \tag{50}$$

for the matrix-norm is invariant with respect to the similarity transformation. Thus

$$\|\xi(t)\| \leqq \|\mathbf{T}^{\frac{1}{2}}\|\mathbf{T}^{-\frac{1}{2}}\|\exp\{t-t_{2i})\mathbf{\Lambda}(\mathbf{B})\}\|\,\|\xi_0\| \prod_{\varrho=1}^{i} \|\mathbf{E}_{2\varrho}(\mathbf{B})\|\,\|\mathbf{E}_{2\varrho}(\mathbf{C})\| =$$

(51)

$$= \|\mathbf{T}^{-\frac{1}{2}}\|\,\|\mathbf{T}^{\frac{1}{2}}\|\,\|\xi_0\|\,\|e^{(t-t_{2i})\mathbf{\Lambda}(\mathbf{B})}\| \prod_{\varrho=1}^{i} \|\exp\{(t_{2\varrho+1}-t_{2\varrho})\mathbf{\Lambda}(\mathbf{B})\}\|\,\|\exp\{(t_{2\varrho}-t_{2\varrho+1})\mathbf{\Lambda}(\mathbf{C}\}\|$$

Similarly

$$\|\eta(t)\| \leqq \|\mathbf{T}^{-\frac{1}{2}}\|^2\,\|\xi_0\|\,\|\exp\{(t-t_{2\tau+1})\mathbf{\Lambda}(\mathbf{C})\| \prod_{\varrho=1}^{i} \|\exp\{(t_{2\varrho}-t_{2\varrho-1})\mathbf{\Lambda}(\mathbf{C})\| \cdot$$

(52)

$$\cdot \|\exp\{(t_{2\varrho-1}-t_{2\varrho-2})\mathbf{\Lambda}(\mathbf{B})\|$$

and our statement easily follows from (51) and (52).

If the eigen-values of $\mathbf{B}$ and $\mathbf{C}$ are multiple roots of the minimal polynomial then in our considerations we must use the corresponding main-vectors instead of the eigen-vectors and in this case $\mathbf{\Lambda}(\mathbf{B})$ and $\mathbf{\Lambda}(\mathbf{C})$ are not diagonal but Jordan-matrices.

However, if each eigen-value of $\mathbf{B}$ and $\mathbf{C}$ has a negative real-part then the statement of our theorem follows from the products of the expressions in (49) and (50).

By this we have proved our theorem.

It is easy to see that our theorem and even the order of ideas in its proof can be carried over to the more general case when $\mathbf{A}(t)$ is equal to $\mathbf{B}_1, \mathbf{B}_2, \ldots \mathbf{B}_k$ on a certain successive sequence of intervals with *nonnegativ* length. In this case we must change our proof in the following way: Coming over from $\mathbf{B}_2$ to $\mathbf{B}_3$ the transformation-matrix will be $\mathbf{T}_2 \cdot \mathbf{T}_1^{-1}$; coming over from $\mathbf{B}_3$ to $\mathbf{B}_4$ it will be $\mathbf{T}_3 \cdot \mathbf{T}_2^{-1}$ etc. and finally coming over from $\mathbf{B}_k$ to $\mathbf{B}_1$ it will be $\mathbf{T}_1\mathbf{T}_{k-1}^{-1}$ ($\mathbf{T}_1$ denotes the transformation-matrix we use when coming over from $\mathbf{B}_1$ to $\mathbf{B}_2$)

Thus the theorem is valid:

THEOREM 5. *If the matrix* $\mathbf{A}(t)$ *of the differential equation* (35) *is equal to constant matrices* $\mathbf{B}_1, \mathbf{B}_2 \ldots \mathbf{B}_k, \mathbf{B}_1, \mathbf{B}_2, \ldots$ *on intervals* (t_1, t_2) $(t_2, t_3) \ldots$ *resp. and each eigen-value of any* $\mathbf{B}_i$ *has a negative real-part then the solution* $\mathbf{x} \equiv \mathbf{o}$ *of* (35) *is asymptotically stable* $(t_1 \leqq t_2;\ t_2 \leqq t_3;\ \ldots)$.

Finally we mention that — as it follows from (51) and (52) — we can guarantee the Ljapunoff-stability and in the case of a suitable combination even the asymptotic stability of the solution $\mathbf{x} \equiv \mathbf{o}$, when the real-part of some eigen-values is zero and also in the case when — with suitable restrictions concerning the subintervals — some eigen-values have positive real parts. E.g. if the equation (35) has a periodic and Riemann-integrable coefficient matrix $\mathbf{A}(t)$ — approximating $\mathbf{A}(t)$ by a sequence of piecewise constant matrices — we can immediately prove the

THEOREM 6. *Let the Riemann-integrable matrix* $\mathbf{A}(t)$ *of* (35) *have period T. Denote* $\lambda_j(t)$ *ist eigen-values depending on t. When the relation*

$$\int_{t_0}^{t_0+T} \sup_{j(\xi)} \operatorname{Re} \lambda_j(\xi)\, d\xi < 0 \tag{53}$$

is valid then any characteristic exponent belonging to (35) *has a negative real-parts i.e. the solution* $\mathbf{x} \equiv \mathbf{o}$ *is asymptotically stable.*

We also mention — that we can see from the proof of Theorem 4 — that $\int_{t_0}^{t_0+T} \sup_{j(\xi)} \operatorname{Re} \lambda_j(\xi) d\xi$ is an upper bound of the real-parts of the corresponding characteristic exponents.

§ 4. Proof of Aizerman's Conjecture

Let us consider the equation

$$\dot{\mathbf{x}} = [\mathbf{A} + \mathbf{B}(\mathbf{x}) + \mathbf{C}(\mathbf{x})]\mathbf{x} + \mathbf{D}(\mathbf{x}) \cdot \boldsymbol{\varphi}(\mathbf{x}) \tag{54}$$

where $\mathbf{A}$ is a constant matrix and $\mathbf{B}(\mathbf{x})$ is also constant while expressions formed by components of $\mathbf{x}$ keep their sign and it changes discontinuously with the sign-changings, so that the product $\mathbf{B}(\mathbf{x}) \cdot \mathbf{x}$ satisfies a Lipschitz-condition. $\mathbf{C}(\mathbf{x})$ is a matrix satisfying the condition

$$\int_0^t \|\mathbf{C}(\mathbf{x}(\tau))\| d\tau \leqq t^{\nu(\varepsilon)} \tag{55}$$

($\nu < 1$, if $\varepsilon > 0$) from some fixed t_s (e.g. from $t_s = 0$), when $\mathbf{x}(t) \in \mathbf{C}$, $\|\mathbf{x}(t)\| \leqq \exp\{-\varepsilon t\}$, $\varepsilon > 0$. A similar condition refers to $\mathbf{D}(\mathbf{x})$, namely

$$\int_0^\infty \|\mathbf{D}(\mathbf{x}(\tau))\| d\tau < \infty \tag{56}$$

when $\mathbf{x}(t) \in \mathbf{C}$, $\|\mathbf{x}(t)\| \leqq \exp\{-\varepsilon t\}$.

At last we assume that the expression $\mathbf{C}(\mathbf{x}) \cdot \mathbf{x} + \mathbf{D}(\mathbf{x}) \cdot \boldsymbol{\varphi}(\mathbf{x})$ satisfies the condition of the type

$$\|\mathbf{C}(\mathbf{x}_2)\mathbf{x}_2 + \mathbf{D}(\mathbf{x}_2)\boldsymbol{\varphi}(\mathbf{x}_2) - \mathbf{C}(\mathbf{x}_1)\mathbf{x}_1 - \mathbf{D}(\mathbf{x}_1)\boldsymbol{\varphi}(\mathbf{x}_1)\| \leqq K\psi(\|\mathbf{x}_2 - \mathbf{x}_1\|), \tag{57}$$

where ψ is a continuous concave monotone increasing function, for which $\psi(0) = 0$, and

$$\lim_{\varepsilon \to 0} \int_\varepsilon^1 \frac{d\xi}{\psi(\xi)} = \infty \tag{58}$$

Further let $\boldsymbol{\varphi}(\mathbf{x})$ also satisfy a condition of the same type.

Besides we assume that $\mathbf{C}$ and $\mathbf{D}$ are bounded in $\mathbf{x}$.

It is obvious that (54) — under the conditions (55)—(58) — includes the problems (1)—(3) as a special case if we assume in addition that for arbitrary $\mathbf{x}(t) \in C$, $\|\mathbf{x}(t)\| \leqq K_1$ the eigen-values λ_ν ($\nu = 1, \ldots, m$) of the piecewise constant matrices $\mathbf{A}$ and $\mathbf{F}(t) = \mathbf{A} + \mathbf{B}(\mathbf{x}(t))$ satisfy for any t the following condition:

$$\begin{aligned} \operatorname{Re}\{\lambda_\nu(\mathbf{A})\} &\leqq -\varepsilon_0 \\ \operatorname{Re}\{\lambda_\nu(\mathbf{F})\} &\leqq -\varepsilon_0 < 0. \end{aligned} \tag{59}$$

We can see that (54) — under the conditions (55)—(59) — is more general than (1)—(3). Every control system containing switching elements and elements of non-linear characteristics which can be well approximated by linear pieces, can be described in the above way if the effects of the switching elements is „sufficiently continuous" and every linearized piece is equivalent to an asymptotically stable and „absolutely linear" system. Under these conditions holds the following

THEOREM 7. *Let the conditions* (55)—(59) *be satisfied. Then the solution* $\mathbf{x} \equiv \mathbf{0}$ *of the equation* (54) *is asymptotically stable if* $\mathbf{C}$ *and* $\mathbf{D}$ *are bounded in* $\mathbf{x}$.

PROOF. First of all we remark that we can prove — as a special case of a theorem of section 3 of [5] — that the successive approximations used for the equation (54) converge to the unique fixed point of the equation (54). But we are going to use successive approximations of another type. Let the starting elements (with the initial condition $\mathbf{x}(0) = \mathbf{x}_0$) be the solution of the following initial-value problem:

$$\dot{\mathbf{z}} = \mathbf{A}\mathbf{z} \quad \mathbf{z}(0) = \mathbf{x}(0). \tag{60}$$

We construct the next approximation $\mathbf{x}_1$ from $\mathbf{x}_0 = \mathbf{z}(t)$ (resp. $\mathbf{x}_{n+1}$ from $\mathbf{x}_n$) in the following way:

$$\mathbf{x}_{n+1}(t) = \mathbf{z}(t) + \int_0^t \mathbf{Y}(t-\tau)\{[\mathbf{B}(\mathbf{x}_n(\tau)) + \mathbf{C}(\mathbf{x}_n(\tau))] \cdot \mathbf{x}_{n+1}(\tau) + \mathbf{D}(\mathbf{x}_n(\tau))\boldsymbol{\varphi}(\mathbf{x}_n(\tau))\}\, d\tau \tag{61}$$

where $\mathbf{Y}(t)$ is the fundamental matrix for the equation (60) satisfying the initial condition $\mathbf{Y}(0) = \mathbf{E}$, i.e. while knowing $\mathbf{x}_n(t)$, $\mathbf{x}_{n+1}(t)$ is the solution of the equation

$$\dot{\mathbf{x}}_{n+1} = \{\mathbf{A} + \mathbf{B}(\mathbf{x}_n) + \mathbf{C}(\mathbf{x}_n)\}\mathbf{x}_{n+1} + \mathbf{D}(\mathbf{x}_n)\boldsymbol{\varphi}(\mathbf{x}_n) \tag{62}$$

under the corresponding initial condition.

To prove our assertion it will be first verified that on every finite interval the sequence $\mathbf{x}_n(t)$ uniformly converges to a uniquely determined solution of (54). Then we shall prove that $\|\mathbf{x}_n(t)\|$ can be estimated by $C \cdot \exp\left\{-\frac{n+1}{2n+1}\,\varepsilon_0 t\right\}$ from a bound T sufficiently large and independent of n, hence their limit satisfies the inequality

$$\|\mathbf{x}(t)\| \leqq C \exp\{-\tfrac{1}{2}\,\varepsilon_0 t\}.$$

At last, considering this inequality we can prove that $\mathbf{x}(t)$ satisfies the inequality

$$\|\mathbf{x}(t)\| \leqq C_1 \exp\{-\varepsilon_0^* t\},$$

for arbitrary $0 < \varepsilon_0^* < \varepsilon_0$, which completely proves our theorem.

We start our proof verifying the second assertion in an inductive way. It is obviously true for $\mathbf{x}_0(t) = \mathbf{z}(t)$. Let us assume that it is true also for every $n \leqq N$. On the basis of (61) we can estimate $\|\mathbf{x}_{N+1}(t)\|$ as follows: $\mathbf{x}_{N+1}(t)$ is the solution of an inhomogeneous linear differential equation, which consists of a corresponding solution of the homogeneous equation and the particular solution of the inhomogeneous equation. The function $\mathbf{x}_{N+1}^{(h)}$ of the homogeneous part of the solution which satisfies the initial condition, too, can be estimated — considering Theorems

1., 2., 6., and the boundedness of $\mathbf{C}$ — in the following way:

$$\|\mathbf{x}_{N+1}^{(h)}(t)\| \leqq C \exp\left\{t^{\nu}\left(\frac{N+1}{2N+1} E_0\right)\right\} \exp(-E_0 t) \tag{63}$$

where C is a constant depending on $\mathbf{x}_0$, the upper bound of $\mathbf{C}$ and the transformations occurring in $\mathbf{B}$. $\nu(\varepsilon)$ is a decreasing function of ε. Let $\nu_0 = \nu(\frac{1}{2}\varepsilon_0)$ and let T_1 be the bound from which the inequality

$$\exp(t^{\nu_0}) \exp(-\varepsilon_0 t) \leqq \tfrac{1}{2} \exp(-\tfrac{3}{4}\varepsilon_0 t)$$

is valid.

Then a fortiori $\|\mathbf{x}_{N+1}^{(h)}(t)\| \leqq \frac{1}{2} C \exp\left\{-\frac{N+2}{2N+3}\varepsilon_0 t\right\}$

for $t \geqq T_1$.

We consider the inhomogeneous part in the following form:

$$\mathbf{x}_{N+1}^{(i)}(t) = \int_0^t \mathbf{Y}(t-\tau)\mathbf{D}[\mathbf{x}_N(\tau)]\boldsymbol{\varphi}(\mathbf{x}_N(\tau))d\tau,$$

where $\|\mathbf{Y}(t-\tau)\| \leqq C_2 \exp\{-\varepsilon_0(t-\tau)\}$, further $\|\mathbf{D}[\mathbf{x}_N(\tau)]\| = s_1(\tau)$ has a bounded integral on $(0, \infty)$ because of the hypotheses valid for $\mathbf{x}_N(t)$; finally $\|\boldsymbol{\varphi}(\mathbf{x}_N(t))\| \leqq$ $\leqq \|\mathbf{x}_N(\tau)\| \cdot \log^2 \|\mathbf{x}_N(\tau)\|$ if τ is sufficiently large (e.g. $\tau \geqq T_2$), since in the opposite case $\int_0^1 \frac{1}{\psi(\xi)} d\xi < \infty$ were valid.

Hence for a sufficiently large t (e.g. for $t \geqq T_3$)

$$\left\| \int_0^t \mathbf{Y}(t-\tau)\mathbf{D}[\mathbf{x}_N(\tau)]\boldsymbol{\varphi}(\mathbf{x}_N(\tau))d\tau \right\| \leqq$$

$$\leqq C_4 e^{-E_0 t} \int_0^t \exp\left\{\frac{N+1}{2N+1} E_0 \tau\right\} \frac{s_1(\tau)}{(1+\tau)^2} d\tau \leqq \frac{1}{2} C_5 \exp\left\{-\frac{N+2}{2N+3}\varepsilon_0 t\right\}. \tag{65}$$

Consequently by (64) and (65) for $t \geqq \max\{T_1, T_3\}$ holds the inequality

$$\|\mathbf{x}_{N+1}(t)\| \leqq C_6 \exp\left\{-\frac{N+2}{2N+3}\varepsilon_0 t\right\} \tag{66}$$

which guarantees the possibility of the induction. Thus the second assertion has been proved.

We may see immediately — considering our assertion already proved, the boundedness of $\mathbf{D}$, and the relation (61) — that the functions $\mathbf{x}_n(t)$ in the interval $[0, \infty)$ are uniformly bounded and satisfy uniformly a Lipschitz-condition (a fortiori they are equicontinuous).

On the basis of the Ascoli-lemma there exists a subsequence $\mathbf{x}_{n_k}(t)$ of the sequence $\mathbf{x}_n(t)$ which tends uniformly to the continuous limit-function $\mathbf{x}^{(1)}(t)$. So on account of (61) also the subsequence $\mathbf{x}_{n_k+1}(t)$ converges uniformly to a function

$\mathbf{x}^{(2)}(t)$ satisfying the relation

$$(67)\quad \mathbf{x}^{(2)} = \mathbf{z} + \int_0^t \mathbf{Y}(t-\tau)\{\langle \mathbf{B}[\mathbf{x}^{(1)}(\tau)] + \mathbf{C}[\mathbf{x}^{(1)}(\tau)]\rangle \mathbf{x}^{(2)}(\tau) + \mathbf{D}[\mathbf{x}^{(1)}(\tau)]\boldsymbol{\varphi}(\mathbf{x}^{(1)}(\tau))\}\, d\tau.$$

Consequently if we verify that $\lim_{n\to\infty} \|\mathbf{x}_{n+1}(t) - \mathbf{x}_n(t)\| = 0$, then — considering (67) — we get the proof of our first assertion.

Then we have

$$\mathbf{x}_{n+1}(t) - \mathbf{x}_n(t) = \int_0^t \mathbf{Y}(t-\tau)\{\mathbf{B}[\mathbf{x}_n(\tau)]\mathbf{x}_{n+1}(\tau) - \mathbf{B}[\mathbf{x}_{n-1}(\tau)]\mathbf{x}_n(\tau) +$$

$$+ \mathbf{C}[\mathbf{x}_n(\tau)]\mathbf{x}_{n+1}(\tau) - \mathbf{C}[\mathbf{x}_{n-1}(\tau)]\mathbf{x}_n(\tau) + \mathbf{D}[\mathbf{x}_n(\tau)]\boldsymbol{\varphi}(\mathbf{x}_n(\tau)) - \mathbf{D}[\mathbf{x}_{n-1}(\tau)]\boldsymbol{\varphi}(\mathbf{x}_{n-1}(\tau))\}\, d\tau =$$

$$(68)\quad = \int_0^t \mathbf{Y}(t-\tau)\{\mathbf{B}[\mathbf{x}_n(\tau)]\langle \mathbf{x}_{n+1}(\tau) - \mathbf{x}_n(\tau)\rangle - \mathbf{B}[\mathbf{x}_{n-1}(\tau)]\langle \mathbf{x}_n(\tau) - \mathbf{x}_{n-1}(\tau)\rangle +$$

$$+ \langle \mathbf{B}[\mathbf{x}_n(\tau)]\mathbf{x}_n(\tau) - \mathbf{B}[\mathbf{x}_{n-1}(\tau)]\mathbf{x}_{n-1}(\tau)\rangle + \mathbf{C}[\mathbf{x}_n(\tau)]\langle \mathbf{x}_{n+1}(\tau) - \mathbf{x}_n(\tau)\rangle -$$

$$- \mathbf{C}[\mathbf{x}_{n-1}(\tau)]\langle \mathbf{x}_n(\tau) - \mathbf{x}_{n-1}(\tau)\rangle + \langle \mathbf{C}[\mathbf{x}_n(\tau)]\mathbf{x}_n(\tau) - \mathbf{C}[\mathbf{x}_{n-1}(\tau)]\mathbf{x}_{n-1}(\tau)\rangle +$$

$$\mathbf{D}[\mathbf{x}_n(\tau)]\boldsymbol{\varphi}[\mathbf{x}_n(\tau)] - \mathbf{D}[\mathbf{x}_{n-1}(\tau)]\boldsymbol{\varphi}[\mathbf{x}_{n-1}(\tau)]\}\, d\tau$$

Denoting $\|\mathbf{x}_{n+1}(t) - \mathbf{x}_n(t)\|$ by $\Delta_{n+1}(t)$ and considering the hypotheses concerning the boundedness and the moduli of continuity of $\|\mathbf{B}\|$, $\|\mathbf{C}\|$ and $\|\mathbf{D}\|$ (B, C, D denote the upper bound of $\|\mathbf{B}\|$, $\|\mathbf{C}\|$, $\|\mathbf{D}\|$ resp.) so from (68) we have the inequality

$$\Delta_{n+1}(t) \leqq \int_0^t \exp\{-\varepsilon_0(t-\tau)\}\{B\Delta_{n+1}(\tau) + B\Delta_n(\tau) + L_1\Delta_n(\tau) +$$

$$+ C\Delta_{n+1}(\tau) + C\Delta_n(\tau) + K\psi(\Delta_n(\tau))\}\, d\tau$$

i.e. the inequality

$$(69)\quad \Delta_{n+1}(t)e^{\varepsilon_0 t} \leqq \int_0^t (B+C)e^{\varepsilon_0 t}\Delta_{n+1}(t)\, dt + \int_0^t \{[B+C+L_1]\Delta_n(\tau) + K\psi(\Delta_n(\tau))\}\, d\tau$$

Owing to the GRONWALL-lemma we have the inequality

$$(70)\quad \Delta_{n+1}(t) \leqq \left(\int_0^t \{\langle B+C+L_1\rangle \Delta_n(\tau) K\psi(\Delta_n(\tau)\} \cdot e^{\varepsilon_0\tau}\, d\tau\right) \exp\{B+C-\varepsilon_0\}\Big) t$$

By the monotonity of ψ and positivity of the coefficients figuring in (70) it follows that $\vartheta_n(t) \geqq \Delta_n(t)$ for every n if $\vartheta_1(t) \geqq \Delta_1(t)$ and the sequence $\{\vartheta_n\}$ satisfies the equation

$$(71)\quad \vartheta_{n+1}(t) = \exp(B+C-\varepsilon_0)t \cdot \int_0^t \{\langle B+C+L_1\rangle \vartheta_n(\tau) + K\psi(\vartheta_n(\tau))\} e^{\varepsilon_0\tau}\, d\tau$$

But (71) is the construction of the solution of the equation

$$e^{-(B+C-\varepsilon_0)t}[\dot{\xi} - (B+C-\varepsilon_0)\xi] = \{[B+C+L_1]\xi + K\psi(\xi)\} \exp(\varepsilon_0 t)$$

by means of successive approximations considering it as an inhomogeneous linear differential equation. If we introduce the new variable $\eta = e^{-(B+C-E_0)t}$, our equation turns into the equation

$$\dot{\eta}(t) = (B+C+L_1)e^{(B+C)t} \cdot \eta + Ke^{\varepsilon_0 t} \cdot \psi(e^{(B+C-\varepsilon_0)t} \cdot \eta) \tag{72}$$

When we chose $\eta_1(t) = e^{-(B+C-\varepsilon_0)t} \cdot \vartheta_1(t)$ the solution of (72) by successive approximations gives the sequence $\eta_n(t) = e^{-(B+C-\varepsilon_0)t} \cdot \vartheta_n(t)$. Since $\Delta_1(0) = 0$, the function ϑ_1 and η_1 can be chosen so that $\vartheta_1(0) = 0$ and $\boldsymbol{\eta}_1(0) = 0$ resp. These initial conditions are satisfied only by the unique solution $\boldsymbol{\eta} \equiv 0$. So by our hypotheses the sequence $\Delta_n(t)$ converges uniformly to this solution on every finite interval (see e.g. [4] and [5]). Hence also $\Delta_n(t)$ converges uniformly to zero on every finite interval that proves our first assertion. These two assertions together show that $\mathbf{x}(t)$ — the solution of (54) — satisfies the relation (66) (Putting that in the matrices occurring in (54), the solution of (54) satisfies the equation

$$\dot{\mathbf{x}} = \{\mathbf{A} + \mathbf{B}[\mathbf{x}(t)] + \mathbf{C}[\mathbf{x}(t)]\}\mathbf{x} + \mathbf{D}[\mathbf{x}(t)]\boldsymbol{\varphi}(\mathbf{x}) \tag{73}$$

where

$$\int_0^t \|\mathbf{C}[x(\tau)]\| d\tau < t^{\nu_0}; \quad \int_0^\infty \|\mathbf{D}[\mathbf{x}(\tau)]\| d\tau < \infty \tag{74}$$

Then, by theorems 1., 2., and 6. the solution of (73) satisfies the estimation

$$\|\mathbf{x}(t)\| \leqq \exp(-\varepsilon_0^* t) \quad \text{if} \quad \varepsilon_0^* < \varepsilon.$$

So we have our theorem completely proved.

We may mention that this theorem can also be proved by means of the same idea if **B**, **C** and **D** — satisfying suitable conditions — depend on t also.

REFERENCES

[1] MINORSKY, : *Nonlinear Oscillations*, Princeton, 1962.

[2] CESARI, L.: Un nuovo criterio di stabilita per le solutioni delle equazioni differentiali lineari, *Ann. Scuola Norm. Sup. Pisa* **9** (1940).

[3] CESARI, L.: *Asymptotic Behaviour and Stability Problems in Ord. Diff. Equ.*, Springer, 1959.

[4] CODDINGTON, E. A. and LEVINSON, N.: *Theory of Ord. Diff. Equ.* McGraw-Hill, 1955.

[5] FREY, T.: Über die Konvergenz von Iterationsfolgen mit veränderlichen Operatoren, *Studia Sci. Math. Hungar.* **2** (1967) 91—141.

[6] BIHARI, I.: A generalization of a lemma of Bellman..., Acta Math. Sci. Hungar. **7** (1956) 81—94.

[7] BIHARI, I.: Researches of the Boundedness and Stability of the Solutions of non-linear Diff. Equ. Acta Math. Sci. Hung. **8** (1957) 261—278.

[8] FREY, T.: Einige neue Methoden zur Berechnung von Eigenwerten, *Apl. Mat.* **10** (1965) 206—212.

COMPUTING CENTER OF THE HUNGARIAN ACADEMY OF SCIENCES,
BUDAPEST

(Received December 10, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 385—390.

NEGATIVE RESULTS IN THE THEORY OF RATIONAL APPROXIMATION

by

J. SZABADOS

In 1964 D. J. NEWMAN [1] proved his famous theorem that the function $|x|$ can be approximated by rational functions of degree n uniformly in $[-1, +1]$ with an error $3e^{-\sqrt{n}}$ at most. Since then a lot of classes of functions were constructed for which the rational approximation is better than the polynomial one. However, the classical class of functions $\mathrm{Lip}\,\alpha$ $(0<\alpha\leqq 1)$ has not this advantage. As D. J. NEWMAN [2] proved, there exists a function $f(x)\in\mathrm{Lip}\,\alpha$ $(0<\alpha<1)$ for which the order of rational approximation is not better than $n^{-\alpha}(\log n)^{-3}$ (It is well-known that the order of the polynomial approximation in this class is $O(n^{-\alpha})$.)

In this note we improve this negative result and give a generalization of the problem for more general classes of functions.

Let $\omega(h)\not\equiv 0$ be an arbitrary module of continuity in the interval $[-1, +1]$ and denote by $C(\omega)$ that class of functions $f(x)$ for which

$$\sup_{0<h\leqq 2}\frac{\omega(f, h)}{\omega(h)}<\infty$$

where $\omega(f, h)$ denotes the module of continuity of $f(x)$. (If $\omega(h)=h^{\alpha}$ $(0<\alpha\leqq 1)$ then evidently $C(\omega)=\mathrm{Lip}\,\alpha$.) Further let $r_n(x)$ be the best approximating rational function to $f(x)$ of degree n at most, and

$$R_n(f)=\max_{-1\leqq x\leqq +1}|f(x)-r_n(x)| \qquad (n=0, 1, 2, \ldots).$$

THEOREM 1. *If*

$$\lim_{h=+0}\frac{h}{\omega(h)}=0 \tag{1}$$

then there exists a function $f(x)\in C(\omega)$ *for which*

$$\limsup_{n\to\infty}\frac{R_n(f)}{\omega\left(\frac{1}{n}\right)}>0.$$

PROOF. Consider the indices $0=n_0<n_1<\ldots$ for which the following two conditions hold:

$$2\omega\left(\frac{1}{9^{n_i+1}}\right)\leqq\omega\left(\frac{1}{9^{n_i}}\right), \qquad (i=0, 1, \ldots) \tag{2}$$

and

$$2 \cdot 9^{n_i} \omega\left(\frac{1}{9^{n_i}}\right) \leqq 9^{n_{i+1}} \omega\left(\frac{1}{9^{n_{i+1}}}\right) \qquad (i = 0, 1, \ldots) \tag{3}$$

(the latter is possible because of (1)). Let

$$T_n(x) = \cos(n \arccos x),$$

$$f(x) = \sum_{i=0}^{\infty} \omega\left(\frac{1}{9^{n_i}}\right) T_{9^{n_i}}\left(\frac{x}{2}\right) \qquad (-1 \leqq x \leqq +1). \tag{4}$$

Here the right hand series converges uniformly in $[-1, +1]$ because of (2). We prove that $f(x) \in C(\omega)$. Let $-1 \leqq x < x+h \leqq +1$ and

$$9^{n_j} \leqq \frac{2}{h} < 9^{n_{j+1}}.$$

Then, using (2), (3), (4), the Lagrange's mean-value theorem, the Markov's inequality and the relations

$$\max_{-1 \leqq x \leqq +1} |T_n(x)| = 1, \qquad \frac{t\omega(T)}{T} \leqq 2\omega(t) \qquad (t \leqq T),$$

we get

$$|f(x+h) - f(x)| \leqq \sum_{i=0}^{j} \omega\left(\frac{1}{9^{n_i}}\right) \left| T_{9^{n_i}}\left(\frac{x+h}{2}\right) - T_{9^{n_i}}\left(\frac{x}{2}\right) \right| +$$

$$+ \sum_{i=j+1}^{\infty} \omega\left(\frac{1}{9^{n_i}}\right) \left| T_{9^{n_i}}\left(\frac{x+h}{2}\right) - T_{9^{n_i}}\left(\frac{x}{2}\right) \right| \leqq \frac{h}{2} \sum_{i=0}^{j} \omega\left(\frac{1}{9^{n_i}}\right) \max_{-\frac{1}{2} \leqq x \leqq \frac{1}{2}} |T'_{9^{n_i}}(x)| +$$

$$+ 2 \sum_{i=j+1}^{\infty} \omega\left(\frac{1}{9^{n_i}}\right) \leqq \frac{h}{\sqrt{3}} \sum_{i=0}^{j} 9^{n_i} \omega\left(\frac{1}{9^{n_i}}\right) + 2 \sum_{i=j+1}^{\infty} \omega\left(\frac{1}{9^{n_{j+1}}}\right) \frac{1}{2^{i-j-1}} \leqq$$

$$\leqq \frac{h}{\sqrt{3}} 9^{n_j} \omega\left(\frac{1}{9^{n_j}}\right) \sum_{i=0}^{j} \frac{1}{2^{j-i}} + 4\omega(h) \leqq \frac{4}{\sqrt{3}} \omega\left(\frac{h}{2}\right) + 4\omega(h) \leqq 7\omega(h),$$

i.e. $f(x) \in C(\omega)$.

Now consider the polynomial

$$p_{9^{n_{k-1}}}(x) = \sum_{i=0}^{k-1} \omega\left(\frac{1}{9^{n_i}}\right) T_{9^{n_i}}\left(\frac{x}{2}\right)$$

of degree $9^{n_{k-1}}$. Let

$$x_j = 2 \cos \frac{j\pi}{9^{n_k}} \qquad \left(\frac{1}{3} \cdot 9^{n_k} \leqq j \leqq \frac{2}{3} \cdot 9^{n_k}\right) \tag{5}$$

then $-1 \leqq x_j \leqq +1$, and

$$T_{9^{n_i}}\left(\frac{x_j}{2}\right) = \cos(9^{n_i - n_k} j\pi) = (-1)^j \qquad \left(i \geqq k, \ \frac{1}{3} \cdot 9^{n_k} \leqq j \leqq \frac{2}{3} \cdot 9^{n_k}\right). \tag{6}$$

Thus

$$f(x_j)-p_{9^{n_k-1}}(x_j)=\sum_{i=k}^{\infty}\omega\left(\frac{1}{9^{n_i}}\right)T_{9^{n_i}}\left(\frac{x_j}{2}\right)=(-1)^j\sum_{i=k}^{\infty}\omega\left(\frac{1}{9^{n_i}}\right) \tag{7}$$

i.e. $f(x)-p_{9^{n_k-1}}(x)$ assumes the maximum of its absolute value in $[-1, +1]$ with alternative signs at the points x_j. The number of these points is $\frac{1}{3}\cdot 9^{n_k}+1$. Now we can apply the following theorem of Čebyšev: Let $f(x)$ be a continuous function in $[-1, +1]$ and $p_n(x)$ a polynomial of degree n at most. If $f(x)-p_n(x)$ assumes the maximum of its absolute value in $[-1, +1]$ with alternative signs at $2n+2$ consequitive points in $[-1, +1]$ then $p_n(x)$ is the best approximating rational function of degree n at most to $f(x)$ in $[-1, +1]$. Being

$$2\cdot 9^{n_k-1}+2\leqq 2\cdot 9^{n_k-1}+2\leqq\frac{1}{3}\cdot 9^{n_k}+1 \qquad (k=1,2,\ldots) \tag{8}$$

we conclude that $p_{9^{n_k-1}}(x)=r_{9^{n_k-1}}(x)$ is the best approximating rational function to $f(x)$ of degree 9^{n_k-1} at most, and by (7)

$$R_{9^{n_k-1}}(f)=\sum_{i=k}^{\infty}\omega\left(\frac{1}{9^{n_i}}\right)>\omega\left(\frac{1}{9^{n_k}}\right)\geqq\frac{1}{10}\,\omega\left(\frac{1}{9^{n_k-1}}\right) \qquad (k=1,2,\ldots). \tag{9}$$

Hence Theorem 1 is proved.

Especially, for the class $\operatorname{Lip}\alpha$ $(0<\alpha<1)$ we have a stronger result:

COROLLARY. There exists a function $f(x)\in\operatorname{Lip}\alpha$ $(0<\alpha<1)$ for which

$$\liminf_{n\to\infty} n^{\alpha}R_n(f)>0.$$

Namely, we may apply Theorem 1 with $n_i=\left[\frac{i}{\alpha(1-\alpha)}\right]$, $\omega(h)=h^{\alpha}$. Then evidently

$$\frac{3}{4\alpha(1-\alpha)}\leqq n_{i+1}-n_i\leqq\frac{5}{4\alpha(1-\alpha)}. \tag{10}$$

Thus, if n is an arbitrary positive integer, and $9^{n_k-1}\leqq n<9^{n_{k+1}-1}$ then by (9) and (10) (putting $\omega(h)=h^{\alpha}$)

$$R_n(f)\geqq R_{9^{n_{k+1}-1}}(f)\geqq\frac{1}{10}\cdot\frac{1}{9^{(n_{k+1}-1)\alpha}}=\frac{1}{10}\cdot\frac{1}{9^{(n_{k+1}-n_k)\alpha}}\cdot\frac{1}{9^{(n_k-1)\alpha}}\geqq$$

$$\geqq\frac{1}{10\cdot 3^{\frac{5\alpha}{2(1-\alpha)}}}\cdot\frac{1}{n^{\alpha}}.$$

REMARKS. 1. It is well-known that for all $f(x)\in C(\omega)$

$$R_n(f)\leqq E_n(f)=O\left(\omega\left(\frac{1}{n}\right)\right)$$

holds, where $E_n(f)$ denotes the best approximation of $f(x)$ by polynomials of degree

n at most. Therefore Theorem 1 shows that the rational approximation in $C(\omega)$ ($\neq$ Lip 1) is generally not better than the polynomial.

2. It can be seen from the proof of our theorem that the best approximating polynomial and rational function of degree n is the same if

$$9^{n_k-1} \leqq n \leqq 9^{n_k}-1 \qquad (k = 1, 2, \ldots).$$

It is worthy of note that the interval $(9^{n_k-1}, 9^{n_k}-1)$ can be as long as we want.

3. E. P. DOLZHENKO [3] published the following theorem (without proof). *Let $\omega(h)$ be an arbitrary module of continuity. Then there exist a continuous function $f(x)$ and constants c_1, c_2, c_3 for which the relations* a) $c_1\omega(h) \leqq \omega(f, h) \leqq c_2\omega(h)$, b) $E_n(f) \leqq c_3 R_n(f)$, c) $E_{2\cdot 9^k}(f) = R_{2\cdot 9^k}(f)$ *hold.* This theorem states that the best approximating rational function of degree $2\cdot 9^k$ can be a polynomial, but does not give information about the exact order of $R_n(f)$.

Now we turn to the class Lip 1 which was excluded in Theorem 1 by the condition (1). It is well-known that for all $f(x) \in$ Lip 1

$$R_n(f) \leqq E_n(f) = O\left(\frac{1}{n}\right). \tag{11}$$

THEOREM 2. *Let $\varepsilon_n > 0$ $(n = 1, 2, \ldots)$ and assume that ε_n converges to 0 arbitrarily slowly. Then there exists a function $f(x) \in$ Lip 1 for which the relation*

$$R_n(f) = O\left(\frac{\varepsilon_n}{n}\right) \tag{12}$$

does not hold.

PROOF. Let us define the indices $n_1 < n_2 < \ldots$ such that

$$\varepsilon_{9^{n_i}-1} \leqq \frac{1}{i^4} \qquad (i = 1, 2, \ldots).$$

Consider

$$f(x) = \sum_{i=1}^{\infty} \frac{\sqrt{\varepsilon_{9^{n_i}-1}}}{9^{n_i}} T_{9^{n_i}}\left(\frac{x}{2}\right) \qquad (-1 \leqq x \leqq +1).$$

Evidently the right hand side series converges uniformly in $[-1, +1]$. Let $-1 \leqq x \leqq x+h \leqq +1$ then

$$|f(x+h)-f(x)| \leqq \frac{\sqrt{\varepsilon_{9^{n_i}-1}}}{9^{n_i}} \left| T_{9^{n_i}}\left(\frac{x+h}{2}\right) - T_{9^{n_i}}\left(\frac{x}{2}\right) \right| \leqq$$

$$= \frac{h}{\sqrt{3}} \sum_{i=1}^{\infty} \sqrt{\varepsilon_{9^{n_i}-1}} \leqq \frac{h}{\sqrt{3}} \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6\sqrt{3}} h$$

i.e. $f(x) \in$ Lip 1. The polynomial

$$p_{9^{n_k}-1}(x) = \sum_{i=1}^{k-1} \frac{\sqrt{\varepsilon_{9^{n_i}-1}}}{9^{n_i}} T_{9^{n_i}}\left(\frac{x}{2}\right)$$

is of degree 9^{n_k-1} at most. As above, it is easy to see — by (5), (6) and (8) — that $p_{9^{n_k-1}}(x)=r_{9^{n_k-1}}(x)$ is the best approximating rational function of degree 9^{n_k-1} at most, and

$$R_{9^{n_k-1}}(f)=\sum_{i=k}^{\infty}{}'\frac{\sqrt{\varepsilon_{9^{n_i-1}}}}{9^{n_i}}\geqq\frac{1}{9}\cdot\frac{\sqrt{\varepsilon_{9^{n_k-1}}}}{9^{n_k-1}}$$

i. e.

$$\frac{9^{n_k-1}\cdot R_{9^{n_k-1}}(f)}{\varepsilon_{9^{n_k-1}}}\geqq\frac{1}{9\sqrt{\varepsilon_{9^{n_k-1}}}}\to\infty\qquad(k\to\infty)$$

which proves (12).

REMARK. D. J. NEWMAN asked (cf [4], p. 189) whether for all $f(x)\in\mathrm{Lip}\,1$ the relation $R_n(f)=o\left(\frac{1}{n}\right)$ holds. Theorem 2 does not solve totally this problem but gives a sharp lower estimation for $\sup_{f\in\mathrm{Lip}\,1}R_n(f)$. The above mentioned DOLZHENKO's theorem does not give information about this question (we do not know the order of $E_{2\cdot 9^k}(f)=R_{2\cdot 9^k}(f)$).

Finally, we investigate the Zygmund's class Z ($f(x)\in Z$ if $|f(x+h)-2f(x)+f(x-h)|=O(h)$) which is between all $\mathrm{Lip}\,\alpha\,(0<\alpha<1)$ and $\mathrm{Lip}\,1$. It is well-known that (11) holds for all $f(x)\in Z$, too.

THEOREM 3. *There exists a function $f(x)\in Z$ for which*

$$\liminf_{n\to\infty}n\cdot R_n(f)>0. \tag{13}$$

PROOF. Let

$$f(x)=\sum_{i=0}^{\infty}{}'\frac{1}{9^i}T_{9^i}\left(\frac{x}{2}\right)\qquad(-1\leqq x\leqq+1).$$

Here the right hand side series converges uniformly in $[-1,\ +1]$. We have for $-1\leqq x-h\leqq x+h\leqq+1$

$$\left|T_n\left(\frac{x+h}{2}\right)-2T_n\left(\frac{x}{2}\right)+T_n\left(\frac{x-h}{2}\right)\right|=\frac{h}{2}|T_n'(\xi)-T_n'(\eta)|=\frac{h|\xi-\eta|}{2}|T_n''(\zeta)|\leqq$$

$$\leqq\frac{h^2}{2}\max_{-\frac{1}{2}\leqq x\leqq+\frac{1}{2}}|T_n''(x)|\leqq\frac{h^2n}{\sqrt{3}}\max_{-\frac{1}{2}\leqq x\leqq+\frac{1}{2}}|T_n'(x)|\leqq\frac{2h^2n^2}{3}$$

$$\left(-\frac{1}{2}\leqq\frac{x-h}{2}\leqq\eta\leqq\zeta\leqq\xi\leqq\frac{x+h}{2}\leqq+\frac{1}{2}\right).$$

Now let

$$9^j\leqq\frac{2}{h}<9^{j+1}$$

then for $-1 \leqq x-h \leqq x+h \leqq +1$

$$|f(x+h)-2f(x)+f(x-h)| = \sum_{i=0}^{\infty} \frac{1}{9^i} \left| T_{9^i}\left(\frac{x+h}{2}\right) - 2T_{9^i}\left(\frac{x}{2}\right) + T_{9^i}\left(\frac{x-h}{2}\right) \right| =$$

$$= \sum_{i=0}^{j} + \sum_{i=j+1}^{\infty} \leqq \sum_{i=0}^{j} \frac{1}{9^i} \cdot \frac{2h^2 9^{2i}}{3} + \sum_{i=j+1}^{\infty} \frac{3}{9^i} = \frac{2h^2}{3} \cdot 9^j \sum_{i=0}^{j} \frac{1}{9^{j-i}} +$$

$$+ \frac{3}{9^{j+1}} \sum_{i=j+1}^{\infty} \frac{1}{9^{i-j-1}} \leqq \frac{4h}{3} \cdot \frac{9}{8} + \frac{3h}{2} \cdot \frac{9}{8} = \frac{51}{16} h$$

i.e. $f(x) \in Z$. The polynomial

$$p_{9^{k-1}}(x) = \sum_{i=0}^{k-1} \frac{1}{9^i} T_{9^i}\left(\frac{x}{2}\right)$$

is of degree 9^{k-1}. From (5), (6) and (8) — by the substitution $n_k = k$, $n_i = i$ — we see that $p_{9^{k-1}}(x) = r_{9^{k-1}}(x)$ is the best approximating rational function of degree 9^{k-1} at most. Thus

$$R_{9^{k-1}}(f) = \sum_{i=k}^{\infty} \frac{1}{9^i} = \frac{1}{8} \cdot \frac{1}{9^{k-1}}. \tag{14}$$

Now let n be an arbitrary positive integer, and $9^{k-1} \leqq n < 9^k$. Then by (14)

$$R_n(f) \geqq R_{9^k}(f) = \frac{1}{8} \cdot \frac{1}{9^k} \geqq \frac{1}{72} \cdot \frac{1}{n} \qquad (n = 1, 2, \ldots),$$

i.e. (13) holds, qu.e.d.

I am deeply indebted to Mr. G. FREUD for his valuable remarks in preparation of this paper.

REFERENCES

[1] NEWMAN, D. J.: Rational approximation to $|x|$, *Michigan Math. J.* **11** (1964) 11—14.
[2] SZÜSZ, P. és TURÁN, P.: A konstruktív függvénytan egy újabb irányáról (Hungarian), *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.* **16** (1966) 33—45.
[3] DOLZHENKO, E. P.: Uniform approximation by rational functions (algebraic and trigonometrical), and global functional properties (Russian), *Dokl. Akad. Nauk SSSR*, **166** (1966) 526—529.
[4] *On approximation theory* International series of Numerical Mathematics, Vol. 5; Birkhäuser Verlag, Basel, 1964.

EÖTVÖS L. UNIVERSITY, BUDAPEST

(Received December 10, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 391—398.

ÜBER DIE ANZAHL DER KNOTENPUNKTE EINES LÄNGSTEN KREISES IN PLANAREN, KUBISCHEN, DREIFACH KNOTENZUSAMMENHÄNGENDEN GRAPHEN

von

H. WALTHER

Es bezeichne G einen *planaren, kubischen, dreifach knotenzusammenhängenden Graphen*, $V(G)$ sei die *Anzahl seiner Knotenpunkte*, $K(G)$ sei die *Anzahl der Knotenpunkte eines längsten Kreises*, $M(G)$ *sei die Anzahl der Knotenpunkte, die in allen längsten Kreisen liegen*, $P(G)$ sei die *Anzahl der Knotenpunkte, die in keinen längsten Kreis liegen.* $|A|$ ist die *Anzahl der in* der Menge A *liegenden Elemente.*

In der vorliegenden Arbeit werden wir einen Satz beweisen, der Auskunft über die Anzahl der Knotenpunkte gibt, die in keinem bzw. allen längsten Kreisen der Graphen einer nachstehend konstruierten Graphenfolge $\{E_n\}$ liegen.

HAUPTSATZ. *Es gibt eine Folge* $\{E_n\}$ *von planaren, kubischen, dreifach knotenzusammenhängenden Graphen, für die gilt:*

(a) $$K(E_n) = M(E_n) \qquad (n = 1, 2, \ldots)$$

(b) $$\lim_{n\to\infty} \frac{K(E_n)}{V(E_n)} = 0$$

(c) $$\lim_{n\to\infty} \frac{P(E_n)}{V(E_n)} = 1$$

Mit anderen Worten: *Der relative Anteil der Knotenpunkte eines längsten Kreises von* E_n *an der Gesamtknotenzahl strebt nach Null, während der relative Anteil der in keinem längsten Kreis von* E_n *liegenden Knotenpunkte an der Gesamtknotenanzahl nach Eins strebt. Ferner liegt ein Knotenpunkt genau dann in einem längsten Kreis, wenn er in allen längsten Kreisen von* E_n *liegt.*

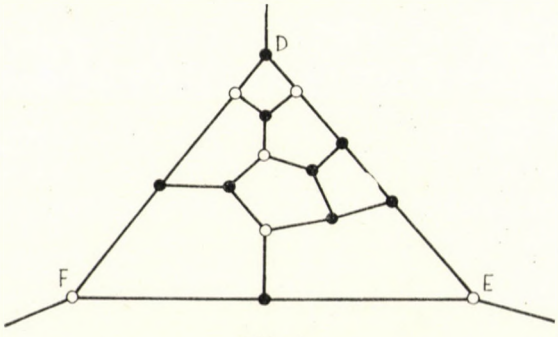


Abb. 1

Zum Beweis des Satzes benötigen wir zunächst einige Hilfssätze.

HILFSSATZ 1. *Ist das Gebilde von Abb. 1 Teil eines Graphen G mit einem Hamiltonkreis H, dann enthält H den Kantenansatz D.* Den Beweis lieferte W. T. TUTTE [1]. Aus diesem Hilfssatz folgt unmittelbar der

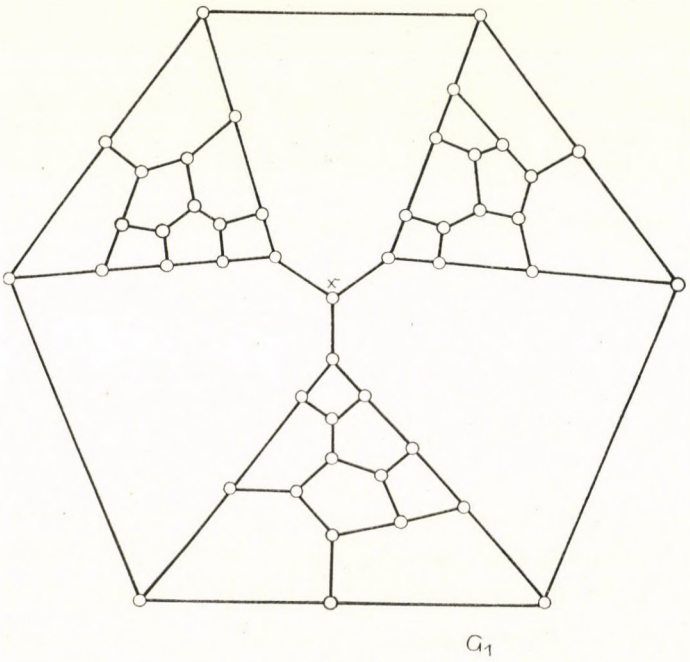


Abb. 2

HILFSSATZ 2. *Der Graph G_1 von Abb. 2 besitzt keinen Hamiltonkreis.* Denn wäre der Hilfssatz falsch, so müßte der Knotenpunkt x mit drei in dem Hamiltonkreis liegenden Kanten inzidieren, was aber unmöglich ist.

HILFSSATZ 3. *Ist das Gebilde H_1* (Abb. 3 bzw. 4 bzw. 5) *Teil eines Graphen G mit einem längsten Kreis K und liegen Knotenpunkte von H_1 in K, dann liegen genau* 54 *der* 55 *Knotenpunkte von H_1 in K.*

BEWEIS. Da nicht alle Knotenpunkte von G in H_1 liegen, kann K nicht ganz in H_1 verlaufen. Es liegen also genau zwei der drei Kantenansätze A, B, C in K. Angenommen, alle 55 Knotenpunkte von H_1 liegen in K. Dann liegen nach Hilfssatz 1 alle drei Kantenansätze D (Abb. 1) in K. Das ist aber ein Widerspruch. Es liegen also höchstens 54 der 55 Knotenpunkte von H_1 in K. Wie die Abb. 3, 4, 5 zeigen, gibt es auch Kreise K, die genau 54 Knotenpunkte von H_1 enthalten, sofern sie überhaupt Knotenpunkte von H_1 enthalten.

Wie die Abb. 3, 4, 5 zeigen, ist es obendrein möglich, den Teil von K in H_1 so zu wählen, daß der Knotenpunkt z nicht in K liegt, welche zwei der drei Kantenansätze A, B, C auch in K liegen.

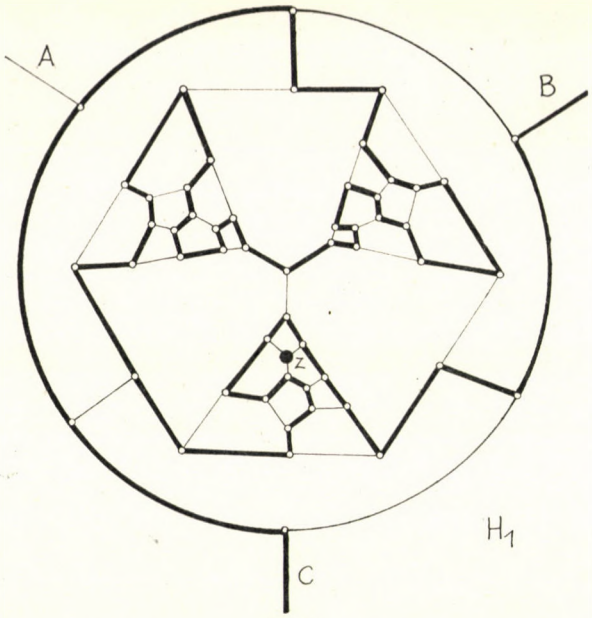


Abb. 3

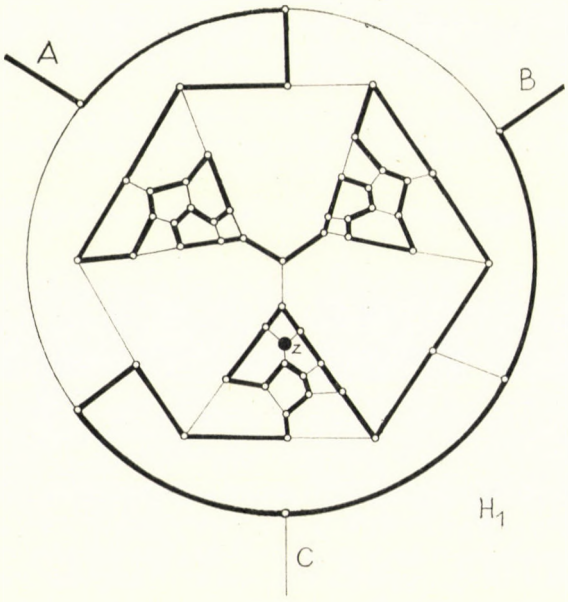


Abb. 4

HILFSSATZ 4. *Das Gebilde H_2* (Abb. 6) *entstehe dadurch, daß jeder von z verschiedene Knotenpunkt des Gebildes H_1 durch ein Dreieck ersetzt wird, der Knotenpunkt z aber ungeändert bleibt. Ist H_2 Teil eines Graphen G mit einem längsten Kreis K und liegen Knotenpunkte von H_2 in K* (es liegen genau zwei der drei Kantenansätze A', B', C' in K), *dann liegen genau* 162 *der* 163 *Knotenpunkte von H_2 in K, der Knotenpunkt z aber liegt nicht in K.*

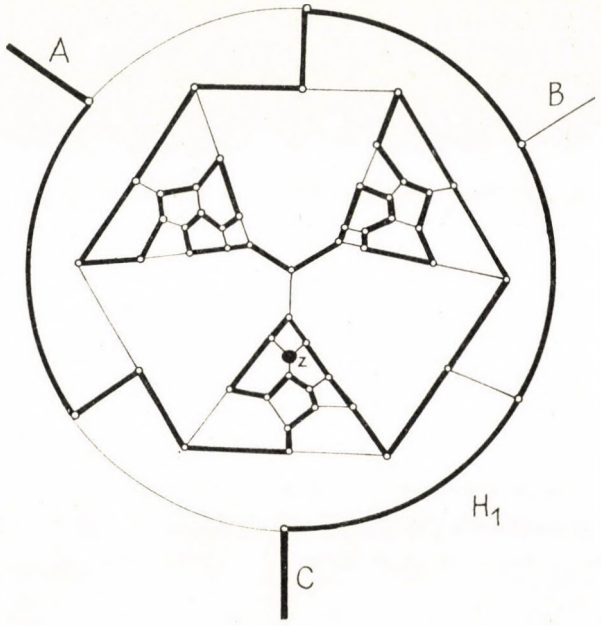


Abb. 5

BEWEIS. Der in H_2 liegende Teil von K sei der Weg W, der zwei der drei Kantenansätze von H_2 miteinander verbindet.

(a) W enthält nicht alle 163 *Knotenpunkte.* Angenommen, W enthielte alle 163 Knotenpunkte. Zieht man H_2 auf H_1 zusammen (das Einsetzen der Dreiecke in H_1, das zu H_2 führte, wird rückgängig gemacht), so geht W in einen zwei Kantenansätze von H_1 verbindenden Weg W' über, der alle Knotenpunkte von H_1 enthält. Das ist aber ein *Widerspruch zu Hilfssatz* 3.

(b) Es gibt einen Weg W, der 162 *Knotenpunkte von H_2 enthält, jedoch nicht z.* Die Abb. 3, 4, 5 zeigen Wege W_1', W_2', W_3', die jeweils zwei der drei Kantenansätze A, B, C von H_1 verbinden, 54 Knotenpunkte enthalten, jedoch nicht den Knotenpunkt z. Ist nun x ein Knotenpunkt, der in W_i' liegt, so kann man auch einen Weg W_i in H_2 finden, der die 3 Knotenpunkte des in x eingesetzten Dreiecks enthält. Aus den Wegen W_i' kann man also Wege W_i von H_2 konstruieren, die 162 Knotenpunkte von H_2 enthalten, jedoch nicht den Knotenpunkt z.

(c) Es gibt keinen längsten Weg W in H_2, der zwei der drei Kantenansätze verbindet und den Knotenpunkt z enthält. Angenommen, *(c)* wäre falsch. Da 162 Knotenpunkte in W liegen (wegen *(b)*) und unter ihnen der Knotenpunkt z, liegt

ein von z verschiedener Knotenpunkt x nicht in W, der gemäß der Konstruktion von H_2 in einem Dreieck Δ liegt. Es liegen jedoch Knotenpunkte dieses Δ in W, da W andernfalls weniger als 162 Knotenpunkte enthielte. Liegen aber Knotenpunkte von Δ in W, dann liegen gemäß den Überlegungen zu *(b)* alle drei Knotenpunkte von Δ in W, da W ein längster Weg ist. Das ist aber ein Widerspruch. Damit ist der Hilfssatz vollständig bewiesen.

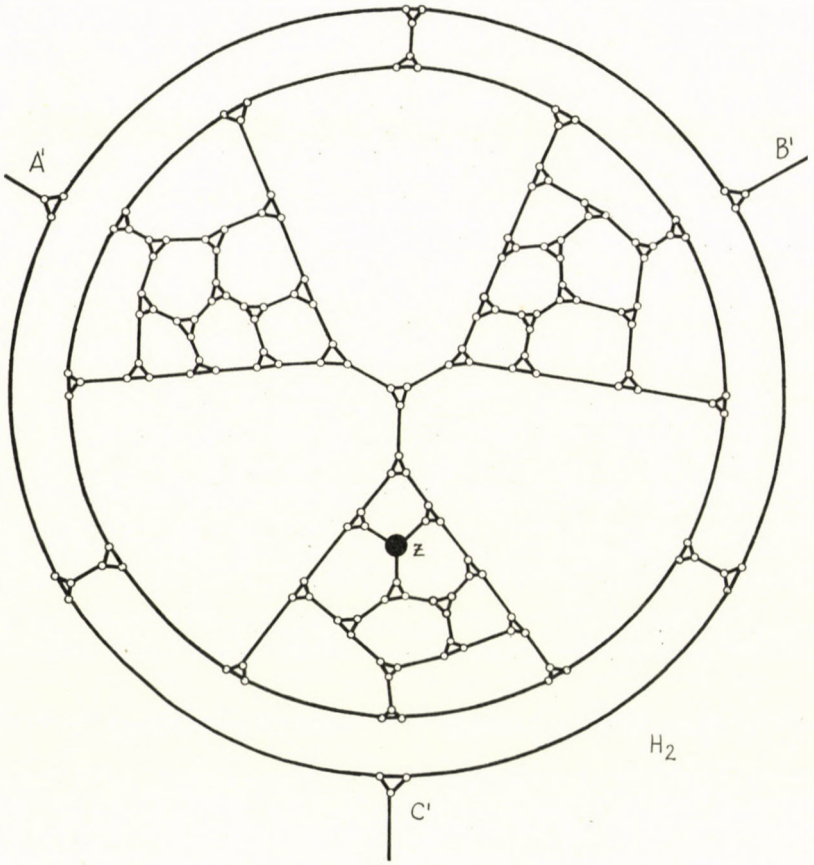


Abb. 6

Wir konstruieren nun die Graphenfolge $\{E_n\}$. E_1 sei der Graph von Abb. 7. Er entsteht aus dem TUTTEschen Graphen von Abb. 2, indem jeder Knotenpunkt außer z durch ein Dreieck ersetzt wird, z aber ungeändert bleibt. E_1 hat die Eigenschaft, daß jeder längste Kreis K_1 von E_1 alle Knotenpunkte von E_1 mit Ausnahme von z enthält (folgt aus Hilfssatz 2). E_2 entstehe, indem jeder Knotenpunkt von E_1 durch ein Gebilde H_2 (Abb. 6) ersetzt wird ... E_n entstehe, indem jeder Knotenpunkt von E_{n-1} durch ein Gebilde H_2 (Abb. 6) ersetzt wird.

DEFINITION. Ein Knotenpunkt x_j von E_j heißt *Nachkomme* eines Knotenpunktes x_i von E_i $(i \leqq j)$, wenn beim Zusammenziehen von E_j auf E_i (das Einsetzen der Gebilde H_2, das von E_i zu E_j führte, wird rückgängig gemacht) der Knotenpunkt x_j in den Knotenpunkt x_i übergeht. Entsprechend wollen wir x_i *Vorfahren* von x_j nennen. Ein Knotenpunkt ist also auch sein eigener Nachkomme und Vorfahr.

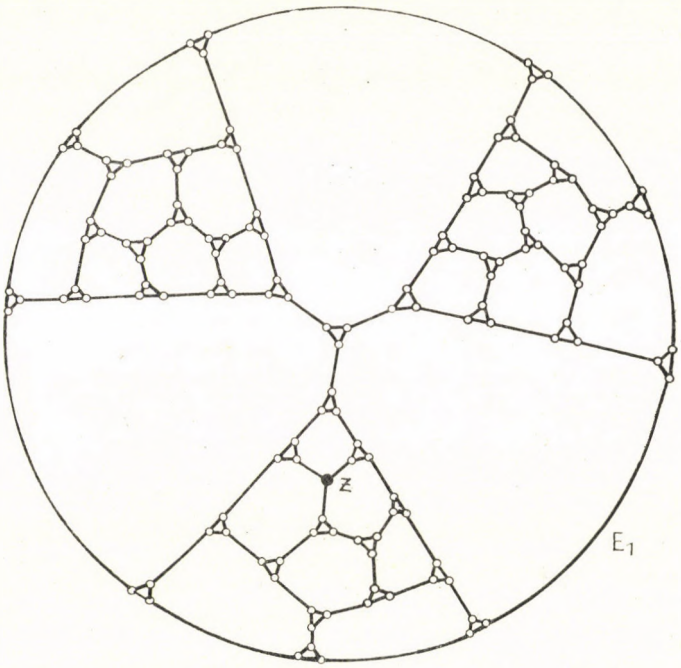


Abb. 7

HILFSSATZ 5. *Sei K_n ein längster Kreis von E_n. Dann geht K_n beim Zusammenziehen von E_n auf E_{n-1} in einen längsten Kreis K_{n-1} von E_{n-1} über.*

BEWEIS. Zunächst ist klar, daß das Gebilde K'_{n-1}, in das K_n beim Zusammenziehen von E_n auf E_{n-1} übergeht, ein Kreis ist. Angenommen, K'_{n-1} ist kein längster Kreis von E_{n-1}, es gelte also $|K'_{n-1}| < |K_{n-1}|$. Da jeder Knotenpunkt von E_{n-1} beim Übergang zu E_n durch ein Gebilde H_2 ersetzt wurde, enthält K_n Knotenpunkte aus $|K'_{n-1}|$ Gebilden H_2. Wegen Hilfssatz 4 enthält K_n aus jedem dieser Gebilde H_2 genau 162 Knotenpunkte, also gilt $|K_n| = |K'_{n-1}| \cdot 162$. Entsprechend kann man aus dem längsten Kreis K_{n-1} von E_{n-1} einen Kreis K'_n von E_n konstruieren, der aus jedem der $|K_{n-1}|$ Gebilde H_2 genau 162 Knotenpunkte enthält. Es gilt also

$$|K'_n| - |K_n| = 162|K_{n-1}| - 162|K'_{n-1}| = 162(|K_{n-1}| - |K'_{n-1}|) > 0.$$

Das ist aber ein Widerspruch, da K_n ein längster Kreis von E_n ist. Damit ist der Hilfssatz 5 bewiesen.

HILFSSATZ 6. *Ein Knotenpunkt x_n aus E_n, der Nachkomme eines Knotenpunktes z aus E_i $(i \leqq n)$ ist, liegt in keinem längsten Kreis von E_n.*

BEWEIS. Angenommen, der Hilfssatz wäre falsch, x_n liege in einem längsten Kreis K_n von E_n und x_n sei Nachkomme eines Knotenpunktes z von E_i. Beim Zusammenziehen von E_n auf E_i geht K_n gemäß Hilfssatz 5 in einen K_i über, der den Knotenpunkt z enthält. *Das ist* aber *ein Widerspruch zu Hilfssatz* 4 bzw., im Falle $i=1$ zu Hilfssatz 2. Damit ist der Hilfssatz 6 bewiesen.

HILFSSATZ 7. *Ein Knotenpunkt x_n aus E_n, der keinen Vorfahren z aus E_i $(i=1, 2, \ldots, n)$ besitzt, liegt in jedem längsten Kreis von E_n.*

BEWEIS. Angenommen, der Hilfssatz wäre falsch, es gäbe also einen Knotenpunkt x_n aus E_n, der keinen Vorfahren z besitzt und einen längsten Kreis K_n von E_n, dem x_n nicht angehört. Nach Hilfssatz 4 liegt kein Knotenpunkt desjenigen Gebildes H_2, in dem x_n liegt, in K_n. Beim Zusammenziehen von E_n auf E_{n-1} geht K_n in einen längsten Kreis K_{n-1} von E_{n-1} über, und das Gebilde H_2, in dem x_n liegt, gehe in den Knotenpunkt x_{n-1} von E_{n-1} über. Dann liegt x_{n-1} nicht in K_{n-1}. Nach Hilfssatz 4 liegt kein Knotenpunkt des Gebildes H_2, in dem x_{n-1} liegt, in K_{n-1}, da x_{n-1} kein Nachkomme eines z ist. Beim Zusammenziehen von E_2 auf E_1 geht K_2 in einen längsten Kreis K_1 von E_1 über, und das Gebilde H_2, in welchem x_2 liegt, gehe in den Knotenpunkt x_1 über. Dann liegt x_1 nicht in K_1. Da x_1 kein z ist, liegen also in dem längsten Kreis K_1 von E_1 weder der Knotenpunkt x noch der Knotenpunkt z. Das ist aber ein Widerspruch, da ein längster Kreis von E_1 genau 135 der 136 Knotenpunkte enthält. Damit ist der Hilfssatz bewiesen.

Damit ist bereits (a) bewiesen, denn wir haben gesehen, daß die Menge der Knotenpunkte x von E_n, die keinen Vorfahren z besitzen, in allen längsten Kreisen liegen, jedoch die Menge der Knotenpunkte x, die wenigstens einen Knotenpunkt z als Vorfahren haben, keinem längsten Kreis angehören.

Zum Beweis von *(b)* und *(c)* berechnen wir nun die Knotenpunktanzahl von E_n und die von K_n.

Da jeder Knotenpunkt von E_n beim Übergang zu E_{n+1} durch ein Gebilde H_2 mit 163 Knotenpunkten ersetzt wurde, gilt

$$V(E_{n+1}) = 163 \cdot V(E_n) = 163^n \cdot V(E_1).$$

Entsprechend gilt: Liegt ein Knotenpunkt x_n in K_n, dann liegen von den 163 in x_n eingesetzten Knotenpunkten des Gebildes H_2 genau 162 in K_{n+1}, es gilt also

$$K(E_{n+1}) = 162 \cdot K(E_n) = 162^n \cdot K(E_1).$$

Mit $V(E_1)=136$, $K(E_1)=135$ erhält man

$$(b) \qquad \lim_{n\to\infty} \frac{K(E_n)}{V(E_n)} = \lim_{n\to\infty} \frac{135 \cdot 162^{n-1}}{136 \cdot 163^{n-1}} = 0.$$

Aus *(a)* und *(b)* folgt unmittelbar *(c)*. Der dreifache Knotenzusammenhang der E_n ergibt sich aus der Konstruktion und der Tatsache, daß E_1 dreifach zusammenhängend ist.

Damit ist der Hauptsatz vollständig bewiesen.

In einer folgenden Arbeit werden wir einen entsprechenden Satz für die längsten Wege eines planaren, kubischen, dreifach zusammenhängenden Graphen bewesen.i

Eine Graphenfolge, die die Eigenschaft (b) des Hauptsatzes hat, wurde bereits von B. GRÜNBAUM und T. S. MOTZKIN (Longest simple paths in polyhedral graphs, J. London Math. Soc. 37(1962, 152—160) angegeben. Da dem Autor diese Arbeit erst nach der Drucklegung dieses Artikels zugänglich war, wird eine eingehende Diskussion in einer folgenden Arbeit von R. LANG und H. WALTHER erfolgen.

LITERATURVERZEICHNIS

[1] TUTTE, W. T.: On Hamiltonian circuits, *J. London Math. Soc.* **21** (1946) 98—101.
[2] KÖNIG, D.: *Theorie der endlichen und unendlichen Graphen,* Leipzig, 1936.

TECHNISCHE HOCHSCHULE, ILMENAU

(Eingegangen: 15. Dezember, 1966.)

Studia Scienttarum Mathematicarum Hungarica 2 (1967) 399—402.

A NEW PROOF OF THE THEOREM OF G. KATONA AND G. TUSNÁDY[1]

by

M. JACOBSEN

We consider an information source $[\mathscr{X}, P_X]$ where $\mathscr{X}$ is the space of all sequences $(\xi_1, \xi_2, \ldots)$ of letters from the n-letter alphabet X and where P_X is a probability measure on the cylinder σ-field on $\mathscr{X}$.

The message $(\xi_1, \xi_2, \ldots)$ is coded letter by letter into a sequence of letters from the alphabet Y which consists of m letters. The code is defined by a mapping c from X into the set of all finite sequences of elements from Y. $\mathscr{Y}$ is the space of all infinite sequences of elements from Y and P_Y is the probability induced by the code on the cylinder σ-field on $\mathscr{Y}$. (As proved in the paper [1], the mapping from $\mathscr{X}$ to $\mathscr{Y}$ determined by c is measurable with respect to the two cylinder σ-fields.)

The code is required to be uniquely decodable in the following sense: if $x', x'' \in X$, $x' \neq x''$, then neither of the finite sequences cx' nor cx'' must be a segment of the other. This condition is stronger than the one imposed in the paper [1], but it turns out to be essential for the proof.

The following notation is introduced: η denotes the function $\eta(t) = -t \log t$; $c(x_1, \ldots, x_N)$ is $(cx_1, \ldots, cx_N)$ considered as a finite sequence of elements from Y; $H(\mathscr{X}^N)$ is the entropy $\sum_{x_1, \ldots x_N \in X} \eta(P_X[x_1, \ldots, x_N]$ where $P_X[x_1, \ldots, x_N] = P_X\{\xi_1 = x_1, \ldots, \ldots, \xi_N = x_N\}$ while $H(\mathscr{X})$ denotes the entropy $\lim_{N \to \infty} \frac{1}{N} H(\mathscr{X}^N)$ provided this limit exists; $H(\mathscr{Y}^N)$ and $H(\mathscr{Y})$ are the corresponding entropies on $\mathscr{Y}$; finally $l^{(N)}$ is the random variable denoting the length of the sequence obtained by coding the first N letters from the message.

It is assumed that the entropy $H(\mathscr{X})$ and the average code length L do exist $\left(\text{the latter meaning that } \frac{1}{N} l^{(n)} \xrightarrow[\text{prob.}]{} L \text{ as } N \to \infty\right)$. The theorem may now be stated as follows:

Under the above assumptions the entropy $H(\mathscr{Y})$ does exist and

$$H(\mathscr{Y}) = \frac{1}{L} H(\mathscr{X})$$

In the proof we shall use the following two inequalities: if $a_1, \ldots, a_r$ are non-negative real numbers and if $A = \sum_{i=1}^{r} a_i$, then

$$\sum_{i=1}^{r} \eta(a_i) \leqq A(\log r - \log A) \tag{1}$$

[1] See [1].

if furthermore $a_i = \sum_{j=1}^{r_i} b_{ij}$, $b_{ij} \geqq 0$, then

$$\sum_{i=1}^{r} \sum_{j=1}^{r_i} \eta(b_{ij}) \geqq \sum_{i=1}^{r} \eta(a_i) \tag{2}$$

Of these (2) is obvious while (1) follows like this: $\sum_{i=1}^{r} \eta(a_i) = A \sum_{i=1}^{r} \eta\left(\frac{a_i}{A}\right) - A \log A \leqq$

$\leqq A(\log r - \log A)$ since $\sum_{i=1}^{r} \frac{a_i}{A} = 1$.

Another important result to be used in the sequel is that

$$P_X[x_1, \ldots, x_N] = P_Y[c(x_1, \ldots, x_N)] \tag{3}$$

for any choice of N and $x_i \in X$. This is easily verified, but it should be noted that here, for the first time, the strong condition of unique decodability is used.

We shall now proceed with the proof itself. So, let $\varepsilon > 0$ be given. Because of (3) we have

$$H(\mathcal{X}^N) = \sum_{\substack{x_1, \ldots, x_N \in X \\ l^{(N)} < (L+\varepsilon)N}} \eta\big(P_Y[c(x_1, \ldots, x_N)]\big) + \sum_{\substack{x_1, \ldots, x_N \in X \\ l^{(N)} \geqq (L+\varepsilon)N}} \eta(P_X[x_1, \ldots, x_N])$$

(Here, given $x_1, \ldots, x_N$, $l^{(N)}$ denotes the length of the sequence $c(x_1, \ldots, x_N)$). If we denote by S the first sum and by S' the second, we get using (1), that

$$S' \leqq a_N(N \log n - \log a_N)$$

for

$$a_N = P_X\left\{\frac{l^{(N)}}{N} \geqq L+\varepsilon\right\} = \sum_{\substack{x_1, \ldots, x_N \in X \\ l^{(N)} \geqq (L+\varepsilon)N}} P_X[x_1, \ldots, x_N]$$

To get an upper bound for S we proceed as follows: for a fixed sequence $(x_1, \ldots, x_N)$ such that $l^{(N)} < (L+\varepsilon)N$ we form all sequences of length $[(L+\varepsilon)N]$ from $Y^{[(L+\varepsilon)N]}$ having the sequence $c(x_1, \ldots, x_N)$ as a segment; thus each of the terms $P_Y[c(x_1, \ldots, x_N)]$ is split up into a sum and we can apply the inequality (2). Because of the unique decodability condition it follows that any two of the sequences from $Y^{[(L+\varepsilon)N]}$ thus formed corresponding to two different $(x_1, \ldots, x_N)$ — sequences will differ. Hence (2) gives an upper bound for S of the form $\sum \eta(P_Y[y_1, \ldots, y_{[(L+\varepsilon)N]}]$ where the sum is taken over some of the sequences from $Y^{[(L+\varepsilon)N]}$. Combining the above results we therefore obtain the following inequality:

$$H(\mathcal{X}^N) \leqq H(\mathcal{Y}^{[(L+\varepsilon)N]}) + a_N(N \log n - \log a_N). \tag{4}$$

In a similar fashion we shall get a lower bound for $H(\mathcal{X}^N)$. We have:

$$H(\mathcal{X}^N) \geqq T$$

where $T = \sum_{\substack{x_1, \ldots, x_N \in X \\ l^{(N)} > (L-\varepsilon)N}} \eta\big(P_Y[c(x_1, \ldots, x_N)]\big)$. Now, consider an arbitrary sequence $(y_1, \ldots, y_{[(L-\varepsilon)N]})$ from $Y^{[(L-\varepsilon)N]}$. With this sequence we associate all sequences $(x_1, \ldots, x_N)$ from X^N for which $l^{(N)} \geqq (L-\varepsilon)N$ and $(y_1, \ldots, y_{[(L-\varepsilon)N]})$ is a segment of $c(x_1, \ldots, x_N)$. Let A_N be the set of those sequences from $Y^{[(L-\varepsilon)N]}$ for which

there exist at least one such sequence $(x_1, \ldots, x_N)$ and let A_N^c denote the complement of A_N. Then because of (2)

$$\sum_{(y_1, \ldots, y_{[(L-\varepsilon)N]}) \in A_N} \eta(P_Y[y_1, \ldots, y_{[(L-\varepsilon)N]}]) \leqq T$$

as is seen by extending the sequences $(y_1, \ldots, y_{[(L-\varepsilon)N]})$ in all possible ways to sequences of the form $c(x_1, \ldots, x_N)$.

If $(y_1, \ldots, y_{[(L-\varepsilon)N]}) \in A_N^c$ it means that either is it impossible to obtain this sequence as the beginning of a coded message or there must exist a sequence $(x_1, \ldots, x_N)$ from X^N with $l^{(N)} < (L-\varepsilon)N$ such that $c(x_1, \ldots, x_N)$ is a segment of $(y_1, \ldots, y_{[(L-\varepsilon)N]})$. Because of the unique decodability, such a sequence, if it exists, is uniquely determined. If furthermore $(y_1, \ldots, y_{[(L-\varepsilon)N]})$ varies over the whole of A_N^c it is obvious that any sequence from X^N with $l^{(N)} < (L-\varepsilon)N$ is obtained by this correspondence. We can therefore conclude that

$$\sum_{(y_1, \ldots, y_{[(L-\varepsilon)N]}) \in A_N^c} P_Y[y_1, \ldots, y_{[(L-\varepsilon)N]}] = P_X\left\{\frac{1}{N} l^{(N)} - L < -\varepsilon\right\}$$

Denoting this probability by b_N and using (1) we obtain

$$\sum_{(y_1, \ldots, y_{[(L-\varepsilon)N]}) \in A_N^c} \eta(P_Y[y_1, \ldots, y_{[(L-\varepsilon)N]}]) - b_N([(L-\varepsilon)N] \log m - \log b_N)$$

Combining the various inequalities we get

$$H(\mathcal{X}^N) \geqq H(\mathcal{Y}^{[(L-\varepsilon)N]}) - b_N([(L-\varepsilon)N] \log m - \log b_N) \tag{5}$$

To obtain the desired result from (4) and (5) is not difficult. First, let us note that because of the definition of L $\lim_{N\to\infty} a_N = \lim_{N\to\infty} b_N = 0$. Defining $d_N(\varepsilon) = \left[\frac{N}{L+\varepsilon}\right]$ we have $[(L+\varepsilon)d_N(\varepsilon)] \leqq N$ and from (4) it then follows that

$$\begin{aligned} H(\mathcal{Y}^N) &\geqq H(\mathcal{Y}^{[(L+\varepsilon)d_N(\varepsilon)]}) \\ &\geqq H(\mathcal{X}^{d_N(\varepsilon)}) - a_{d_N(\varepsilon)}(d_N(\varepsilon) \log n - \log a_{d_N(\varepsilon)}) \end{aligned}$$

and therefore

$$\liminf_{N\to\infty} \frac{1}{N} H(\mathcal{Y}^N) \geqq \frac{1}{L+\varepsilon} H(\mathcal{X}) \tag{6}$$

Similarly, putting $e_N(\varepsilon) = \left[\frac{N}{L-\varepsilon}\right] + 3$, we have $[(L-\varepsilon)e_N(\varepsilon)] > N + 2(L-\varepsilon) - 1 \geqq N$ provided $\varepsilon \leqq \frac{1}{2}$ (since $L \geqq 1$) and hence

$$\begin{aligned} H(\mathcal{Y}^N) &\leqq H(\mathcal{Y}^{[(L-\varepsilon)e_N(\varepsilon)]}) \\ &\leqq H(\mathcal{X}^{e_N(\varepsilon)}) + b_{e_N(\varepsilon)}([(L-\varepsilon)e_N(\varepsilon)] \log m - \log b_{e_N(\varepsilon)}) \end{aligned}$$

using (5). Therefore

$$\limsup_{N\to\infty} \frac{1}{N} H(\mathcal{Y}^N) \leqq \frac{1}{L-\varepsilon} H(\mathcal{X}) \tag{7}$$

From (6) and (7) it finally follows for $\varepsilon \to 0$ that $H(\mathcal{Y})$ is defined and equals $\frac{1}{L} H(\mathcal{X})$.

Q.E.D.

REFERENCE

[1] KATONA, G. and TUSNÁDY, G.: The principle of conservation of entropy in a noiseless channel, *Studia Sci. Math. Hungar.* **2** (1967) 29—35.

INSTITUT FOR MATEMATISK STATISTIK, KØBENHAVN

(Received December 20, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 403—414.

A NEW PROOF OF A. F. TIMAN'S APPROXIMATION THEOREM

by

G. FREUD and P. VÉRTESI

1. Introduction

In his paper [3] the first of us introduced a new type of interpolation process, which is the object of our present investigation.

Let

$$x_{k,n} = \cos\frac{2k-1}{2n}\pi \qquad (k = 1, 2, \ldots, n) \tag{1}$$

be the roots of the Čebyšev polynomial

$$l_{kn}(x) = \frac{(-1)^{k+1}\sqrt{1-x_{kn}^2}}{n} \cdot \frac{T_n(x)}{x - x_{kn}} \tag{2}$$

the fundamental polynomials of Lagrange interpolation based on the nodes x_{kn} and

$$h_{kn}(x) = \left[1 - \frac{x_{kn}}{1-x_{kn}^2}(x - x_{kn})\right] l_{kn}^2(x) \equiv v_{kn}(x) l_{kn}^2(x) \tag{3}$$

the fundamental polynomials of Hermite—Fejér interpolation.

Further, let

$$\psi_n(u, v) = \frac{2}{n}\sum_{r=1}^{n-1} T_r'(u) T_r(v). \tag{4}$$

Now, the fundamental polynomials of the interpolation introduced by the first of us in [3] are

$$\varphi_{kn}(x) = v_{kn}(x) l_{kn}^4(x) + 2(x - x_{kn}) l_{kn}^3(x) \psi_n(x_{kn}, x). \tag{5}$$

His result is as follows:

Let $f(x)$ be an arbitrary continuous function in $[-\frac{1}{2}, \frac{1}{2}]$ and let us extend this function by the convention $f(x) = f(+\frac{1}{2})$ for $x > \frac{1}{2}$ resp. $f(x) = f(-\frac{1}{2})$ for $x < -\frac{1}{2}$ to $[-1, +1]$. Finally let us construct the polynomial

$$J_n(f; x) = f(0) + \sum_{k=1}^{n} \varphi_{kn}(x)[f(x_{kn}) - f(0)] \tag{6}$$

of degree $4n - 3$ at most. It was shown in [3], that we have with an absolute constant C

$$|f(x) - J_n(f; x)| \leqq C\omega\left(f; \frac{1}{4n}\right), \qquad -\frac{1}{2} \leqq x \leqq \frac{1}{2} \tag{7}$$

Here

$$\omega(f;\delta)=\max_{\substack{|h|\leqq\delta\\ |x|\leqq\frac{1}{2}\\ |x+h|\leqq\frac{1}{2}}}|f(x+h)-f(x)|$$

denotes — as usual — the continuity modulus of $f(x)$.

The proof of (7) is straightforward, so that it furnishes a new interpolatory approach to Jackson's approximation theorem. The method was later extended by M. SALLAY [4] to the case, when the nodes are the roots of Legendre polynomials (resp. orthogonal polynomials which are very near to Legendre polynomials) and by R. B. SAXENA [5] to the case, when the nodes are the zeros of the Čebyšev polynomials of the second kind $U_n(x)$. A very essential improvement of R. B. SAXENA was that he — using the corresponding expression built on the zeros of $U_n(x)$ — he replaced (6) by the formula

$$(8)\qquad J_n^*(f;x)=\frac{1+x}{2}f(1)+\frac{1-x}{2}f(-1)+$$
$$+\sum_{k=1}^{n}\left\{f(x_{kn})-\left[\frac{1+x}{2}f(1)+\frac{1-x}{2}f(-1)\right]\right\}\varphi_{kn}(x).$$

Using the zeros of $U_n(x)$ as nodes of interpolation, R. B. SAXENA proved that this polynomials approximate even on the whole interval of the interpolation $[-1, +1]$ every continuous function with an error $C\omega\left(f;\frac{1}{4n}\right)$ at most. The similar problem with the roots of $T_n(x)$ as nodes of interpolation was treated in [7] by the second of us and he was able to prove — as an improvement to [3] — that

$$|f(x)-J_n^*(f;x)|\leqq C\omega\left(f;\frac{1}{4n}\right)\quad\text{for}\quad x\in[-1,+1].$$

In the present paper we make the further improvement

$$(9)\qquad |f(x)-J_n^*(f;x)|\leqq C\left[\omega\left(f;\frac{\sqrt{1-x^2}}{4n}\right)+\omega\left(f;\frac{1}{(4n)^2}\right)\right].$$

As a consequence of (8) our polynomials $J_n^*(f;x)$ are of degree $4n-2$ at most and they take the same values as $f(x)$ at the points x_{kn} $(k=1, 2, \ldots, n)$.

The first construction of a sequence of polynomials for which the error of approximation is estimated by the right side of (9) is due to A. F. TIMAN [6]; it is his merit to have found the precise form of the error term of polynomial approximation, valid even in the neighbourhood of ± 1. It was shown later by V. K. DZIADIK [1] — in the way of proving the reverse theorem for the classes Lip α $(\alpha<1)$ — that this result of A. F. TIMAN is — up to the value of the constant C — best-possible.

By showing (9) we give a new proof of TIMAN's theorem. It seems to be of considerable advantage that for a fixed n our polynomials $J_n^*(x)$ depend on a finite set of values of $f(x)$ only. Expressions of this kind can be calculated much easier than those which include integrals.

2. Preliminary Estimations

For sake of brevity let us fix n and $f(x)$ and let us denote x_{kn} by x_k, $l_{kn}(x)$ by $l_k(x)$ etc. and we denote $\omega(f;\delta)$ by $\omega(\delta)$. We put further $x_0=1$, $x_{n+1}=-1$, $x=\cos\vartheta$, $x_k=\cos\vartheta_k$, so that

$$\vartheta_0=0, \quad \vartheta_k=\frac{2k-1}{2n}\pi \qquad (k=1,2,\ldots,n), \quad \vartheta_{n+1}=\pi.$$

By „c" we denote in the sequal positive absolute constants, which need not be the same.

We recall the formula

$$\sum_{k=1}^{n}\varphi_k(x)=\left\{\frac{1}{n}\left[1+2\sum_{r=1}^{n-1}\cos^2 r\vartheta\right]\right\}^2$$

(see (8) and (9) on page 229 of [3]) and conclude that

$$\left|\sum_{k=1}^{n}\varphi_k(x)-1\right|\leqq 3 \qquad (-1\leqq x\leqq 1) \tag{10}$$

and

$$\sin\vartheta\left|\sum_{k=1}^{n}\varphi_k(\cos\vartheta)-1\right|\leqq\frac{3}{n} \qquad (0\leqq\vartheta\leqq\pi) \tag{11}$$

hold (See [7], formulas (2.1) and (2.2)). Further we recall the elementary inequalities

$$\frac{\sin\vartheta_k}{|\cos\vartheta-\cos\vartheta_k|}\leqq\frac{1}{\sin\dfrac{|\vartheta-\vartheta_k|}{2}} \qquad (0\leqq\vartheta\leqq\pi,\ k=1,2,\ldots,n) \tag{12}$$

(see (3. 5) in [7]).

From (3. 14) in [7] we get

$$|\psi_n(x_k,x)|\leqq\frac{c}{\sin\vartheta_k}\cdot\frac{1}{\sin\dfrac{|\vartheta-\vartheta_k|}{2}}. \tag{13}$$

Besides (13) we need a more precise estimation of $|\psi_n(x_k,x)|$:

LEMMA I. We have

$$|\psi_n(x_k,x)|\leqq\frac{c}{\sin\vartheta_k}\left(\frac{1}{n}\frac{1}{\sin^2\dfrac{\vartheta-\vartheta_k}{2}}+1\right)+\frac{c\sin\vartheta}{\sin\vartheta_k}\frac{1}{|\cos\vartheta-\cos\vartheta_k|}. \tag{14}$$

PROOF. From (4) we obtain (see (3.14) in [7])

$$\psi_n(x_k, x) = \frac{1}{n \sin \vartheta_k} \left\{ \left[\sin(2n-1) \frac{\vartheta_k - \vartheta}{2} \cos \frac{\vartheta_k - \vartheta}{2} - \right.\right.$$

$$\left. - (2n-1) \cos (2n-1) \frac{\vartheta_k - \vartheta}{2} \sin \frac{\vartheta_k - \vartheta}{2} \right] \frac{1}{4 \sin^2 \frac{\vartheta_k - \vartheta}{2}} +$$

$$+ \left[\sin (2n-1) \frac{\vartheta_k + \vartheta}{2} \cos \frac{\vartheta_k + \vartheta}{2} - (2n-1) \cos (2n-1) \frac{\vartheta_k + \vartheta}{2} \sin \frac{\vartheta_k + \vartheta}{2} \right] \cdot$$

$$\left. \cdot \frac{1}{4 \sin^2 \frac{\vartheta_k + \vartheta}{2}} \right\}.$$

Let us denote

$$\psi_n^{(1)}(x_k, x) = \frac{1}{n \sin \vartheta_k} \left[\sin (2n-1) \frac{\vartheta_k - \vartheta}{2} \cos \frac{\vartheta_k - \vartheta}{2} \cdot \frac{1}{4 \sin^2 \frac{\vartheta_k - \vartheta}{2}} + \right.$$

$$\left. + \sin (2n-1) \frac{\vartheta_k + \vartheta}{2} \cos \frac{\vartheta_k + \vartheta}{2} \cdot \frac{1}{4 \sin^2 \frac{\vartheta_k + \vartheta}{2}} \right],$$

$$\psi_n^{(2)}(x_k, x) = -\frac{1}{n \sin \vartheta_k} \left[(2n-1) \cos (2n-1) \frac{\vartheta_k - \vartheta}{2} \sin \frac{\vartheta_k - \vartheta}{2} \cdot \frac{1}{4 \sin^2 \frac{\vartheta_k - \vartheta}{2}} + \right.$$

$$\left. + (2n-1) \cos (2n-1) \frac{\vartheta_k + \vartheta}{2} \sin \frac{\vartheta_k + \vartheta}{2} \frac{1}{4 \sin^2 \frac{\vartheta_k + \vartheta}{2}} \right].$$

Obviously

$$\psi_n(x_k, x) = \psi_n^{(1)}(x_k, x) + \psi_n^{(2)}(x_k, x). \tag{15}$$

At first we estimate $\psi_n^{(1)}(x_k, x)$. Taking into account that

$$\frac{1}{\sin \frac{\vartheta_k + \vartheta}{2}} \leqq \frac{1}{\sin \frac{|\vartheta_k - \vartheta|}{2}} \qquad (0 \leqq \vartheta \leqq \pi, \ 0 \leqq \vartheta_k \leqq \pi),$$

$$\left| \sin (2n-1) \frac{\vartheta_k \mp \vartheta}{2} \cos \frac{\vartheta_k \mp \vartheta}{2} \right| \leqq 1,$$

we get

$$|\psi_n^{(1)}(x_k, x)| \leqq \frac{c}{n \sin \vartheta_k} \cdot \frac{1}{\sin^2 \frac{\vartheta_k - \vartheta}{2}}. \tag{16}$$

In order to estimate $\psi_n^{(2)}(x_k, x)$ we get

$$\cos(2n-1)\frac{\vartheta_k \pm \vartheta}{2} = \cos n(\vartheta_k \pm \vartheta)\cos\frac{\vartheta_k \pm \vartheta}{2} + \sin n(\vartheta_k \pm \vartheta)\sin\frac{\vartheta_k \pm \vartheta}{2}.$$

Using the notations

$$\psi_n^*(x_k, x) = -\frac{1}{n\sin\vartheta_k}\left[(2n-1)\sin n(\vartheta_k-\vartheta)\sin^2\frac{\vartheta_k-\vartheta}{2}\cdot\frac{1}{4\sin^2\dfrac{\vartheta_k-\vartheta}{2}} + \right.$$

$$\left. + (2n-1)\sin n(\vartheta_k+\vartheta)\sin^2\frac{\vartheta_k+\vartheta}{2}\cdot\frac{1}{4\sin^2\dfrac{\vartheta_k+\vartheta}{2}}\right],$$

$$\psi_n^{**}(x_k, x) = -\frac{1}{n\sin\vartheta_k}\left[(2n-1)\cos n(\vartheta_k-\vartheta)\cos\frac{\vartheta_k-\vartheta}{2}\frac{1}{4\sin\dfrac{\vartheta_k-\vartheta}{2}} + \right.$$

$$\left. + (2n-1)\cos n(\vartheta_k+\vartheta)\cos\frac{\vartheta_k+\vartheta}{2}\frac{1}{4\sin\dfrac{\vartheta_k-\vartheta}{2}}\right],$$

we can write

$$\psi_n^{(2)}(x_k, x) = \psi_n^*(x_k, x) + \psi_n^{**}(x_k, x).$$

Obviously

$$|\psi_n^*(x_k, x)| \leqq \frac{c}{n\sin\vartheta_k}. \tag{17}$$

Now we turn to critical part of $\psi_n(x_k, x)$, $\psi_n^{**}(x_k, x)$. The splitting of $\psi_n^{**}(x_k, x)$ and its estimation belong to the essential steps of our proof.

We remind that

$$\cos(n\vartheta_k \mp n\vartheta) = \cos\left[(2k-1)\frac{\pi}{2} \mp n\vartheta\right] = \pm(-1)^{k+1}\sin n\vartheta$$

so from the formula of $\psi_n^{**}(x_k, x)$ we obtain

$$|\psi_n^{**}(x_k, x)| \leqq \frac{c\sin\vartheta}{\sin\vartheta_k}\frac{1}{|\cos\vartheta - \cos\vartheta_k|}^{1} \tag{18}$$

Considering (15), (16), (17) and (18), we can see that (14) is true. —

[1] In detail:

$$\left|\frac{\cos\dfrac{\vartheta_k+\vartheta}{2}}{\sin\dfrac{\vartheta_k+\vartheta}{2}} - \frac{\cos\dfrac{\vartheta_k-\vartheta}{2}}{\sin\dfrac{\vartheta_k-\vartheta}{2}}\right| = \left|\operatorname{ctg}\frac{\vartheta_k+\vartheta}{2} - \operatorname{ctg}\frac{\vartheta_k-\vartheta}{2}\right| = \frac{\sin\vartheta}{\left|\sin\dfrac{\vartheta_k+\vartheta}{2} - \sin\dfrac{\vartheta_k-\vartheta}{2}\right|} =$$

$$= \frac{1}{2}\frac{\sin\vartheta}{|\cos\vartheta - \cos\vartheta_k|}.$$

We quote further from the classical paper of L. FEJÉR [2] the inequality

$$|l_k(x)| \leqq \sqrt{2} \qquad (-1 \leqq x \leqq 1;\ k = 1, 2, \ldots, n) \tag{19}$$

At least we remind the well-known inequalities

$$\omega(\lambda\delta) \leqq 2\lambda\omega(\delta) \quad \text{if} \quad \lambda \geqq 1 \tag{20}$$

$$\omega(\lambda\delta) \leqq (\lambda+1)\omega(\delta) \quad \text{if} \quad \lambda \geqq 0. \tag{21}$$

3. Estimation of $J_n^*(f; x) - f(x)$

We consider the expression

$$\sum_{k=1}^{n} |\varphi_k(x)|\,|f(x)-f(x_k)| = \sum_{k=1}^{n} |v_k(x) l_k^4(x) + 2(x-x_k) l_k^3(x)\psi_n(x_k, x)| \cdot \cdot |f(x)-f(x_k)| \tag{22}$$

Our first aim is to prove

$$\sum_{k=1}^{n} |\varphi_k(x)|\,|f(x)-f(x_k)| \leqq c\left[\omega\left(\frac{\sqrt{1-x}}{n}\right)+\omega\left(\frac{1}{n^2}\right)\right] \qquad (0 \leqq x \leqq 1). \tag{23}$$

Let us now suppose

$$0 \leqq \vartheta \leqq \frac{\pi}{2} \quad \text{and} \quad \vartheta_i \leqq \vartheta < \vartheta_{i+1} \tag{24}$$

Using the inequality

$$|f(\cos\vartheta)-f(\cos\vartheta_k)| \leqq \omega(|\cos\vartheta-\cos\vartheta_k|) \leqq \omega\left[2\sin\frac{|\vartheta-\vartheta_k|}{2}\left(\sin\frac{\vartheta}{2}+\sin\frac{\vartheta_k}{2}\right)\right],$$

we obtain

$$|f(x)-f(x_k)| \leqq c\omega\left(\sin\frac{|\vartheta-\vartheta_2|}{2}\cdot\sin\frac{\vartheta}{2}\right) \quad \text{if} \quad \vartheta_k \leqq 2\vartheta \tag{25}$$

$$|f(x)-f(x_k)| \leqq c\omega\left(\sin\frac{|\vartheta-\vartheta_k|}{2}\cdot\sin\frac{\vartheta_k}{2}\right) \quad \text{if} \quad \vartheta_k > 2\vartheta \tag{26}$$

This dissection of which we introduce in the estimation of (22) is — besides the estimation of ψ_n^{**} — the second essential feature of our calculation.

In order to prove (23) we insert some lemmas.

Let us denote

$$\varphi_k^*(x) = l_k^4(x) + 2(x-x_k) l_k^3(x)\psi_n(x_k, x) \tag{27}$$

$$\varphi_k^{**}(x) = \frac{x_k}{1-x_k^2}(x_k - x) l_k^4(x), \tag{28}$$

so that

$$\varphi_k(x) = \varphi_k^*(x) + \varphi_k^{**}(x). \tag{29}$$

LEMMA II. If $0 \leqq \vartheta \leqq \frac{\pi}{2}$, $\vartheta_i \leqq \vartheta < \vartheta_{i+1}$ then

$$|\varphi_k^*(x)|\,|f(x)-f(x_k)| \leqq c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)\frac{1}{n^2}\frac{1}{|\vartheta-\vartheta_k|^2} \quad \text{if} \quad \vartheta_k < 2\vartheta,\ k \neq i,\ i+1 \tag{30}$$

$$|\varphi_k^*(x)|\,|f(x)-f(x_k)| \leqq c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right) \quad \text{if} \quad k = i,\ i+1 \tag{31}$$

$$|\varphi_k^*(x)|\,|f(x)-f(x_k)| \leqq c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)\frac{1}{n^2}\frac{1}{|\vartheta-\vartheta_k|^2} + \tag{32}$$

$$+ c\omega\left(\frac{1}{n^2}\right)\frac{1}{n}\frac{\vartheta_k}{|\vartheta-\vartheta_k|} \quad \text{if} \quad \vartheta_k > 2\vartheta,\ k > i+1.$$

PROOF. We have obviously

$$|\vartheta-\vartheta_k| > \frac{c}{n} \quad \text{or} \quad |\vartheta-\vartheta_k|n > c \quad \text{if} \quad 1 \leqq k < i, \quad i+1 < k \leqq n \tag{33}$$

If $\vartheta_k \leqq 2\vartheta$, $k \neq i, i+1$ then using (2), (12), (13), (20), (25), and (33) we obtain

$$|\varphi_k^*(x)|\,|f(x)-f(x_k)| = |l_k^4(x)+2(x-x_k)l_k^3(x)\psi_n(x_k,x)|\,|f(x)-f(x_k)| \leqq$$

$$\leqq c\left|\frac{\sin^4\vartheta_k\cos^4 n\vartheta}{n^4(\cos\vartheta-\cos\vartheta_k)^4}+2|\cos\vartheta-\cos\vartheta_k|\frac{\sin^3\vartheta_k|\cos^3 n\vartheta|}{n^3|\cos\vartheta-\cos\vartheta_k|^3}\cdot\frac{1}{\sin\vartheta_k}\frac{1}{\sin\frac{|\vartheta-\vartheta_k|}{2}}\right| \cdot$$

$$\cdot\,\omega\left(\sin\frac{|\vartheta-\vartheta_k|}{2}\sin\frac{\vartheta}{2}\right) \leqq c\left[\frac{\sin^4\vartheta_k}{n^4(\cos\vartheta-\cos\vartheta_k)^4}+\frac{\sin^2\vartheta_k}{n^3(\cos\vartheta-\cos\vartheta_k)^2}\cdot\frac{1}{\sin\frac{|\vartheta-\vartheta_k|}{2}}\right] \cdot$$

$$\cdot\, n\sin\frac{|\vartheta-\vartheta_k|}{2}\,\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right) \leqq$$

$$\leqq c\left[\frac{1}{n^3\sin^3\frac{|\vartheta-\vartheta_k|}{2}}+\frac{1}{n^2\sin^2\frac{\vartheta-\vartheta_k}{2}}\right]\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right) \leqq$$

$$\leqq c\left[\frac{1}{n^3|\vartheta-\vartheta_k|^3}+\frac{1}{n^2|\vartheta-\vartheta_k|^2}\right]\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right) \leqq$$

$$\leqq c\,\frac{1}{n^2}\,\frac{1}{|\vartheta-\vartheta_k|^2}\,\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)$$

If $k=i$ (or $i+1$) using (2), (12), (13), (20), (25) and L. FEJÉR's inequality (19), we get

$$|\varphi_i^*(x)|\,|f(x)-f(x_i)| \leqq c|l_i^4(x)+2(x-x_i)l_i^3(x)\psi_n(x_i,x)|\,\omega\left(\sin\frac{|\vartheta-\vartheta_i|}{2}\sin\frac{\vartheta}{2}\right)\leqq$$

$$\leqq c\left[\frac{\sin^4\vartheta_i\cos^4 n\vartheta}{n^4(\cos\vartheta-\cos\vartheta_i)^4}+2\frac{\sin^2\vartheta_i\cos^3 n\vartheta}{n^3(\cos\vartheta-\cos\vartheta_i)^2}\frac{1}{\sin\dfrac{|\vartheta-\vartheta_i|}{2}}\right]n\sin\frac{|\vartheta-\vartheta_i|}{2}\cdot$$

$$\cdot\,\omega\left(\frac{\sin\dfrac{\vartheta}{2}}{n}\right)\leqq c[|l_i^3(x)|+l_i^2(x)]\,\omega\left[\frac{\sin\dfrac{\vartheta}{2}}{n}\right]\leqq c\omega\left(\frac{\sin\dfrac{\vartheta}{2}}{n}\right).$$

Let us turn now to the case $\vartheta_k>2\vartheta$, $k>i+1$.

By the help of (14), (26) and the previously used formulas

$$|\varphi_k^*(x)|\,|f(x)-f(x_k)| \leqq c\left|l_k^4(x)+2|x-x_2|\,|l_k^3(x)|\left[\frac{c}{\sin\vartheta_k}\left(\frac{1}{n}\frac{1}{\sin^2\dfrac{\vartheta_k-\vartheta}{2}}+1\right)\right]\right|\cdot$$

$$\cdot\,\omega\left(\sin\frac{|\vartheta_k-\vartheta|}{2}\sin\frac{\vartheta_k}{2}\right)+c|x-x_k|\left|l_k^3(x)\frac{\sin\vartheta}{\sin\vartheta_k}\frac{1}{|\cos\vartheta-\cos\vartheta_k|}\right|\omega(|\cos\vartheta-\cos\vartheta_k|)\leqq$$

$$\leqq c\left[\frac{\sin^4\vartheta_k}{n^4(\cos\vartheta-\cos\vartheta_k)^4}+\frac{\sin^3\vartheta_k}{n^3|\cos\vartheta-\cos\vartheta_k|^2}\cdot\frac{c}{\sin\vartheta_k}\left(\frac{1}{n}\frac{1}{\sin^2\dfrac{\vartheta_k-\vartheta}{2}}+1\right)\right]\cdot$$

$$\cdot\,n^2\sin\frac{|\vartheta_k-\vartheta|}{2}\sin\frac{\vartheta_k}{2}\,\omega\left(\frac{1}{n^2}\right)+\frac{c}{n^3}\frac{\sin^3\vartheta_k}{(\cos\vartheta-\cos\vartheta_k)^2}\frac{\sin\vartheta}{\sin\vartheta_k}\frac{1}{|\cos\vartheta-\cos\vartheta_k|}\cdot$$

$$\cdot\frac{n}{\sin\vartheta}|\cos\vartheta-\cos\vartheta_k|\,\omega\left(\frac{\sin\vartheta}{n}\right)\leqq c\omega\left(\frac{1}{n^2}\right)\cdot$$

$$\cdot\left[\frac{1}{n^2}\frac{\vartheta_k}{|\vartheta-\vartheta_k|^3}+\frac{1}{n^2}\frac{\vartheta_k}{|\vartheta-\vartheta_k|^3}+\frac{1}{n}\frac{\vartheta_k}{|\vartheta-\vartheta_k|}\right]+c\omega\left(\frac{\sin\vartheta}{n}\right)\frac{1}{n^2}\frac{1}{(\vartheta-\vartheta_k)^2}\leqq$$

$$\leqq c\omega\left(\frac{\sin\dfrac{\vartheta}{2}}{n}\right)\frac{1}{n^2}\frac{1}{|\vartheta-\vartheta_k|^2}+c\omega\left(\frac{1}{n^2}\right)\frac{1}{n}\frac{\vartheta_k}{|\vartheta-\vartheta_k|}.$$ [2]

So the proof of Lemma II is ended. —

[2] Here we used

$$\omega\left(\frac{\sin\vartheta}{n}\right)=\omega\left(\frac{2\sin\dfrac{\vartheta}{2}\cos\dfrac{\vartheta}{2}}{n}\right)\leqq c\omega\left(\frac{\sin\dfrac{\vartheta}{2}}{n}\right).$$

LEMMA III. If $0 \leqq \vartheta \leqq \frac{\pi}{2}$ then

$$\sum_{k=1}^{n} |\varphi_k^*(x)|\,|f(x)-f(x_k)| \leqq c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)+c\omega\left(\frac{1}{n^2}\right). \tag{34}$$

PROOF. Considering (30), (31) and (32) we can write

$$\sum_{k=1}^{n} |\varphi_k^*(x)|\,|f(x)-f(x_k)| \leqq c \sum_{\substack{\vartheta_k \leqq 2\vartheta \\ k \neq i, i+1}} \omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)\frac{1}{n^2}\,\frac{1}{|\vartheta-\vartheta_k|^2}+c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)+$$

$$+c\sum_{\substack{\vartheta_k > 2\vartheta \\ k > i+1}}\left[\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)\frac{1}{n^2}\,\frac{1}{|\vartheta-\vartheta_k|^2}+\omega\left(\frac{1}{n^2}\right)\frac{1}{n}\,\frac{\vartheta_k}{|\vartheta-\vartheta_k|}\right].$$

But if $1 \leqq k < i,\ i+1 < k \leqq n$

$$\frac{1}{|\vartheta-\vartheta_k|} = O\left(\frac{n}{|i-k|}\right) \qquad (\vartheta_k \leqq 2\vartheta \quad \text{or} \quad \vartheta_k > 2\vartheta,\ k \neq i,\ i+1), \tag{35}$$

so we get

$$\sum_{k=1}^{n} |\varphi_k^*(x)|\,|f(x)-f(x_k)| \leqq c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)\sum_{\substack{\vartheta_k \leqq 2\vartheta \\ k \neq i, i+1}}\frac{1}{n^2}\,\frac{n^2}{(i-k)^2}+c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)+$$

$$+c\omega\left(\frac{1}{n^2}\right)\sum_{\substack{\vartheta_k > 2\vartheta \\ k > i+1}}\frac{\frac{k}{n}}{\frac{|i-k|}{n}}\cdot\frac{1}{n}=c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)\sum_{\substack{\vartheta_k \leqq 2\vartheta \\ k \neq i, i+1}}\frac{1}{(i-k)^2}+c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)+$$

$$+c\omega\left(\frac{1}{n^2}\right)\sum_{\substack{\vartheta_k > 2\vartheta \\ k > i+1}}\frac{k}{|i-k|}\,\frac{1}{n} \leqq c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)+c\omega\left(\frac{1}{n^2}\right),$$

as we asserted.

After investigation of $\varphi_k^*(x)$ we turn to $\varphi_k^{**}(x)$ (see (28)).

LEMMA IV. In the whole interval $0 \leqq \vartheta \leqq \pi$

$$\sum_{k=1}^{n} |\varphi_k^{**}(x)|\,|f(x)-f(x_k)| \leqq c\omega\left(\frac{1}{n^2}\right). \tag{36}$$

PROOF. Like previously, we can write

$$\sum_{k=1}^{n} |\varphi_k^{**}(x)|\,|f(x)-f(x_k)| = \sum_{k=1}^{n} l_k^4(x)\left|\frac{x_k}{1-x_k^2}(x-x_k)\right||f(x)-f(x_k)| \leqq$$

$$\leqq c\sum_{k=1}^{n} l_k^4(x)\frac{|x-x_k|}{1-x_k^2}\,\omega(|x-x_k|) \leqq c\sum_{k=1}^{n} l_k^4(x)\frac{(x-x_k)^2}{1-x_k^2}\,n^2\,\omega\left(\frac{1}{n^2}\right) \leqq$$

$$\leqq c\sum_{k=1}^{n}\frac{\sin^4\vartheta_k\cos^4 n\vartheta}{n^4(\cos\vartheta-\cos\vartheta_k)^4}\,\frac{(\cos\vartheta-\cos\vartheta_k)^2}{\sin^2\vartheta_k}\,n^2\,\omega\left(\frac{1}{n^2}\right) \leqq$$

$$\leqq c\sum_{\substack{k=1\\ k\neq i,i+1}}^{n}\frac{\sin^2\vartheta_k}{n^2(\cos\vartheta-\cos\vartheta_k)^2}\,\omega\left(\frac{1}{n^2}\right)+c\sum_{k=i}^{i+1}\frac{\sin^2\vartheta_k\cos^2 n\vartheta}{n^2(\cos\vartheta-\cos\vartheta_k)^2}\,\omega\left(\frac{1}{n^2}\right) \leqq$$

$$\leqq c\omega\left(\frac{1}{n^2}\right)\sum_{\substack{k=1\\ k\neq i,i+1}}^{n}\frac{1}{(\vartheta-\vartheta_k)^2}\,\frac{1}{n^2}+c\omega\left(\frac{1}{n^2}\right)[l_i^2(x)+l_{i+1}^2(x)] \leqq c\omega\left(\frac{1}{n^2}\right)$$

as it was stated.

From the former Lemmas we conclude the following ones:

LEMMA V.

$$|f(x)-J_n^*(f;x)| \leqq c\left[\omega\left(\frac{\sqrt{1-x}}{n}\right)+\omega\left(\frac{1}{n^2}\right)\right] \qquad 0\leqq x\leqq 1. \tag{37}$$

PROOF. By the aid of (29), (34) and (36) we can write

$$\sum_{k=1}^{n}|\varphi_k(x)|\,|f(x)-f(x_k)| \leqq c\left[\omega\left(\frac{\sqrt{1-x}}{n}\right)+\omega\left(\frac{1}{n^2}\right)\right] \qquad 0\leqq x\leqq 1. \tag{38}$$

In [5] we obtained

$$|f(x)-J_n^*(x)| = \left|\frac{1+x}{2}[f(x)-f(1)]\left[1-\sum_{k=1}^{n}\varphi_k(x)\right]+\right.$$
$$\left.+\frac{1-x}{2}[f(x)-f(-1)]\left[1-\sum_{k=1}^{n}\varphi_k(x)+\sum_{k=1}^{n}[f(x)-f(x_k)]\varphi_k(x)\right]\right|. \tag{39}$$

Now we shall estimate the remaining parts of the sum (39). Using (10), (11) and (21) we get

$$\left|\frac{1+x}{2}[f(x)-f(1)]\left[1-\sum_{k=1}^{n}\varphi_k(x)\right]\right| \leqq \frac{1+x}{2}\,\omega(|x-1|)\left|1-\sum_{k=1}^{n}\varphi_k(x)\right| \leqq$$

$$\leqq \frac{1+x}{2}\left[1+n|x-1|\,\frac{1}{\sin\vartheta}\right]\omega\left(\frac{\sin\vartheta}{n}\right)\left|1-\sum_{k=1}^{n}\varphi_k(x)\right| \leqq$$

$$\leqq \omega\left(\frac{\sin\vartheta}{n}\right)\left|1-\sum_{k=1}^{n}\varphi_k(x)\right|+cn\,\frac{(1+x)(1-x)}{\sin\vartheta}\,\omega\left(\frac{\sin\vartheta}{n}\right)\cdot\left|1-\sum_{k=1}^{n}\varphi_k(x)\right| \leqq$$

$$\leqq 3\omega\left(\frac{\sin\vartheta}{n}\right)+cn\sin\vartheta\left|1-\sum_{k=1}^{n}\varphi_k(x)\right|\omega\left(\frac{\sin\vartheta}{n}\right) \leqq c\omega\left(\frac{\sin\vartheta}{n}\right) \leqq c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)$$

Similarly we have

$$\left|\frac{1-x}{2}[f(x)-f(-1)]\left[1-\sum_{k=1}^{n}\varphi_k(x)\right]\right| \leqq c\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)$$

Using these and (38), (39) we get

$$|f(x)-J_n^*(x)| \leqq c\left[\omega\left(\frac{\sin\frac{\vartheta}{2}}{n}\right)+\omega\left(\frac{1}{n^2}\right)\right], \qquad 0 \leqq x \leqq 1.$$

But

$$\sin\frac{\vartheta}{2} = \sqrt{\frac{1-\cos\vartheta}{2}} = \sqrt{\frac{1-x}{2}}$$

so the formula (37) is proved.

4. Proof of the Main Theorem

THEOREM. *We have on the whole interval* $-1 \leqq x \leqq 1$

$$|f(x)-J_n^*(x)| \leqq c\left[\omega\left(\frac{\sqrt{1-x^2}}{4n}\right)+\omega\left(\frac{1}{(4n)^2}\right)\right]. \tag{40}$$

Using the fact that the zeros of $T_n(x)$ are located symmetrically to $x=0$, we get

$$J_n^*[f(-t);\, x] = J_n^*[f(t);\, -x]. \tag{41}$$

From Lemma V we conclude

$$|J_n^*[f(-t);\, x]-f(x)| \leqq c\left[\omega\left(\frac{\sqrt{1-x}}{n}\right)+\omega\left(\frac{1}{n^2}\right)\right] \qquad 0 \leqq x \leqq 1.$$

We use (41) and replace x by $-x$:

$$|J_n^*[f(t);\, x]-f(x)| \leqq c\left[\omega\left(\frac{\sqrt{1+x}}{n}\right)+\omega\left(\frac{1}{n^2}\right)\right] \qquad -1 \leqq x \leqq 0 \tag{42}$$

Combining (37), (42) and (20) we see that (40) is satisfied, as it was stated.

REFERENCES

[1] Дзядык, В. К.: О конструктивной характеристике функций удовлетворяющих условию Lip α $(0<\alpha<1)$ на конечнам отрезке вещественный оси, *Изв. Акад. Наук СССР Сер. Мат.* **20** (1956) 623—642.

[2] FEJÉR, L.: Lagrangesche Interpolation und die zugehöriges konjugierten Punkte, *Math. Ann.* **106** (1932)

[3] FREUD, G.: Über ein Jacksonsches Interpolationsverfahren, *On Approximation Theory,* Birkhäuser Verlag, Basel und Stuttgart, 1964; 227—232.

[4] SALLAY, M.: On an interpolation process, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **9** (1964) 607—615.

[5] SAXENA, R. B.: On a polynomial of interpolation, *Studia Sci. Math. Hungar.* **2** (1967) 167—183.

[6] Тиман, А. Ф.: Усиление теоремы Джексона о наилущим приближении непрерывных функций многочленами на конечном отрезке вещественной оси, *Докл. Акад. Наук СССР* **78** (1951) 17—20.

[7] VÉRTESI, P.: Jackson tételének bizonyítása interpolációs úton, Mat. Lapok **18** (1967)

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST
EÖTVÖS L. UNIVERSITY, BUDAPEST

(Received December 27, 1966.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 415—418.

ON TWO PROBLEMS ON EXCHANGEABLE EVENTS

by

C. J. RIDLER-ROWE

The two problems considered here were given by Professor D. G. KENDALL in his paper [2] "On finite and infinite sequences of exchangeable events". The author wishes to thank Professor D. G. KENDALL and Professor G. E. H. REUTER for their assistance.

The problems are now stated and the notation follows that of KENDALL [2]. For each $v=1, 2, \ldots$ let $(\Omega_v, \mathscr{A}_v, \mathrm{pr}_v)$ be a probability space carrying a finite number M_v of exchangeable events $\{A_r^{(v)}: r=1, 2, \ldots, M_v\}$, where $v \leqq M_v$. Let X_v be the number of the first v events $A_1^{(v)}, A_2^{(v)}, \ldots, A_v^{(v)}$ which occur and let

$$\alpha_j^{(v)} = \mathrm{pr}_v(A_{r_1}^{(v)} \cap A_{r_2}^{(v)} \cap \ldots \cap A_{r_j}^{(v)}),$$

where $r_1, r_2, \ldots, r_j$ are distinct. Then what conditions on the rate of growth of the sequence $\{M_v\}$ imply that

$$\lim_{v\to\infty} \mathrm{pr}_v(X_v = s) = \frac{\mu^s e^{-\mu}}{s!}, \qquad s = 0, 1, \ldots, \tag{1}$$

if

$$\lim_{v\to\infty} v^j \alpha_j^{(v)} = \mu^j, \qquad j = 1, 2, \ldots, J, \tag{2}$$

(where the case $J=2$ is of especial interest)? Also if $M_v \equiv v$, does (1) follow if $v\alpha_1^{(v)}$ and $v^2\alpha_2^{(v)}$ converge sufficiently rapidly to μ and μ^2 respectively as $v \to \infty$?

THEOREM. *If* (2) *is true for* $J=2$, *then* (1) *follows if*

$$\lim_{v\to\infty} \frac{M_v}{v} = \infty. \tag{3}$$

PROOF. Using equations (16) and (19) in KENDALL [2], the probability that exactly s of the first v events $A_1^{(v)}, \ldots, A_v^{(v)}$ occur, and exactly $k-s$ of the remaining events $A_{v+1}^{(v)}, \ldots, A_{M_v}^{(v)}$ occur, is

$$\binom{v}{s}\binom{M_v - v}{k-s} \delta^{M_v-k}\alpha_k^{(v)} = \tilde{\omega}_k^{(v)} g_{ks}^{(v)},$$

where $\delta\alpha_k^{(v)} = \alpha_k^{(v)} - \alpha_{k+1}^{(v)}$, $\tilde{\omega}_k^{(v)}$ is the probability that exactly k of the M_v events occur, and

$$g_{ks}^{(v)} = \binom{v}{s}\binom{M_v - v}{k-s} \Big/ \binom{M_v}{k}. \tag{4}$$

Hence on summing over k

$$\text{pr}_v(X_v = s) = \sum_{k=s}^{M_v - v + s} \tilde{\omega}_k^{(v)} g_{ks}^{(v)}, \qquad s = 0, 1, \ldots, . \tag{5}$$

Equations (20) in KENDALL [2] give

$$\sum_{k=0}^{M_v} (k)_j \tilde{\omega}_k^{(v)} = (M_v)_j \alpha_j^{(v)}, \qquad j = 0, 1, \ldots, M_v, \tag{6}$$

(where $(x)_j = x(x-1) \ldots (x-j+1)$ for any real number x and non-negative integer j). For each v let ϱ_v and σ_v^2 be the mean and variance respectively of the distribution $\{\tilde{\omega}_k^{(v)}\}$. Then from assumptions (2), with $J=2$, and (3)

$$\varrho_v = (1+\varepsilon_v)\lambda_v, \quad \sigma_v^2 = \eta_v \lambda_v^2, \tag{7}$$

where $\lambda_v = \dfrac{M_v \mu}{v}$ and, as $v \to \infty$, $\lambda_v \to \infty$, $\varepsilon_v \to 0$ and $\eta_v \to 0$. Let

$$Z_v = \{k : |k - \varrho_v| \leqq \eta_v^{1/3} \lambda_v\}. \tag{8}$$

Then from (7) using ČEBYŠEV's inequality,

$$\sum_{k \in Z_v}{}' \tilde{\omega}_k^{(v)} \to 1, \quad \text{as} \quad v \to \infty. \tag{9}$$

From (4), $g_{ks}^{(v)}$ is the probability that a random sample of size v, taken without replacement from a set consisting of k objects of type I and $M_v - k$ objects of type II, contains exactly s objects of type I. Hence

$$0 \leqq g_{ks}^{(v)} \leqq 1, \quad \text{for all } k, s \text{ and } v. \tag{10}$$

Also (e.g. see FELLER [1], Chapter VI, § 10, problem 36) it is known that

(11)

$$g_{ks}^{(v)} \to \frac{\mu^s e^{-\mu}}{s!}, \quad \text{if} \quad \frac{M_v}{v} \to \infty \quad \text{and} \quad k \sim \frac{M_v \mu}{v} \quad \text{as} \quad v \to \infty, \qquad (s = 0, 1, \ldots).$$

Let s be fixed. Then from (5) and (10)

$$\left| \text{pr}_v(X_v = s) - \frac{\mu^s e^{-\mu}}{s!} \right| \leqq \sum_{k \in Z_v} \tilde{\omega}_k^{(v)} d_v + \sum_{k \notin Z_v} \tilde{\omega}_k^{(v)}, \tag{12}$$

where

$$d_v = \max_{k \in Z_v} \left| g_{ks}^{(v)} - \frac{\mu^s e^{-\mu}}{s!} \right|.$$

But from (7), (8) and (11), $d_v \to 0$ as $v \to \infty$. The result then follows from (9) on letting v tend to infinity in (12).

If (2) is assumed only for $J=1$, then there is no condition on $\{M_v\}$ which implies (1). For if, for each $v \geqq \mu$, the distribution $\{\tilde{\omega}_k^{(v)}\}$ assigns masses $1 - \mu/v$ to $k=0$ and μ/v to $k = M_v$, then $v\alpha_1^{(v)} = \mu$ and, from (5), $\text{pr}_v(X_v = 0) \to 1$ as $v \to \infty$.

It is now shown that the theorem does not remain true if the condition on $\{M_v\}$ is made weaker (including the case $M_v \equiv v$), even if the conditions on $\{\alpha_j^{(v)}\}$ are made much stronger to include those of the second problem.

EXAMPLE. Let $\{M_\nu: \nu = 1, 2, \ldots\}$ be any sequence such that $\nu \leqq M < \infty$ and $\liminf_{\nu\to\infty} \frac{M_\nu}{\nu} < \infty$. Let J be any fixed positive integer. A sequence of exchangeable events is now constructed such that X_ν does not have a Poisson limit distribution though

$$\nu^j \alpha_j^{(\nu)} = \mu_j, \quad \text{for} \quad \nu \geqq \mu \quad \text{and} \quad 0 \leqq j \leqq M_\nu \quad \text{with} \quad j \neq J.$$

From section 4 of KENDALL [2] we know that, for each ν, a set $(\alpha_0^{(\nu)}, \alpha_1^{(\nu)}, \ldots, \alpha_{M_\nu}^{(\nu)})$ can be associated with a set of M_ν exchangeable events if $\{\tilde{\omega}_k^{(\nu)}\}$ is a proper probability distribution, where

$$\tilde{\omega}_k^{(\nu)} = \binom{M_\nu}{k} \delta^{M_\nu - k} \alpha_k^{(\nu)}, \qquad k = 0, 1, \ldots, M_\nu. \tag{13}$$

Let $\lambda = \liminf_{\nu\to\infty} \frac{M_\nu \mu}{\nu}$. Then $\lambda < \infty$. Let an infinite sequence S be chosen such that $\frac{M_\nu \mu}{\nu} \to \lambda$ as $\nu \to \infty$ with $\nu \in S$. If $\nu > J$ and

$$\alpha_j^{(\nu)} = \begin{cases} \left(\frac{\mu}{\nu}\right)^j, & \text{for} \quad 0 \leqq j \leqq M_\nu \quad \text{and} \quad j \neq J, \\ \left(\frac{\mu'}{\nu}\right)^J, & \text{for} \quad j = J, \end{cases} \tag{14}$$

where μ' is chosen later, then from (13), for $J < k \leqq M_\nu$

$$\tilde{\omega}_k^{(\nu)} = \binom{M_\nu}{k} \left(1 - \frac{\mu}{\nu}\right)^{M_\nu - k} \left(\frac{\mu}{\nu}\right)^k \to \frac{\lambda^k e^{-\lambda}}{k!}, \text{ as } \nu \to \infty \text{ with } \nu \in S, \tag{15}$$

whilst for $k \leqq J$

$$\tilde{\omega}_k^{(\nu)} = \binom{M_\nu}{k} \left\{ \left(1 - \frac{\mu}{\nu}\right)^{M_\nu - k} \left(\frac{\mu}{\nu}\right)^k + (-1)^{J-k} \binom{M_\nu - k}{J - k} \left[\left(\frac{\mu'}{\nu}\right)^J - \left(\frac{\mu}{\nu}\right)^J \right] \right\} \tag{16}$$

$$\to \frac{\lambda^k e^{-\lambda}}{k!} + O\left(\left(\frac{\mu'}{\mu}\right)^J - 1 \right), \quad \text{as} \quad \nu \to \infty \quad \text{with} \quad \nu \in S.$$

Let μ' be chosen such that $\mu' \neq \mu$ and all the limits in (16) are strictly positive. Then let ν_0 be chosen such that, if $\nu \in S$ and $\nu \geqq \nu_0$, then in (15) and (16) $\tilde{\omega}_k^{(\nu)} \geqq 0$ for $k = 0, 1, \ldots, M_\nu$. Also since $\alpha_0^{(\nu)} = 1$, it follows from equations (20) in KENDALL [2] that $\sum_{k=0}^{M_\nu} \tilde{\omega}_k^{(\nu)} = 1$. Hence if $\nu \in S$ and $\nu \geqq \nu_0$ then $\{\tilde{\omega}_k^{(\nu)}\}$ as defined by (15) and (16) is a proper probability distribution, and in this case we may let a set of exchangeable events be associated with (14). However if $\nu \geqq \mu$ and if $\nu < \nu_0$ or $\nu \notin S$, then clearly we may let a set of exchangeable events be associated with

$$\alpha_j^{(\nu)} = \left(\frac{\mu}{\nu}\right)^j, \qquad j = 0, 1, \ldots, M_\nu,$$

whilst the events may be defined arbitrarily for $\nu < \mu$.

Since $\frac{M_v\mu}{v} \to \lambda$ as $v \to \infty$ with $v \in S$, it follows from (4) that for $j \geqq s$

$$g_{js}^{(v)} \to \binom{j}{s}\left(1-\frac{\mu}{\lambda}\right)^{j-s}\left(\frac{\mu}{\lambda}\right)^{s}, \quad \text{as} \quad v \to \infty \quad \text{with} \quad v \in S, \tag{17}$$

$$\left(\text{putting } \left(1-\frac{\mu}{\lambda}\right)^{0} = 1 \quad \text{if} \quad \mu = \lambda\right).$$

LEMMA. If (i) $p^{(v)}$ is a sequence of probability distributions on the integers such that $\lim_{v\to\infty} p_j^{(v)} = p_j$ for each j and $\sum_j p_j = 1$, and (ii) $f_j^{(v)}$ is a sequence of functions, on the integers such that $\lim_{v\to\infty} f_j^{(v)} = f_j$ for each j and $f_j^{(v)}$ is bounded for all v and j, then

$$\lim_{v\to\infty} \sum_j p_j^{(v)} f_j^{(v)} = \sum_j p_j f_j.$$

From (5), (15), (17) and the lemma it follows that for $s > J$

$$\mathrm{pr}_v(X_v = s) \to \frac{\mu^s e^{-\mu}}{s!}, \quad \text{as} \quad v \to \infty \quad \text{with} \quad v \in S. \tag{18}$$

But since $\mu' \neq \mu$ it follows from (16) that $\tilde{\omega}_J^{(v)} \nrightarrow \frac{\lambda^J e^{-\lambda}}{J}$ as $v \to \infty$ with $v \in S$. Hence by a similar argument

$$\mathrm{pr}_v(X_v = J) \nrightarrow \frac{\mu^J e^{-\mu}}{J!}, \quad \text{as} \quad v \to \infty \quad \text{with} \quad v \in S.$$

Thus X_v has no Poisson limit distribution.

REMARK. An argument similar to that giving (18) shows that

$$\lim_{v\to\infty} \mathrm{pr}_v(X_v = s) = \frac{\mu^s e^{-\mu}}{s!}, \quad s = 0, 1, \ldots,$$

$$\text{if } \lim_{v\to\infty} \frac{M_v}{v} = \frac{\lambda}{\mu} \text{ and } \lim_{v\to\infty} \mathrm{pr}_v(Y_v = s) = \frac{\lambda^s e^{-\lambda}}{s!}, \quad s = 0, 1, \ldots,$$

where Y_v is the number of the M_v events which occur.

REFERENCES

[1] FELLER, W.: *Introduction to Probability Theory and its Applications*, Vol I, (2nd ed.) New York, 1957.

[2] KENDALL, D. G.: On finite and infinite sequences of exchangeable events, *Studia Sci. Math. Hungar.* **2** (1967).

DEPARTMENT OF MATHEMATICS, IMPERIAL COLLEGE, LONDON

(Received January 11, 1967.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 419—423.

A CONTRIBUTION TO THE PROBLEM OF RATIONAL APPROXIMATION OF REAL FUNCTIONS

by

G. FREUD

1. Statement of the Problem

Let $f(x)$ be a real continuous function defined on $x \in [-1, 1]$, and let

$$R_n(f) = \inf_{r_n} \max_{x \in [-1,+1]} |f(x) - r_n(x)|$$

where $r_n(x)$ runs through the rational functions, where both nominator and denominator are polynomials of degree n at most.

P. SZÜSZ and P. TURÁN [4] mentioned the following problem: Let $-1 \leqq \xi_0 < \xi_1 < \ldots < \xi_m \leqq 1$ be a finite pointset, let $f(x)$ belong to the class Lip α in any closed part of the open intervals (ξ_{k-1}, ξ_k) and let $f(x)$ satisfy a Lip β-condition $(\beta < \alpha)$ in the neighbourhood of the points ξ_k $(k = 0, 1, \ldots m)$. P. TURÁN conjectured that under conditions of this type $R_n(f) = O(n^{-\alpha})$ is valid. In the present paper we are proving a theorem of this kind.[1]

2. The Weak Localization Theorem

Our main tool is the following

LOCALIZATION THEOREM. Let there exist m sequences of polynomials $\{\pi_n^{(k)}(x); n = 0, 1, \ldots\}$ $(k = 1, 2, \ldots m)$ so that the degree of $\pi_n^{(k)}$ is at most n and we have

$$|f(x) - \pi_n^{(k)}(x)| \leqq \varepsilon_n \quad (x \in [\xi_{k-1}, \xi_k], k = 1, 2, \ldots m).$$

Under this hypotheses we have

$$R_n(f) \leqq c(f)\left(\varepsilon_{\nu_n} + e^{-\frac{1}{2}\sqrt{\nu_n}}\right); \quad \nu_n = 2\left[\frac{n}{4m}\right]$$

This theorem was proved in G. FREUD [1]. We call it the "weak localization theorem". The reason for this nomenclature is, that it was conjectured on the same place by the author and proved later by J. SZABADOS [3], that in this statement polynomial approximation can be replaced by rational approximation. This theorem of J. SZABADOS implies the weak localization theorem, and for this reason we call

[1] In our conversation on 20. January, 1967, where he mentioned his conjecture, Prof. TURÁN communicated to be in possession of a sketch of a proof, as a result of a collaboration with P. SZÜSZ, but he has no intention to work it out. As an answer to a question of mine, he mentioned that they did not realized the connection of the problem with the localization theorem.

it strong localization theorem. In this latter theorem an additional assumption on the order of the derivatives of the approximating functions occurs, but fortunately this causes no harm. We consider it as a very remarkable fact, that the weak localization theorem applies to TURÁN's problem.

3. The Function Class $C^*(\omega)$. Statement of the Theorem

Let $\eta_0 = -1$, $\eta_j = \frac{1}{2}(\xi_{j-1} + \xi_j)$ $(j = 1, 2, \dots m)$, $\eta_{m+1} = \xi_m$. Let further $\omega(\delta)$ be a continuity modulus, i.e. a nondecreasing function defined for each $\delta \geqq 0$ for which we have $\omega(2\delta) \leqq 2\omega(\delta)$. We say that the function $f(x)$ belongs to the class $C^*(\omega)$ if there exists a finite set $\{\xi_i\} \subset [-1, +1]$ (and corresponding points η_i) and a $K > 0$ so that we have for $x_1, x_2 \in [\eta_j, \xi_j]$ or $x_1, x_2 \in [\xi_j, \eta_{j+1}]$ $(j = 0, 1, \dots m)$.

$$|f(x_1) - f(x_2)| \leqq K\omega\left(\left|\sqrt{|x_1 - \xi_j|} - \sqrt{|x_2 - \xi_j|}\right|\right) \tag{1}$$

THEOREM. *Let* $\omega(\delta) \not\equiv 0$, $f \in C^*(\omega)$ *then*

$$R_n(f) \leqq c_1(f)\omega\left(\frac{1}{n}\right).$$

Before turning to the proof, we discuss the connection of this result with Turán's problem. Let $\omega_\alpha(\delta) = \delta^\alpha$ $(0 < \alpha \leqq 1)$ then each function of the class $C^*(\omega_\alpha)$ belongs to Lip α in each closed part of each open interval (ξ_{j-1}, ξ_j) and satisfies a Lip $\alpha/2$-condition in the neighbourhood of the points ξ_i, so that it reduces to a TURÁN-type case with $\beta = \alpha/2$ (resp. $\beta \geqq \alpha/2$). The problem seems to remain still open for $\beta < \alpha/2$. As an example we mention the function $g_\alpha(x) = |x|^{\alpha/2} \in C^*(\omega_\alpha)$. In this case our theorem asserts $R_n(g_\alpha) = O(n^{-\alpha})$, while the exact order of polynomial approximation is $O(n^{-\alpha/2})$. By the way, it was proved even $R_n(g_\alpha) = O(e^{-cn^{1/3}})$ (see G. FREUD—J. SZABADOS [2]).

4. Proof of the Theorem

Let us consider our function in the interval $[\xi_{k-1}, \xi_k]$ and introduce the transformation

$$x = \frac{\xi_k + \xi_{k-1}}{2} + \frac{\xi_k - \xi_{k-1}}{2}\cos\theta \qquad (0 \leqq \theta \leqq \pi) \tag{2}$$

so that

$$\sqrt{|x - \xi_{k-1}|} = \sqrt{\xi_k - \xi_{k-1}}\cos\frac{\theta}{2} \quad \text{and} \quad \sqrt{|x - \xi_k|} = \sqrt{\xi_k - \xi_{k-1}}\sin\frac{\theta}{2}. \tag{3}$$

From (1) and (3) we deduce, that the modulus of continuity of the even 2π-periodic function

$$F_k(\theta) = f\left(\frac{\xi_k + \xi_{k-1}}{2} + \frac{\xi_k - \xi_{k-1}}{2}\cos\theta\right)$$

satisfies

$$\omega(F_k; \delta) \leqq c_2(f)\omega(\delta).$$

By JACKSON's theorem there exists for each integer n an even trigonometric polynomial $\tau_n(\theta)$ of order n at most for which

$$|F_k(\theta)-\tau_n(\theta)| \leqq c_3(f)\omega\left(\frac{1}{n}\right).$$

Reverting the transformation (2), we obtain from $\tau_n(\theta)$ a rational polynomial $\pi_n^{(k)}(x)$ satisfying

$$|f(x)-\pi_n^{(k)}(x)| \leqq c_4(f)\omega\left(\frac{1}{n}\right), \tag{4}$$

Now our theorem follows from (4) and the weak localization theorem, since $\omega(\delta)\geqq$ $\geqq c_5(f)\delta$ for $f\not\equiv 0$.

5. Some Comments

First of all, let us observe that in the case $m=1$ (i.e. if the exceptional points are only the end-points of $[-1, +1]$) we have a *polynomial* $\pi_n(x)$ of degree n at most with

$$|f(x)-\pi_n(x)| \leqq cK\omega\left(\frac{1}{n}\right),$$

c not depending on f.

As a consequence of (1) the modulus of continuity of $f(x)$ is estimated in order by $\omega(\delta)$ in every closed interval not containing the points ξ_ν and by $\omega(\delta^{1/2})$ near the points ξ_i [2] but these properties do not give a complete description of condition (1). We mention an important case, when a simpler formulation of the condition (1) is possible.

COROLLARY. *Let $\omega_1(\delta)\not\equiv 0$ be a modulus of continuity, let*

$$f(x_1)-f(x_2)=o\{\omega_1(|x_1-x_2|)\} \tag{5}$$

be satisfied uniformly in every closed interval not containing the points ξ_i $(i=0, 1, \ldots m)$ and let

$$f(x_1)-f(x_2)=O\{\omega(|x_1-x_2|)\} \tag{6}$$

uniformly in $[-1, +1]$ *finally let* $\varrho_1(\delta, \eta)=\dfrac{\omega_1(\eta\delta)}{\omega_1(\delta)}$ *be continuous in the triangle* $0\leqq\delta\leqq\eta\leqq 1$ *and* $\varrho_1(0, 0)=0$, *then we have*

$$R_n(f) = o\left\{\omega_1\left(\frac{1}{n}\right)\right\} \tag{7}$$

The exact order of polynomial approximation for the class of functions satisfying the conditions of the corollary for a fixed $\omega_1(\delta)$ is $O\left\{\omega_1\left(\frac{1}{n}\right)\right\}$. The corollary

[2] This can be deduced considering the elementary inequality

$$|\sqrt{a}-\sqrt{b}| \leqq \sqrt{|a-b|}$$

shows that for this class as a whole the order of rational approximation is better than the order of polynomial approximation.

The condition on $\varrho_1(\delta, \eta)$ is clearly satisfied if e.g. $\omega_1(\delta) = A\delta^\alpha$, which is a „Szüsz—Turán-type" case.

In order to deduce the corollary from the theorem, we shell construct an $\omega(\delta)$ satisfying (1) with $K=1$, so that $\omega(\delta) = o\{\omega_1(\delta)\}$. Applying the theorem we will then obtain the corollary.

Let us turn to the construction of $\omega(\delta)$. We define it by

$$\omega(\delta) = \max |f(x_2) - f(x_1)|$$

where x_1 and x_2 runs through all values satisfying the following conditions a) and b):

a) x_1, x_2 are in the same interval $[\eta_j, \xi_j]$ or $[\xi_j, \eta_{j+1}]$,

b) we have $|\sqrt{|x_1 - \xi_j|} - \sqrt{|x_2 - \xi_j|}| \leqq \delta$.

It follows that $\omega(\delta)$ is nondecreasing and $\omega(2\delta) \leqq 2\omega(\delta)$. In what follows, let x_1 and x_2 be the values for which this maximum is attained. We can suppose without loss of generality that $x_1, x_2 \in [\xi_j, \eta_{j+1}]$.

We consider an arbitrary $\varepsilon > 0$ and take δ_1 so small that $\varrho_1(\delta, \eta) \leqq \varepsilon$ for $0 \leqq \delta \leqq \eta \leqq 2\delta_1^{1/2}$. We dessect $[x_1, x_2]$ in $[x_1^{(1)}, x_2^{(1)}] \subseteq [\xi_j, \xi_j + \delta_1]$ and $[x_1^{(2)}, x_2^{(2)}] \subseteq [\xi_j + \delta_1, \eta_{j+1}]$, one of this possibly void.

We obtain

$$|f(x_1^{(1)}) - f(x_2^{(1)})| \leqq \omega_1(|x_1^{(1)} - x_2^{(1)}|) =$$

$$= \omega_1\left\{\left(\sqrt{|\xi_j - x_1^{(1)}|} + \sqrt{|\xi_j - x_2^{(1)}|}\right)\left|\sqrt{|\xi_j - x_1^{(1)}|} - \sqrt{|\xi_j - x_2^{(1)}|}\right|\right\} \leqq$$

$$\leqq \omega_1\left(2\delta_1^{\frac{1}{2}}\left|\sqrt{|\xi_j - x_2^{(1)}|} - \sqrt{|\xi_j - x_1^{(1)}|}\right|\right) =$$

$$= \varrho_1\left(\left|\sqrt{|\xi_j - x_2^{(1)}|} - \sqrt{|\xi_j - x_1^{(1)}|}\right|, 2\delta^{\frac{1}{2}}\right)\omega_1\left(\left|\sqrt{|\xi_j - x_2^{(1)}|} - \sqrt{|\xi_j - x_1^{(1)}|}\right|\right) \leqq$$

$$\leqq \varepsilon\omega_1\left(\left|\sqrt{|\xi_j - x_2^{(1)}|} - \sqrt{|\xi_j - x_1^{(1)}|}\right|\right) \leqq \varepsilon\omega_1(\delta).$$

Let us now consider $|f(x_2^{(2)}) - f(x_1^{(2)})|$ where $x_1^{(2)}, x_2^{(2)} \in [\xi_j + \delta_1, \eta_{j+1}]$, so that

$$\left|\sqrt{|\xi_j - x_2^{(2)}|} - \sqrt{|\xi_j - x_1^{(2)}|}\right| \geqq \frac{|x_1^{(2)} - x_2^{(2)}|}{2\sqrt{\eta_{j+1} - \xi_j}}$$

From (5) we conclude that for a sufficiently small $\Delta(\varepsilon)$ we have for $|x_2^{(1)} - x_1^{(1)}| \leqq \Delta(\varepsilon)$

$$|f(x_1^{(2)}) - f(x_2^{(2)})| \leqq \varepsilon\omega_1(|x_1^{(2)} - x_2^{(2)}|)$$

and let us observe, that the condition imposed on $|x_1^{(2)} - x_2^{(2)}|$ is satisfied provided that

$$\left|\sqrt{|\xi_j - x_2^{(2)}|} - \sqrt{|\xi_j - x_1^{(2)}|}\right| \leqq 2\sqrt{\eta_{j+1} - \xi_j}\,\Delta(\varepsilon)$$

We have then

$$|f(x_1^{(2)}) - f(x_2^{(2)})| \leqq \varepsilon\omega_1(|x_1^{(2)} - x_2^{(2)}|) \leqq$$

$$\leqq \varepsilon\omega_1\left(2\sqrt{\eta_{j+1} - \xi_j}\left|\sqrt{|\xi_j - x_2^{(2)}|} - \sqrt{|\xi_j - x_1^{(2)}|}\right|\right) \leqq$$

$$\leqq 4\varepsilon\sqrt{\eta_{j+1} - \xi_j}\,\omega_1\left(\left|\sqrt{|\xi_j - x_2^{(2)}|} - \sqrt{|\xi_j - x_1^{(2)}|}\right|\right) \leqq \delta\varepsilon\omega_1(\delta).$$

All in all we obtained that under the condition $\delta \leqq \Delta(\varepsilon)$ we have

$$\omega(\delta) \leqq 9\varepsilon\omega_1(\delta),$$

thus our statement follows.

REFERENCES

[1] FREUD, G.: Über Approximation durch rationale gebrochene Functionen, *Acta Math. Acad. Sci. Hungar.* **17** (1966) 313—324.
[2] FREUD, G. and SZABADOS, J.: Rational approximation to $|x^{\alpha}|$, *Acta Math. Ac. Sci. Hungar.*
[3] SZABADOS, J.: Generalization of two theorems of G. Freud concerning rational approximation, Studia Sci. Math. Hungar. **2** (1967) 73—80.
[4] SZÜSZ, P.—TURÁN, P.: A konstruktív függvénytan egy újabb irányáról (Hungarian), *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.* **16** (1966) 33—46.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received January 26, 1967.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 425—429.

A NON-MARKOVIAN QUASI-POISSON PROCESS

by

P. A. P. MORAN

Consider the class of all time-shift invariant point processes on the real line $(-\infty < t < \infty)$. Such a process is defined if we know the joint distribution of $N_1, \ldots, N_n$ the numbers of points occurring in any given set of disjoint intervals $I_1, \ldots, I_n$. It is known that if $N_1, \ldots, N_n$ are independently distributed as Poisson variates with means proportional to the lengths I_i (which we also write as I_i) then the process is a Poisson process. Professor A. RÉNYI [2] has raised the question whether a Poisson process is characterised by the requirement that the number of points occurring in every single interval is a Poisson variate with mean proportional to the length of the interval. We answer this question in the negative by constructing a stationary (i.e. time-shift invariant) point process which is not Markovian but which satisfies this requirement.

Consider first an ordinary Poisson process such that the expected number of points occurring in any interval of length l is equal to l. If $\ldots t_{-1}, t_0, t_1, t_2, \ldots$ are the coordinates of the points it is well known that all the intervals, $I_i = t_i - t_{i-1}$, are independently distributed in negative exponential distribution with probability densities equal to $\exp -x$. We could therefore construct such a point process by constructing a sequence of such random variables I_n $(n = \ldots -1, 0, 1, 2, \ldots)$ and then positioning them in a suitable manner on the line $(-\infty < t < \infty)$.

To carry out the latter procedure requires a little care. It is well known that in a Poisson process such as the above the length of an interval known to cover a specified point, does not have a negative exponential distribution but a distribution with density equal to $x \exp -x$. Such a distribution is the convolution of two variables each independently having a negative exponential distribution.

One way of embedding the intervals in the time axis is therefore to define the points of the process as occurring at $t_i = I_0 + \ldots I_i$ for $i \geqq 0$, and at $t_i = -(I_{-1} + I_{-2} + \ldots + I_i)$ for $i < 0$. The point 0 is omitted and is not regarded as a point of the process.

Another way, which is similar to that which we shall use in the construction of a non-Markovian process, is to suppose that all the I_i have the negative exponential distribution given above except for I_0 which will have a distribution with density $x \exp -x$. Let W be an independent random variable uniformly distributed on the interval $0 \leqq W \leqq 1$. Define $t_0 = -WI_0$, $t_1 = (1 - W)I_0$, and put

$$t_i = t_1 + I_1 + \ldots I_{i-1} \quad \text{for } i > 1,$$
$$= t_0 - I_{-1} - \ldots - I_{-i} \quad \text{for } i < 0.$$

It is then clear that the resulting process is a stationary Poisson point process.

To construct a non-Markovian process with the required property we first construct a sequence of random variables I_n (where $n = \ldots -1, 0, 1, \ldots$) which are not all independent but which are such that each has a negative exponential distribution with density $\exp -x$, and such that the sum of any k of them has a distribution which is the convolution of k negative exponential distributions and therefore has the probability density

$$\Gamma(k)^{-1} e^{-x} x^{k-1}. \tag{1}$$

To do this we first construct a joint distribution of two random non-negative variables, X and Y, such that each has a distribution with density $\exp -x$, their sum has a distribution with density $x \exp -x$, but are such that they are not independent.

Write $f(x, y)$ for the probability density of this joint distribution and put

$$f(x, y) = \exp -(x+y) + f_1(x, y).$$

Let ε be a number such that

$$0 < \varepsilon < e^{-6}.$$

Denote the square

$$(m \leqq x < m+1,\ n \leqq y < n+1)$$

by (m, n) and define $f_1(x, y)$ to be equal to ε in the squares

$$(0, 2),\ (1, 3)\ (2, 1) \quad \text{and} \quad (3, 0),$$

equal to $-\varepsilon$ in

$$(0, 3),\ (1, 2),\ (2, 0), \quad \text{and} \quad (3, 1),$$

and zero elsewhere.

Then $f(x, y) > 0$, and is a probability distribution density of the required type because we can easily verify that

$$\int_0^\infty f_1(x, y)\,dy = \int_0^\infty f_1(x, y)\,dx = \int_0^A f_1(x, A-x)\,dx = 0 \tag{2}$$

for all positive x, y, and A.

It follows that

$$\int_0^\infty f(x, y)\,dx = e^{-y}, \qquad \int_0^\infty f(x, y)\,dy = e^{-x},$$

and

$$\int_0^\alpha f(x, \alpha - x)\,dx = \alpha e^{-\alpha}. \tag{3}$$

We now define the sequence of random intervals $\{I_n\}$ $(n = 0, \pm 1, \ldots)$ as follows. We suppose that (I_n, I_{n+1}) $(n = 0, \pm 2, \pm 4, \ldots)$ are pairs of random variables having joint distributions with density $f(x, y)$ defined above so that

$$\Pr(I_n < x, I_{n+1} < y) = \int_0^x \int_0^y f(u, v)\,du\,dv$$

for $n=0, \pm 2, \pm 4, \ldots$, and we impose the further condition that different pairs are independently distributed. Then from the manner of construction it will be seen that the sum of any k specified intervals chosen out of this sequence will have a distribution with the probability density (1) even though some of these intervals may not be independent in pairs.

We now have to construct a point process on the real line such that if the intervals between successive points are denoted by J_n they will be distributed as the sequence $\{I_n\}$ defined above or as the sequence $\{I_{n+1}\}$, depending on whether J_0 happens to be the left hand or right hand member of a pair.

To do this we first consider the sequence of intervals $\{K_n\}$ where K_n is the interval made up of I_n and I_{n+1} (the symbol K_n is defined for even values of n only). The intervals K_n are distributed independently in distributions with the density $x \exp -x$ and can be regarded as the intervals between the points of a stationary renewal process. To embed them on the line $(-\infty < t < \infty)$ we use the fact that in such a stationary renewal process the length of the interval covering any prescribed time point t will have a probability distribution with density $\frac{1}{2}x^2 \exp -x$. Moreover conditionally on the length of the interval having a given value l, the distance of the point t from the left hand end of the interval will be uniformly distributed over the interval $(0, l)$ (See, for example, COX and MILLER [1] p. 356). We can use this to embed the sequence of intervals K_n on the time axis as follows.

Consider the point $t=0$. We suppose that this is covered by an interval K_0 of length l having the probability distribution with density $\frac{1}{2}x^2 \exp -x$. As we want to identify K_0 with the sum of intervals I_0 and I_1 we proceed as follows. If X and Y have the joint distribution whose density is $f(x, y)$ defined above, write $g(x|z)$ for the probability density of the distribution of X when $Z=X+Y$ has the value Z. Let U be a random variable uniformly distributed on the interval (0, 1). Let l be a random variable with the distribution $\frac{1}{2}x^2 \exp -x$, and X be a random variable with the distribution $g(x|l)$. Then we define K_0 to be the interval

$$(-Ul, (1-U)l)$$

and suppose it to be composed of the two intervals

$$I_0 = (-Ul, -Ul+X),$$

$$I_1 = (-Ul+X, (1-U)l).$$

We then fill up the real line $(-\infty < t < \infty)$ to the right and left by adding pairs of intervals (I_2, I_3), (I_4, I_5), ... and (I_{-2}, I_{-1}), (I_{-4}, I_{-3}), ...

It is clear from the theory of stationary renewal processes that the sequence of intervals $\{K_n\}$, thus embedded on the time axis, is a stationary renewal process, and the distribution of the K-interval covering any specified point t will have a length with the distribution $\frac{1}{2}x^2 \exp -x$ and will be positioned relatively to t in the same way that K_0 is positioned relative to $t=0$.

We now have to prove that the distribution of the number of end points of I-intervals lying in any specified interval of length L is a Poisson distribution with mean L. Since the process is stationary we can suppose this interval to be the interval $(0, L)$. We now determine the distribution of the k-th point to the right of $t=0$.

Consider first the distribution of the nearest point to the right of the point $t=0$ which we denote by 0. Write A and C for the end points $t=-Ul$, $t=(1-U)l$, and B for the point $t=-Ul+X$ which divides AC into the two intervals I_0 and I_1. Then the nearest point of the process to the right of 0 will be B if B lies in the interval OC, and will be C if B lies in the interval AO.

If α is any number greater than zero, let P_1 be the probability that B lies in OC and OB is greater than α. Remembering that conditional on $AC=l$, AO is uniformly distributed over the interval $(0, l)$, and independently of the position of 0, $X=AB$ has the distribution $g(x|l)$ we see that P_1 is equal to

$$\int_\alpha^\infty dl \int_\alpha^l dx \left(\frac{x-\alpha}{l}\right) \frac{1}{2} l^2 e^{-l} g(x|l). \tag{4}$$

Putting $l=x+y$, changing the variables of integration to x and y, and observing that

$$le^{-l}g(x|l)=f(x, y),$$

P_1 is equal to

$$\frac{1}{2}\int_\alpha^\infty dx \int_0^\infty dy(x-\alpha)f(x,y) = \frac{1}{2}\int_\alpha^\infty (x-\alpha)e^{-x}dx = \frac{1}{2}e^{-\alpha}. \tag{5}$$

Now let P_2 be the probability that B lies in AO and $OC>\alpha$. This is equal to

$$\int_\alpha^\infty dl \int_0^{l-\alpha} dx \left(\frac{l-\alpha-x}{l}\right) \frac{1}{2} l^2 e^{-l} g(x|l) = \frac{1}{2}\int_\alpha^\infty dl \int_0^{l-\alpha} dx(l-\alpha-x)le^{-l}g(x|l). \tag{6}$$

Changing again to the variables (x, y), this is the integral of $l-\alpha-x=y-\alpha$ over the range $x>0$, $y>\alpha$ and is therefore

$$\int_0^\infty dx \int_\alpha^\infty dy(y-\alpha)f(x,y) = \int_\alpha^\infty (y-\alpha)e^{-y}dy = \frac{1}{2}e^{-\alpha}. \tag{7}$$

Adding these cases together we see that the probability that the nearest right hand point to 0 is distant more than α from 0 is $\exp -\alpha$.

The second point to the right of 0 is either C (if B lies in OC), or the dividing point of K_2 if B lies in AO. Consider the second case first. If B lies in AO the distance of 0 from the second point on the right is the sum of the intervals OC and I_2, the left-hand interval of the subdivision of K_2. These intervals are independently distributed with densities $\exp -x$, and their sum has a distribution with density $x \exp -x$ as required.

Let P_3 be the probability that B lies in OC and $OC>\alpha$. This is equal to

$$\frac{1}{2}\int_\alpha^\infty dl \int_0^{l-\alpha} dx\, xle^{-l}g(x|l) + \frac{1}{2}\int_\alpha^\infty dl \int_{l-\alpha}^l dx(l-\alpha)le^{-l}g(x|l).$$

In the first of these integrals the range is over the region $(x+y>\alpha, 0<x<l-\alpha)$ which is equivalent to $(x>0, y>\alpha)$, and in the second the range is over the region $(x+y>\alpha, x+y-\alpha<x<x+y)$ which is equivalent to $(x+y>\alpha, y<\alpha)$. The latter

can be regarded as the difference of the regions $(x>0,\ y<\alpha)$ and $(0<x+y<\alpha)$. Changing the variables of integration, the above sum is therefore equal to

$$\frac{1}{2}\int_0^\infty dx\int_\alpha^\infty dy\, xf(x,y)+\frac{1}{2}\int_0^\infty dx\int_0^\alpha dy(x+y-\alpha)f(x,y)-$$

$$-\frac{1}{2}\int_0^\alpha dx\int_0^{\alpha-x} dy(x+y-\alpha)f(x,y)$$

$$=\frac{1}{2}\int_0^\infty dx\int_0^\infty dy\, xf(x,y)+\frac{1}{2}\int_0^\infty dx\int_0^\alpha dy(y-\alpha)f(x,y)$$

$$-\frac{1}{2}\int_0^\alpha dx\int_0^{\alpha-x} dy(x+y-\alpha)f(x,y)$$

$$=\frac{1}{2}\int_0^\infty xe^{-x}dx+\frac{1}{2}\int_0^\alpha (y-\alpha)e^{-y}dy-\frac{1}{2}\int_0^\alpha (z-\alpha)ze^{-z}dz$$

on using the results (3). The sum of these integrals is easily found to be

$$\frac{1}{2}e^{-\alpha}(1+\alpha)$$

which is half the probability that a random variable with density (1) with $k=2$ exceeds α. Thus the second point to the right has a distance from 0 with the required probability distribution.

Now consider the distribution of the distance of the k-th point to the right where $k>2$. If B lies to the right of 0, C is the second point to the right and the distance to the k-th point to the right is found by adding $k-2$ I-intervals whose sum has the distribution (1) with k replaced by $k-2$. Furthermore the sum of these $k-2$ intervals is distributed independently of OC. Thus the distribution of the distance to the k-th point has the density (1). Similarly if B lies to the left of 0 we add $k-1$ intervals and obtain the same result.

The probability that there are at least k points in the interval (O, L) is the probability that the distance to the k-th point to the right is less than L. Integrating the integral of (1) by parts this is the required tail of the Poisson distribution.

We have therefore constructed a stationary point process which is not a Poisson process but is such that the number of points occurring in any interval of length L has a Poisson distribution with mean L.

REFERENCES

[1] Cox, D. R. and Miller, H. D.: *The Theory of Stochastic Processes*, Methuen, London, 1965.
[2] Rényi, A.: Remarks on the Poisson process, *Studia Sci. Math. Hungar.* **2** (1967) 119—123.

AUSTRALIAN NATIONAL UNIVERSITY, CANBERRA, A. C. T.

(Received January 31, 1967.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 431—434.

ON THE SEQUENCE OF GENERALIZED PARTIAL SUMS OF A SERIES

by

G. TUSNÁDY

RÉNYI in the papers [1] and [2] has investigated the following problem:

Let $\sum_{k=0}^{\infty} a_k$ be a series and denote by A_n the sum

$$A_n = a_{k_1} + a_{k_2} + \ldots + a_{k_r} \tag{1}$$

provided that

$$n = 2^{k_1} + 2^{k_2} + \ldots + 2^{k_r} \tag{2}$$

is the representation of n in the binary number system. That is, we consider all finite sums of the terms of the series $\sum a_k$, and arrange them in lexicographic order. The problem is the relation between the convergence properties of the series $\sum a_k$ and those of the sequence $\{A_n\}$. It is trivial that to a convergent series $\sum a_k$ may correspond a divergent sequence $\{A_n\}$. RÉNYI in the paper [1] has proved that the sequence $\{A_n\}$ has a $(C, 1)$-limit if the series $\sum a_k$ is convergent in the ordinary sense, that is if the series $\sum a_k$ is $(C, 0)$-summable. In this direction we can prove in a similar way that the sequence $\{A_n\}$ has a (C, ε)-limit with positive ε, if the series $\sum a_k$ is $(C, 0)$ summable. Some difficulties arise in proving the $(C, \alpha+\varepsilon)$-summability of the sequence $\{A_n\}$, if the series $\sum a_k$ is (C, α) summable.

Another question is what can be said about the series $\sum a_k$ if we know the convergence properties of the sequence $\{A_n\}$. In the paper [1] RÉNYI has proved that from the $(C, 2)$-summability of the sequence $\{A_n\}$ follows the $(C, 0)$-summability of the series $\sum a_k$. In this paper I give an elementary proof for the $(C, 0)$-summability of the series $\sum a_k$, provided that the sequence $\{A_n\}$ is Abel-Summable, under a weaker Tauberian condition than in the paper [2].

THEOREM. *If the sequence $\{A_n\}$ is Abel-summable and $|a_k| \leqq M$ $(k = 1, 2, \ldots)$, then the series $\sum a_k$ is convergent in the ordinary sense, i.e. it is $(C, 0)$-summable.*

PROOF. We may rearrange the absolutely convergent series $\sum A_n x^n$ in the following way:

$$\sum_{n=0}^{\infty} A_n x^n = \sum_{n=0}^{\infty} x^n \sum_{k_i \in B_n} a_{k_i} = \sum_{k=0}^{\infty} a_k \sum_{k \in B_n} x^n = \frac{1}{1-x} \sum_{k=0}^{\infty} a_k \frac{x^{2^k}}{1+x^{2^k}}$$

where B_n is the set $\{k_1, k_2, \ldots, k_r\}$ in (2). Thus from the Abel-summability of the sequence $\{A_n\}$ follows that

$$\lim_{x \to 1-0} \sum_{k=0}^{\infty} a_k \frac{x^{2^k}}{1+x^{2^k}}$$

exists. Let $x_0 \in (0, 1)$. We need only that the above limit exists on the sequence $\{x_0^{2^{-N}}\}_{N=1}^{\infty}$ i.e. the sequence $\{B_N\}$ is convergent, where

$$B_N = \sum_{k=-N}^{\infty} a_{k+N} \frac{x_0^{2^k}}{1+x_0^{2^k}}$$

Let $x_0 = \frac{1}{y_0}$ then $y_0 > 1$ and $B_N = \sum_{k=-N}^{\infty} a_{k+N} f(y_0^{2^k})$ where $f(y) = \frac{1}{1+y}$.

In the first part of our proof we prove that if the sequence $\{B_N\}$ is convergent then $\lim_{n\to\infty} a_n = 0$.

Let α and β be arbitrary real numbers and $f^*(x) = \alpha f(x) + \beta f(x^2)$, then

$$B_N^* = \sum_{k=-N}^{\infty} a_{k+N} f^*(y_0^{2^k}) = \alpha \sum_{k=-N}^{\infty} a_{k+N} f(y_0^{2^k}) + \beta \sum_{k=-N}^{\infty} a_{k+N} f(y_0^{2^{k+1}}) = \alpha B_N + \beta B_{N-1}$$

Thus the convergence of the series $\{B_N\}$ implies the convergence of the series $\{B_N^*\}$. Especially, when $\alpha = -\beta = 1$, then $\lim_{N\to\infty} B_N^* = 0$. We shall step by step determine the functions $f_i(y)$ in order to increase the peak of the sequence $\{f_i(y_0^{2^k})\}_{-\infty}^{\infty}$. We need only a term $f_i(y_0) > \frac{1}{2}$, then (provided that the sequence $\{f_i(y_0^{2^k})\}_{-\infty}^{\infty}$ is positive and its sum is 1) follows that $\lim_{k\to\infty} a_k = 0$. Indeed the following lemma is true:

LEMMA. Let $b_k \geqq 0$ $(k = 0, \pm 1, \pm 2, \ldots)$ $\sum_{k=-\infty}^{\infty} b_k = 1$ and $b_0 > \frac{1}{2}$ then from $\lim_{N\to\infty} B_N = 0$ where $B_N = \sum_{k=-N}^{\infty} a_{k+N} b_k$ follows that $\lim_{k\to\infty} a_k = 0$, provided that $|a_k| < M$.

PROOF OF THE LEMMA. Assume indirectly that

$$\overline{\lim_{k\to\infty}} |a_k| = a > 0$$

then to some $\varepsilon > 0$ and some natural n_0 there is a natural N as large as we need so that $|a_N| > a - \varepsilon$ and $|a_n| < a + \varepsilon$ if $n > n_0$. Then

$$|B_N| = \left| \sum_{k=-N}^{\infty} a_{k+N} b_k \right| = \left| a_N b_0 + \sum_{k=-N}^{-N+n_0} a_{k+N} b_k + \sum_{\substack{k=-N+n_0+1 \\ k\neq 1}}^{\infty} a_{k+N} b_k \right| \geqq$$

$$\geqq b_0(a-\varepsilon) - M \sum_{k=-N}^{-N+n_0} b_k - (a+\varepsilon)(1-b_0) = a(2b_0-1) - M \sum_{k=-N}^{-N+n_0} b_k - \varepsilon$$

where $\sum_{k=-N}^{-N+n_0} b_k$ tends to 0 if n_0 is fixed and N tends to infinity so we get that

$$\overline{\lim_{N\to\infty}} |B_N| \geqq a(2b_0-1) - \varepsilon > 0$$

But $\lim_{N\to\infty} B_N = 0$ so our indirect assumption was false and the lemma is true.

We return to the proof of our theorem. The function $f^*(y)$ is positive if $\alpha f(y) + \beta f(y^2) > 0$, that is

$$\frac{f(y^2)}{f(y)} < -\frac{\alpha}{\beta}$$

provided that $\beta<0$. If the function $f(y^2)/f(y)$ increases with $y\geqq 1$, then we may take

$$\alpha=\lim_{y\to 1+0}\frac{f(y^2)}{f(y)};\qquad \beta=-1$$

According to this remark let

$$f_1(y)=f(y)-f(y^2)$$

$$f_{k+1}(y)=2^{2k-1}f_k(y)-f_k(y^2)\qquad (k=1,2,\ldots).$$

We need only that with this transformation we get a function $f_5(y)$ in the form

$$f_5(y)=\frac{(y-1)^9 h_5(y)}{(y+1)(y^2+1)\ldots(y^{32}+1)}$$

where the coefficients of the function $h_5(y)$ are positive which may be verified by the actual carrying out of the transformation

$$h_1(y)=y$$

$$h_{k+1}(y)=\frac{1}{(y-1)^2}[2^{2k-1}(y^{2^{k+1}}+1)h_k(y)-(y+1)^{2k}h_k(y^2)].$$

According to $f(1)=\frac{1}{2}$ and $\lim_{y\to\infty}f(y)=0$ we get that $s_1=\sum_{k=-\infty}^{\infty}f_1(y_0^{2^k})=\frac{1}{2}$, thus $s_5=\sum_{k=-\infty}^{\infty}f_5(y_0^{2^k})=\frac{1}{2}(2-1)(2^3-1)(2^5-1)(2^7-1)$, and with some numerical calculation we may see that $f_5(4,2)>\frac{s_5}{2}$, and applying our lemma to the series $b_k=\frac{1}{s_5}f_5(y_0^{2^k})$ with $y_0=4,2$ we get that $\lim_{k\to\infty}a_k=0$.

Although in the paper [2] RÉNYI has proved by applying the „high-indices" theorem that if $a_n\to 0$, then from the Abel-summability of the sequence $\{A_n\}$ follows the convergence of the series $\sum a_k$, for the sake of making our proof completely elementary we complete also this part of the proof. This needs only to remark that if we put $c_k=f(y_0^{2^k})$ then

$$\left|\sum_{k=0}^{N}a_k-2C_N\right|=\left|\sum_{k=0}^{N}a_k-2\sum_{k=0}^{\infty}a_kc_{k-N}\right|\leqq\left|\sum_{k=0}^{n}a_k(1-2c_{k-N})\right|+$$

$$+\left|\sum_{k=n+1}^{N}a_k(1-2c_{k-N})\right|+2\left|\sum_{k=N+1}^{\infty}a_kc_{k-N}\right|\leqq nM(1-2c_{n-N})+$$

$$+2\varepsilon\left\{\sum_{k=0}^{\infty}\left(\frac{1}{2}-c_{-k}\right)+\sum_{k=1}^{\infty}c_k\right\}\leqq\eta,$$

where $C_N=\sum_{k=0}^{\infty}a_kc_{k-N}$, $\varepsilon\leqq\frac{1}{4}\eta\left\{\sum_{k=0}^{\infty}\left(\frac{1}{2}-c_{-k}\right)+\sum_{k=1}^{\infty}c_k\right\}^{-1}$; and n is so large that $|a_j|<\varepsilon$ if $j>n$; and N is so large that

$$1-2c_{n-N}\leqq\frac{\eta}{2nM}$$

Hence $\sum_{k=1}^{\infty}a_k=2\lim_{N=\infty}C_N$, and the proof of the theorem is completed.

REFERENCES

[1] RÉNYI, A.: Mathematical notes, II, *Publ. Math. Debrecen.* **5** (1957) 129—141.
[2] RÉNYI, A.: Summation methods and probability theory, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **4** (1959) 389—399.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received February 3, 1967.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 435—439.

ON THE SEQUENCE OF GENERALIZED PARTIAL SUMS OF A SERIES

by

G. HALÁSZ

The aim of this paper is to prove the following conjecture of A. RÉNYI:[1]

THEOREM. *If*

$$f(t) = \sum_{k=0}^{\infty} a_k \frac{e^{-2^k t}}{1+e^{-2^k t}}$$

exists for $t>0$ *and tends to a limit* A *as* $t \to +0$ *then* $\sum_{k=0}^{\infty} a_k$ *is convergent to* $2A$.

There is a general theorem, the high indices theorem of LEVINSON that could be applied as to the form of the summability method. Its basic conditions concern a certain Fourier integral which in our case would be

$$\Gamma(s)\zeta(s)\left(1-\frac{2}{2^s}\right)$$

and by its numerous zeros in $\operatorname{Re} s>0$ it does not satisfy those conditions. Using, however, the special form of the exponents 2^k and the fact that in a small domain this function really does not vanish, we succeed in proving our unrestricted Tauberian theorem.

PROOF. We have

$$f(t) = \sum_{k=0}^{\infty} a_k \frac{e^{-2^k t}}{1+e^{-2^k t}} = \sum_{k=0}^{\infty} a_k \sum_{l=1}^{\infty} (-1)^{l+1} e^{-l2^k t} = \sum_{m=1}^{\infty} e^{-mt} \sum_{2^k | m} a_k (-1)^{\frac{m}{2^k}+1} =$$

$$\stackrel{\text{def}}{=} \sum_{m=1}^{\infty} b_m e^{-mt}$$

According to the number theoretic relation between a_k and b_m, first only formally

$$\sum_{k=0}^{\infty} \frac{a_k}{2^{ks}} \sum_{l=1}^{\infty} \frac{(-1)^{l+1}}{l^s} = \sum_{k=0}^{\infty} \frac{a_k}{2^{ks}} \zeta(s)\left(1-\frac{2}{2^s}\right) = \sum_{m=1}^{\infty} \frac{b_m}{m^s} \tag{1}$$

We see that in case $a_0=1$ which can be assumed, b_m is a multiplicative number theoretical function taking on 1 at each odd number while

$$b_{2^k} = a_k - \sum_{j=0}^{k-1} a_j$$

To prove that the yet formal identity (1) really holds we, nevertheless, use LEVINSON's

[1] For the backgraund of this problem see the previous paper in this issue by G. TUSNÁDY.

high indices theorem [1] to get a preliminary estimation of a_k. It is simpler to deal with b_m:

$$f(t) = \sum_{m=1}^{\infty}{}' b_m e^{-mt} = \sum_{k=0}^{\infty} b_{2^k} \sum_{l=1}^{\infty}{}' e^{-(2l-1)2^k t} = \sum_{k=0}^{\infty} b_{2^k} \frac{e^{-2^k t}}{1-e^{-2^{k+1}t}}$$

Here we assume only $f(t)=o\left(\frac{1}{t}\right)$ as $t \to +0$:

$$f(t)t = \sum_{k=0}^{\infty} b_{2^k} \frac{te^{-2^k t}}{1-e^{-2^{k+1}t}} = \sum_{k=0}^{\infty} \frac{b_{2^k}}{2^k} \cdot \frac{2^k \cdot t \cdot e^{-2^k t}}{1-e^{-2\cdot 2^k t}} = \sum_{k=0}^{\infty} \frac{b_{2^k}}{2^k} N(2^k t) = o(1)$$

with $N(x)=\frac{xe^{-x}}{1-e^{-2x}}$. It has the form required in the cited theorem. To be able to use it, we have to compute

$$\int_0^{\infty} N'(x) x^s \, dx.$$

The integral is convergent for $\operatorname{Re} s > -1$ and represents a regular function there. But first let $\operatorname{Re} s > 1$ where the following partial and termwise integrations are legitimate:

$$\int_0^{\infty} N'(x) x^s \, dx = -s \int_0^{\infty} N(x) x^{s-1} \, dx = -s \int_0^{\infty} \frac{xe^{-x}}{1-e^{-2x}} x^{s-1} \, dx =$$

$$= -s \int_0^{\infty} x^s \sum_{l=1}^{\infty}{}' e^{-(2l-1)x} dx = -s \sum_{l=1}^{\infty}{}' \int_0^{\infty} e^{-(2l-1)x} x^{s-1} \, dx =$$

$$= -s \sum_{l=1}^{\infty}{}' \frac{1}{(2l-1)^{s+1}} \Gamma(s+1) = -\Gamma(s+1)\left(1-\frac{1}{2^{s+1}}\right) s\zeta(s+1).$$

In effect, we achieved a shift of the original half plane into a zero free one. In fact, neither $\Gamma(s+1)\left(1-\frac{1}{2^{s+1}}\right)$ nor $s\zeta(s+1)$ vanish in $\operatorname{Re} s \geqq 0$ and from well-known estimations of these functions follow all the conditions of LEVINSON's theorem. It gives therefore

$$\sum_{k=0}^{n}{}' \frac{b_{2^k}}{2^k} = \sum_{k=0}^{n}{}' \frac{1}{2^k}\left(a_k - \sum_{j=0}^{k-1}{}' a_j\right) = \sum_{j=0}^{n}{}' a_j \left(\frac{1}{2^j} - \sum_{k=j+1}^{n}{}' \frac{1}{2^k}\right) = \frac{1}{2^n} \sum_{j=0}^{n}{}' a_j = o(1)$$

hence $a_k = o(2^k)$. As a consequence, the left of the formal identity (1) converges absolutely for $\operatorname{Re} s > 1$ and so does therefore its right. With that, we proved the identity valid in the half plane $\operatorname{Re} s > 1$. Hence in the same domain we have, by well-known transformation:

$$\sum_{m=1}^{\infty}{}' \frac{b_m}{m^s} = \frac{1}{\Gamma(s)} \int_0^{\infty} \sum_{m=1}^{\infty}{}' b_m e^{-mt} t^{s-1} \, dt = \frac{1}{\Gamma(s)} \int_0^{\infty} f(t) t^{s-1} \, dt,$$

$$\sum_{k=0}^{\infty}{}' \frac{a_k}{2^{ks}} = \frac{1}{\zeta(s)\left(1-\frac{2}{2^s}\right)\Gamma(s)} \int_0^{\infty} f(t) t^{s-1} \, dt \tag{2}$$

$f(t)$ is bounded and tends exponentially to 0 when $t \to +\infty$ so that the integral represents a regular function even for $\operatorname{Re} s > 0$. On the other hand, $\zeta(s)$ is known to have no zero in $\operatorname{Re} s > 0$, $|\operatorname{Im} s| \leqq \dfrac{\pi}{\log 2} = 4{,}5\ldots$. Therefore, $\sum_{k=0}^{\infty}{}' \dfrac{a_k}{2^{ks}}$ is regular in this domain and as it is periodical with period $\dfrac{2\pi}{\log 2} i$ it is regular for all $\operatorname{Re} s > 0$ and we can express the partial sums s_n of $\sum_{k=0}^{\infty} a_k$ as an integral on any vertical segment of length $\dfrac{2\pi}{\log 2}$ in the right half plane. Instead of s_n, let us regard first

$$s_n \pm i s_{n-1} = \frac{\log 2}{2\pi i} \int_{I_\pm} \frac{1 \pm 2^{-s}}{1 - 2^{-s}} \cdot 2^{ns} \sum_{k=0}^{\infty}{}' \frac{a_k}{2^{ks}}\, ds =$$

$$= \frac{\log 2}{2\pi i} \int_0^{\infty} \frac{f(t)}{t} \int_{I_\pm} \frac{1 \pm 2^{-s}}{1 - 2^{-s}} \frac{2^{ns} t^s}{\zeta(s)\left(1 - \dfrac{2}{2^s}\right)\Gamma(s)}\, ds\, dt \overset{\text{def}}{=} \frac{\log 2}{2\pi i} \int_0^{\infty} \frac{f(t)}{t} \int_{I_\pm} F_\pm(s)(2^n t)^s\, ds\, dt$$

where I_+ and I_- are segments of the described type. This has the advantage that $F_\pm(s)$, while regular in $|\operatorname{Re} s| \leqq \dfrac{1}{2}$, $|\operatorname{Im} s| \leqq \dfrac{3\pi}{\log 2}$, vanishes on the imaginary axis because of the factor 1 ± 2^{-s} in case of $+$ at $-i\dfrac{\pi}{2\log 2}$ and $i\dfrac{3\pi}{2\log 2}$ in case of $-$ at $-i\dfrac{3\pi}{2\log 2}$ and $i\dfrac{\pi}{2\log 2}$ and if we choose the endpoints of I_+ and I_- with these imaginary parts and with real part $\dfrac{1}{n}$ then $F_\pm(s)$ vanishes near the endpoints of $I_\pm$.

There is obviously no loss of generality in supposing that the limit $f(+0)$ is 0, and it is enough to prove that then $s_n \pm i s_{n-1} = o(1)$. We may confine ourself to the $+$ case.

To $\varepsilon > 0$ there exists a $0 < \delta < 1$ such that $|f(t)| < \varepsilon$ in $(0, \delta)$. For large n we split the range of integration with respect to t into four parts

$$A: 0 < t \leqq 2^{-n-1}, \quad B: 2^{-n-1} \leqq t \leqq 2^{-n+1}, \quad C: 2^{-n+1} \leqq t \leqq \delta, \quad D: \delta \leqq t < +\infty$$

B is the simplest case. The inner integral is trivially less than

$$c_1 (2^n t)^{\frac{1}{n}} = 2c_1 t^{\frac{1}{n}} < 2c_1$$

and

$$\left| \int_{2^{-n-1}}^{2^{-n+1}} \frac{f(t)}{t} \int_{I_+} \ldots \right| < 2c_1 \varepsilon \int_{2^{-n-1}}^{2^{-n+1}} \frac{1}{t}\, dt = 2c_1 \log 4 \cdot \varepsilon$$

In A where $T \overset{\text{def}}{=} t \cdot 2^n \leqq 2^{-1}$, instead of integrating on I_+, we start horizontally from the lower endpoint and proceed till the line $\operatorname{Re} s = \frac{1}{2}$ from where we integrate upwards on this line till the height of the other endpoint of I_+ and then go back

horizontally. On the horizontal lines, by the mentioned zero of $F_+(s)$

$$\left|\int F_+(s)T^s ds\right| \leqq c_2 \int_{\frac{1}{n}}^{\frac{1}{2}} \sigma e^{-\sigma|\log T|} d\sigma \leqq c_2 \int_0^{\infty} \sigma e^{-\sigma|\log T|} d\sigma =$$

$$= \frac{c_2}{\log^2 T} \int_0^{\infty} \sigma e^{-\sigma} d\sigma = \frac{c_3}{\log^2 T}$$

On the line $\operatorname{Re} s = \frac{1}{2}$ even better

$$\left|\int F_+(s)T^s ds\right| \leqq c_4 T^{\frac{1}{2}} \leqq \frac{c_5}{\log^2 T}$$

This gives for A

$$\left|\int_0^{2^{-n-1}} \frac{f(t)}{t} \int_{I_+} \ldots\right| \leqq c_6 \varepsilon \int_0^{2^{-n-1}} \frac{1}{t^2 \log^2 (t2^n)} dt = c_6 \varepsilon \int_0^{\frac{1}{2}} \frac{1}{t \log^2 t} dt = c_7 \varepsilon.$$

In C and D where $T = t2^n \geqq 2$, the estimation of the inner integral runs the same way, namely we deform the path of integration similarly to the previous case, the line $\operatorname{Re} s = \frac{1}{2}$ replaced by $\operatorname{Re} s = -\frac{1}{2}$. On the horizontal segments

$$\left|\int F_+(s)T^s ds\right| \leqq c_8 \left\{\int_{-\frac{1}{2}}^{0} |\sigma| e^{\sigma \log T} d\sigma + \int_0^{\frac{1}{n}} \sigma T^{\sigma} d\sigma\right\} \leqq$$

$$\leqq c_8 \left\{\int_0^{\infty} \sigma e^{-\sigma \log T} d\sigma + \frac{1}{n^2} T^{\frac{1}{n}}\right\} \leqq c_9 \left\{\frac{1}{\log^2 T} + \frac{1}{n^2} T^{\frac{1}{n}}\right\},$$

while on the vertical one $\left|\int F_+(s)T^s ds\right| \leqq c_{10} T^{-\frac{1}{2}} \leqq \frac{c_{11}}{\log^2 T}$.

In C this latter is the dominating term since

$$\frac{1}{n^2} T^{\frac{1}{n}} = \frac{(t2^n)^{\frac{1}{n}}}{n^2} \leqq \frac{2}{n^2} = \frac{2 \log^2 2}{\log^2 2^n} \leqq \frac{2 \log^2 2}{\log^2 (t2^n)} = \frac{2 \log^2 2}{\log^2 T}$$

and

$$\left|\int_{2^{-n+1}}^{\delta} \frac{f(t)}{t} \int_{I_+} \ldots\right| \leqq c_{12} \varepsilon \int_{2^{-n+1}}^{\delta} \frac{1}{t \cdot \log^2 (t2^n)} dt \leqq c_{12} \varepsilon \int_2^{\infty} \frac{1}{t \log^2 t} dt = c_{13} \varepsilon.$$

In D

$$\frac{1}{\log^2 (t2^n)} \leqq \frac{1}{\log^2 (\delta 2^n)} \leqq \frac{c_{14}(\delta)}{n^2} \leqq \frac{c_{14}(\delta)}{n^2} T^{\frac{1}{n}}$$

and $|f(t)| \leqq c_{15}(\delta)e^{-t}$ hence

$$\left| \int_{\delta}^{\infty} \frac{f(t)}{t} \int_{I_+} \dots \right| \leqq c_{16}(\delta) \int_{\delta}^{\infty} \frac{e^{-t}}{n^2 t} t^{\frac{1}{n}} \leqq \frac{c_{17}(\delta)}{n^2}$$

and we see that the contributions of all the four ranges are small if we choose $\varepsilon > 0$ small and then, depending on it, n large. Q.e.d.

REFERENCE

[1] LEVINSON, N.: Gap and Density Theorems, *Amer. Math. Soc. Coll. Publ.*, Vol. XXVI., Chap. X., Th. LIII.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received February 3. 1967.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 441—446.

ON A NEW CLASS OF UNIMODAL INFINITELY DIVISIBLE DISTRIBUTION FUNCTIONS AND RELATED TOPICS

by

P. MEDGYESSY

1.

A distribution function $F(x)$ is called unimodal with the mode at $x=a$ (or, in short, "(*a*) *unimodal*") if and only if the graph of $F(x)$ is convex in $(-\infty, a)$ and concave in (a, ∞). If $F(x)$ is strictly convex in $(-\infty, a)$ and strictly concave in (a, ∞) then $F(x)$ is called strictly unimodal with the mode at $x=a$ (or, in short, "*strictly* (*a*) *unimodal*"). The point a may be a point of discontinuity.

We remark that if throughout this paper "convex", "concave" and "unimodal" are substituted simultaneously by "strictly convex", "strictly concave", "strictly unimodal" then one gets true assertions again.

Certain linear operations executed on unimodal distribution functions conserve unimodality; in some cases it is also essential that the occurring distribution functions be symmetrical with respect to $x=b$, say (or, in short, "(*b*) *symmetrical*"). — An important type of such linear operations is presented by

THEOREM 1. 1. *Let $F(x; y)$ $(c<y<d)$ be a one-parameter family of* (*a*) *unimodal distribution functions and $A(y)$ be some distribution function. Then the mixture*

$$G(x) = \int_c^d F(x; y)\, dA(y) \tag{1.1}$$

(where c, d are appropriate constants) is an (*a*) *unimodal distribution function. If in addition $F(x, y)$ is* (*a*) *symmetrical, then $G(x)$ is also* (*a*) *symmetrical.*

PROOF. It suffices to prove the unimodality. For $x \leqq a$, $G(x)$ is a mixture of convex functions and, consequently, is itself convex (as to this as well as other properties of convex functions, see [1], p. 299). By a similar reasoning one obtains that $G(x)$ is concave for $x>a$. This completes the proof.

Passing over to characteristic functions we get

COROLLARY 1. 1. *Let $\varphi(t; y)$ be the characteristic function belonging to the* (*a*) *unimodal distribution function $F(x; y)$. Then*

$$\omega(t) = \int_c^d \varphi(t; y)\, dA(y) \tag{1.2}$$

is the characteristic function of an (*a*) *unimodal distribution function $G(x)$. If, in addition, $F(x; y)$ is* (*a*) *symmetrical then $G(x)$ is also* (*a*) *symmetrical.*

An important particular case of (1. 2) is the case $\varphi(t; y)=[\varphi(t)]^y$ provided it is the characteristic function of an (a) unimodal distribution function. We know

that for $y=0, 1, 2, \ldots$ this is again a characteristic function; moreover, if $\varphi(t)$ is the characteristic function of an infinitely divisible distribution function, then $[\varphi(t)]^y$ is a characteristic function for every $y>0$. However, the question: when is $[\varphi(t)]^y$ the characteristic function of an (*a*) *unimodal* distribution function — is rather complicated. Some answer to this question can be obtained if the distribution function belonging to $\varphi(t)$ is (0) unimodal and (0) symmetrical, too. Namely, a theorem of A. WINTNER (see e.g. [2], p. 841) states that the convolution of (0) symmetrical (0) unimodal distribution functions is also (0) symmetrical and (0) unimodal. Hence it follows that if $\varphi(t)$ is the characteristic function of a (0) symmetrical (0) unimodal distribution function then $[\varphi(t)]^n$ $(n=0, 1, 2, \ldots)$ is a characteristic function of the same type. Taking the distribution function belonging to a probability distribution $\{p_r\}$ $(r=0, 1, 2, \ldots)$ for $A(y)$ in Corollary 1. 1. we then obtain

COROLLARY 1. 2. *Let $\Phi(t)$ be the characteristic function belonging to a* (0) *symmetrical* (0) *unimodal distribution function and $\{p_r\}$ $(r=0, 1, 2, \ldots)$ be some probability distribution of generating function $g(z)$. Then*

$$\sum_{r=0}^{\infty} p_r[\Phi(t)]^r = g[\Phi(t)]$$

is the characteristic function belonging to a (0) *symmetrical* (0) *unimodal distribution function.*

REMARK 1. It is easy to see that if the distribution function belonging to $\{p_r\}$ is infinitely divisible then the distribution function belonging to $g[\Phi(t)]$ is also infinitely divisible.

Hence — taking into account that the geometric distribution is infinitely divisible (see e.g. [4], p. 89) — one obtains e.g. that if $g(t)$ is the characteristic function of a (0) unimodal (0) symmetrical distribution function then the function $f(t)=\dfrac{p-1}{p-g(t)}$ $(p>1)$ investigated first by E. LUKÁCS (see e.g. [4], p. 216) is the characteristic function of a (0) unimodal (0) symmetrical infinitely divisible distribution function.

REMARK 2. Corollary 1. 2. is based on the theorem of A. WINTNER referred to. Then the *problem* arises: Are there (0) unimodal *non-symmetrical* distribution functions the convolution powers of which remain (0) unimodal? Obviously the assertion of Corollary 1. 2 would be true also if $\Phi(t)$ were the characteristic function belonging to such a (0) unimodal non-symmetrical distribution function.

As to a continuous mixture of type (1. 2) a theorem analogous to Corollary 1. 2 will be given in Part 3.

2.

Up to the present the following infinitely divisible distribution functions have been proved to be unimodal (cf. [5] and [6]):

1. all stable distribution functions (see [7]),
2. all symmetrical distribution functions from the class $\mathscr{L}$ of distribution functions (see [2], p. 847; as to the definition, see [8], p. 149).

Let the characteristic function $\varphi(t)$ of an infinitely divisible distribution function $F(x)$ have the P. LÉVY canonical representation

$$\varphi(t) = \exp\left[i\gamma t - \frac{\sigma^2 t^2}{2} + \int_{-\infty}^{-0}\left(e^{itu} - 1 - \frac{itu}{1+u^2}\right) dM(u) + \right.$$

$$\left. + \int_{+0}^{\infty}\left(e^{itu} - 1 - \frac{itu}{1+u^2}\right) dN(u)\right]$$

where γ and $\sigma^2 \geqq 0$ are real constants, $M(u)$ and $N(u)$ are non-decreasing in the intervals $(-\infty, 0)$ and $(0, \infty)$, respectively, $M(-\infty) = N(\infty) = 0$ and $\int_{-\varepsilon}^{0} u^2 dM(u) +$ $+ \int_{0}^{\varepsilon} u^2 dN(u) < \infty$ for every finite $\varepsilon > 0$ (cf. [8], p. 84). From a theorem of P. BRAUMANN [9] it follows that $F(x)$ will be (γ) symmetrical if and only if for every $u > 0$ such that u is a continuity point of $N(u)$ and $-u$ is a continuity point of $M(u)$ one has $M(u) = -N(-u)$.

Consequently the canonical representation of the characteristic function $\psi(t)$ of a (γ) *symmetrical* infinitely divisible distribution function will have the form

$$\psi(t) = \exp\left[i\gamma t - \frac{\sigma^2 t^2}{2} + \int_{-\infty}^{-0}\left(e^{itu} - 1 - \frac{itu}{1+u^2}\right) dM(u) - \right.$$

$$\left. - \int_{+0}^{\infty}\left(e^{itu} - 1 - \frac{itu}{1+u^2}\right) dM(-u)\right]$$

where γ and $\sigma^2 \geqq 0$ are real constants, $M(u)$ is non-decreasing in $(-\infty, 0)$, $M(-\infty) = 0$ and $\int_{-\varepsilon}^{0} u^2 dM(u) < \infty$ for every finite $\varepsilon > 0$.

Now let $\mathcal{M}$ denote the class of (γ) symmetrical infinitely divisible distribution functions $S(x)$ enjoying the property that in the canonical representation

$$\psi(t) = \exp\left[i\gamma t - \frac{\sigma^2 t^2}{2} + \int_{-\infty}^{-0}\left(e^{itu} - 1 - \frac{itu}{1+u^2}\right) dM(u) - \right.$$

(2.1)

$$\left. - \int_{+0}^{\infty}\left(e^{itu} - 1 - \frac{itu}{1+u^2}\right) dM(-u)\right]$$

of their characteristic functions $M(u)$ is convex in $(-\infty, 0)$.

Then we have the central

THEOREM 2. 1. *A* (γ) *symmetrical infinitely divisible distribution function belonging to* $\mathcal{M}$ *is* (γ) *unimodal.*

PROOF. It is based on some of the ideas of I. A. IBRAGIMOV and YU. V. LINNIK, utilized in proving the unimodality of the symmetrical stable distribution functions (see [10], p. 88). Let us define the functions $M_n(u)$ $(n=1, 2, 3, \ldots)$ by

$$M_n(u) = \begin{cases} M(u) & \left(u \leqq -\frac{1}{n}\right) \\ M'_+\left(-\frac{1}{n}\right)\left(u+\frac{1}{n}\right)+M\left(-\frac{1}{n}\right) & \left(-\frac{1}{n} < u < 0\right) \end{cases}$$

where $M'_+(u)$ denotes the right-hand derivative of $M(u)$ (it exists because $M(u)$ is convex). Evidently $M_n(u)$ is convex in $(-\infty, 0)$ and the function

$$\psi_n(t) =$$

$$= \exp\left[i\gamma t - \frac{\sigma^2 t^2}{2} + \int_{-\infty}^{-0}\left(e^{itu}-1-\frac{itu}{1+u^2}\right)dM_n(u) - \int_{+0}^{\infty}\left(e^{itu}-1-\frac{itu}{1+u^2}\right)dM(-u)\right] =$$

$$= \exp\left[i\gamma t - \frac{\sigma^2 t^2}{2} + 2M_n(-0) + \int_{-\infty}^{-0} e^{itu}\,dM_n(u) - \int_{+0}^{\infty} e^{itu}\,dM_n(-u)\right]$$

is of the type of (2. 1) i.e. it is the characteristic function of a symmetrical infinitely divisible distribution function $S_n(x)$. Let us put $2M_n(-0)=A_n$ and let us define the distribution function $H_n(x)$ by

$$H_n(x) = \begin{cases} \dfrac{M_n(x)}{A_n} & (x \leqq 0) \\ \dfrac{-M_n(x)+A_n}{A_n} & (x > 0). \end{cases}$$

$H_n(x)$ is (0) unimodal and it is (0) symmetrical, too. Then we have

$$\int_{-\infty}^{-0} e^{itu}\,dM_n(u) - \int_{+0}^{\infty} e^{itu}\,dM_n(-u) = A_n \int_{-\infty}^{\infty} e^{itu}\,dH_n(u).$$

Denoting the characteristic function $\int_{-\infty}^{\infty} e^{itu}dH_n(u)$ by $\chi_n(t)$ we have

$$\psi_n(t) = e^{i\gamma t} e^{-\frac{\sigma^2 t^2}{2}} \sum_{r=0}^{\infty} \frac{e^{-A_n} A_n^r}{r!} [\chi_n(t)]^r. \tag{2. 2}$$

By Corollary 1. 2 the sum on the right-hand side of (2. 2) is the characteristic

function of a (0) symmetrical (0) unimodal distribution function $G_n(x)$. Thus $S_n(x)$ is the convolution of $G_n(x)$ and $\int_{-\infty}^{\frac{x-\gamma}{\sigma}} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy$; then, by the theorem of A. WINTNER (see [2], p. 841), also $S_n(x)$ is (γ) unimodal and (γ) symmetrical.

Now let $n \to \infty$; then $M_n(u) \to M(u)$ and (cf. [8], p. 88) $\psi_n(t) \to \psi(t)$ i.e. $S_n(x) \to S(x)$ at every of its points of continuity. Hence it follows (cf. [8], p. 160) that $S(x)$ is also (γ) unimodal and (γ) symmetrical. This completes the proof.

REMARK 1. Evidently, all (γ) symmetrical stable distribution functions belong to $\mathcal{M}$.

REMARK 2. *There are distribution functions belonging to* $\mathcal{M}$ *which do not belong to* $\mathcal{L}$. For instance let us take $M(u) = e^u$ $(u < 0)$. This function satisfies the conditions of our theorem. However, it does not generate the characteristic function of an infinitely divisible distribution function belonging to the class $\mathcal{L}$ because $u\,M'(u) = ue^u$ $(u < 0)$ is *not* nonincreasing (cf. the characterization of the class $\mathcal{L}$ in [8], p. 149).

Finally we establish the useful

THEOREM 2. 2. *Let $\psi(t)$ be the characteristic function belonging to a (γ) symmetrical infinitely divisible distribution function belonging to $\mathcal{M}$. Then $[\psi(t)]^c$ $(c > 0)$ is the characteristic function of a $(c\gamma)$ symmetrical $(c\gamma)$ unimodal infinitely divisible distribution function.*

PROOF. $\psi(t)$ has the form (2. 1); $[\psi(t)]^c$ has the same form with $c\gamma$ and $cM(u)$ instead of γ and $M(u)$, resp. $cM(u)$ is also convex in $(-\infty, 0)$; hence $[\psi(t)]^c$ belongs to $\mathcal{M}$. Then, by Theorem 2. 1, our assertion follows.

3.

Now it is easy to obtain a generalization of Corollary 2. 1. We have

THEOREM 3. 1. *Let $\psi(t)$ be the characteristic function of a (0) symmetrical infinitely divisible distribution function belonging to $\mathcal{M}$ and let $B(x)$ be a disribution function which is identically 0 for $x \leqq 0$. Then*

$$\omega(t) = \int_0^\infty [\psi(t)]^x dB(x) = L[-\log \psi(t)],$$

where $L(s)$ denotes the Laplace—Stieltjes transform of $B(x)$, is the characteristic function of a (0) unimodal (0) symmetrical distribution function.

PROOF. Let $B_n(x)$ be a sequence of step functions converging to $B(x)$ as $n \to \infty$; let $b_r^{(n)}$ be the rth jump of $B_n(x)$ taking place at $x = a_r$ $(r = 0, 1, 2, \ldots)$. Let us introduce the function $\omega_n(t) = \int_0^\infty [\psi(t)]^x dB_n(x) = \sum_{r=0}^{\infty} b_r^{(n)} [\psi(t)]^{a_r}$. By Theorem 2. 2 $[\psi(t)]^{a_r}$

is the characteristic function of a (0) unimodal (0) symmetrical infinitely divisible distribution function. By Corollary 1. 1 $\omega_n(t)$ is a characteristic function of the same type as $[\psi(t)]^{a_r}$. By a well-known theorem of E. HELLY,

$$\lim_{n\to\infty} \omega_n(t) = \int_0^\infty [\psi(t)]^x \, d[\lim_{n\to\infty} B_n(x)] = \int_0^\infty [\psi(t)]^x \, dB(x)$$

and it is again a (real) characteristic function. By a convergence theorem (see [8], p. 160) $\omega(t)$ is the characteristic function of a (0) unimodal distribution function. Hence our assertion follows.

REMARK. It is easily seen (cf. [3], p. 538) that if $B(x)$ is infinitely divisible then the distribution function belonging to $\omega(t)$ is also infinitely divisible.

REFERENCES

[1] GIRAULT, M.: Les fonctions caractéristiques et leurs transformations, *Publ. Inst. Statist. Univ. Paris.* **4** (1955) 221—299.
[2] WINTNER, A.: Cauchy's stable distributions and an "explicit formula" of Mellin, *Amer. J. Math.* **78** (1956) 819—861.
[3] FELLER, W.: *An introduction to probability theory and its applications,* Volume II, Wiley and Sons, Inc., New York, 1966.
[4] LUKÁCS, E.: *Fonctions caractéristiques,* Dunod, Paris, 1964.
[5] LUKÁCS, E.: Recent developments in the theory of characteristic functions, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. II. Univ. of California Press, Berkeley and Los Angeles, 1961; pp. 307—335.
[6] FISZ, M.: Infinitely divisible distributions: Recent results and applications, *Ann. Math. Statist.* **33** (1962) 68—84.
[7] Ибрагимов, И. А. и Чернин, К. Е.: Об одновершинности устойчивых законов, *Теор. Вероятн. и Примен.*, **4** (1959) 453—456.
[8] GNEDENKO, B. V. and KOLMOGOROV, A. N.: *Limit distributions for sums of independent random variables,* Addison-Wesley Publ. Co., Inc., Cambridge, Mass., 1954.
[9] BRAUMANN, P.: Symmetrische unbeschraenkt teilbare Wahrscheinlichkeitsgesetze, *Rev. Fac. Ciencias Lisboa,* 2.A. Série, A- Ciências Matemàticas **7** (1959) 255—262.
[10] Ибрагимов, И. А. и Линник, Ю. В.: *Независимые и стационарно связанные величины,* Изд. „Наука", Москва, 1965.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received February 7, 1967.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 447—454.

ОДНА ПРЕДЕЛЬНАЯ ТЕОРЕМА В ЗАДАЧЕ ОБСЛУЖИВАНИЯ ПРИ НЕОГРАНИЧЕННО ВОЗРАСТАЮЩЕЙ ИНТЕНСИВНОСТИ ПОТОКА

J. TOMKÓ

1. Введение. В телефонном деле и других областях применения теории очередей возникает следующая задача: На однолинейную систему массового обслуживания поступает простейший поток требований. Максимальное число требований, могущих находиться в системе ограничено, т. е. имеется только конечное число мест для ожидания. Требования, пришедшие в систему, когда все места ожидания заняты, теряются. Преставляет интерес найти вероятности состояний системы, функции распределения времени ожидания и длительности занятого периода прибора. В случае, когда в системе имеется сколько угодно мест для ожидания, упомянутые характеристики системы широко изучены в работах [1], [2], [3]. Однако, как известно, в этом случае ситуация может быть интересной только при интенсивности входящего потока $\lambda \leqq 1$ (среднее значение времени обслуживания мы будем предполагать равной 1). В предлагаемом случае при любых значениях λ, искомые характеристики системы будут невырожденными. В частности, вероятности состояний системы достаточно обширно изучены в работе [4]. В большинстве случаев, закон распределения длительности рабочего периода системы не удается найти в явном виде. Интуитивно ясно, что при увеличении значения λ, это распределение сосредоточится на удаленной части положительной вещественной оси. Возникает вопрос, можно ли надлежащим образом нормировать длительность занятости прибора, чтобы переходя к пределу $\lambda \to \infty$, функция распределения нормированной величины приняла простой вид. Разыскание такого предельного соотношения является целью данной работы.

Перечислим обозначения и условия, которым будем придерживаться во всей работе. Число мест для ожидания выберем n. Моменты поступления требований обозначим через $t_1, t_2, \ldots, t_i, \ldots$, а промежутки между поступлениями $t_{i+1} - t_i$ через τ_i. Длительности обслуживания, которые будут обозначаться через $\xi_1, \xi_2, \ldots, \xi_i, \ldots$, будем считать совместно независимыми и одинаково распределенными по закону $F(x)$ с конечным средним значением $\mathsf{M}\xi_i = 1$. Каждый рабочий период системы вызывается поступлением некоторого требования, заставщего систему пустой. Если данный период занятости вызван поступлением i_1-го требования, то длительность этого периода будем обозначать через $\eta_{i_1}^{(n)}$ (n—указывает на то, что в системе может находиться всего $n+1$ требований). Элементы последовательности $\eta_{i_1}^{(n)}, \eta_{i_2}^{(n)}, \ldots, \eta_{i_k}^{(n)}, \ldots$ суть независимые и одинаково распределенные случайные величины. Их общий закон распределения обозначим через $G_n(x)$, а общее среднее значение через μ_n. Далее, будут приняты следующие обозначения для преобразований Лапласа-Стилтьеса: При $\operatorname{Re} s > 0$

$$\psi(s)=\int_0^\infty e^{-su}dF(u)$$

$$\Phi_n(s)=\int_0^\infty e^{-su}dG_n(u)$$

2. Основное рекуррентное соотношение. В этом пункте, для нахождения структуры случайной величины $\eta^{(n)}$ воспользуемся рассуждениями типа приведенного в [1] (стр. 61). Прежде всего заметим, что распределение длительности занятого периода системы не зависит от порядка обслуживания требований. Поэтому не нарушая общности, можем предполагать, что если требование поступает на систему, когда она пуста, то прибор немедленно начинает его обслуживание, а в дальнейшем, освободившийся прибор переходит на обслуживание того требования, которое поступило на систему последним среди присуствующих в ней. После обслуживания первого требования с длительностью ξ, если за это время поступило ν новых требований ($\nu \leqq n$, случайная величина) система ведет себя как система с $n-\nu+1$ местом ожидания. Пусть $\delta^{(n-\nu+1)}$ обозначает длительность рабочего периода такой системы, имеющая функцию распределения $G_{n-\nu+1}(x)$. Точно также, в момент, когда j-ое из поступивших за время ξ требований ($j \leqq \nu$) включается в процесс обслуживания, начинается новый период занятости $\delta^{(n-\nu+j)}$ системы с $n-\nu+j$ местом ожидания с функцией распределения $G_{n-\nu+j}(x)$.

Ясно, что случайные величины $\delta^{(n-\nu+1)}, \ldots, \delta^{(n)}$ независимые и $\eta^{(n)}$ удовлетворяет „рекуррентному" соотношению

$$\eta^{(n)}=\xi+\sum_{j=1}^{\nu}\delta^{(n-\nu+j)} \tag{1}$$

где ν является целочисленной ($0 \leqq \nu \leqq n$) случайной величиной.

3. Производящая функция средних значений μ_n. Пусть $\mu(z)=\sum_1^\infty \mu_n z^n$. Мы докажем предложение:

Теорема 1. При всех значениях $\lambda>0$ и $n \geqq 1$ математические ожидания μ_n конечны и их производящая функция представляется в виде

$$\mu(z)=\frac{z}{\psi(\lambda(1-z))-z} \tag{2}$$

Доказательство. Допустим, что рабочий период $\eta^{(n)}$ начинается в момент $t=0$. Обозначим через ζ_l сумму $\sum_{l n+1}^{l n+n}\xi_i$, а через E_l событие, состоящее в том, что за время ζ_l поступает по меньшей мере одно требование. Условимся взять E_0 достоверным событием. Пусть далее, ν^* случайная величина, принимающая значение k ($k \geqq 1$) на множестве $E_1E_2\ldots E_{k-1}\bar{E}_k$. Тогда, очевидно

$$\eta^{(n)} \leqq \sum_1^{\nu^*}\zeta_i. \tag{3}$$

Если теперь взять математическое ожидание обоих сторон (3), то можем воспользоваться тождеством Вальда и получаем, что

$$\mu_n \leqq n\mathsf{M}\nu^*$$

Так как

$$\mathsf{P}(\nu^*=k) = \left(\int_0^\infty (1-e^{-\lambda x})\,dF_n(x)\right)^{k-1}\left(1-\int_0^\infty (1-e^{-\lambda x})\,dF_n(x)\right)$$

$$= r^{k-1}(1-r),$$

где $F_n(x)=F^{*(n)}(x)$, то следует $\mathsf{M}\nu^* = 1\dfrac{1}{1-r} < \infty$ и конечность μ_n показана.

Чтобы определить производящую функцию $\mu(z)$, найдем связь между величинами $\mu_1, \mu_2, \ldots \mu_n$ с помощью рекуррентного соотношения (1). Беря математическое ожидание обоих сторон (1) находим, что

$$\mu_n = \sum_{k=0}^{n} \mathsf{M}(\xi_i/\nu=k)\mathsf{P}(\nu=k) + \sum_{k=1}^{n} (\mu_n + \ldots + \mu_{n-k+1})\mathsf{P}(\nu=k),$$

откуда вытекает рекуррентная связь

$$\mu_n = 1 + r_1\mu_n + \ldots + r_n\mu_1, \tag{4}$$

где r_s обозначает сумму вероятностей

$$p_j = \mathsf{P}(\nu=j) = \int_0^\infty \frac{(\lambda x)^j}{j!}\,dF(x) \qquad (j \geqq 0)$$

от s до ∞. Умножив обе стороны (4) на z^n, просуммируя по $n \geqq 1$ и изменив порядок суммирования на правой стороне, получаем:

$$\mu(z) = \frac{z}{1-z} + r_1\mu(z) + zr_2\mu(z) + \ldots + z^{i-1}r_i\mu(z) + \ldots$$

$$= \frac{z}{1-z} + \mu(z)[r_1 + r_2 z + \ldots + r_i z^{i-1} + \ldots]$$

Простыми вычислениями находим, что

$$\sum_1^\infty r_i z^{i-1} = \frac{1}{1-z}\left[1-\psi(\lambda(1-z))\right]$$

в силу которого получается представление (2), и этим Теорема I доказана.

Заметим одно асимптотическое свойство роста μ_n при $\lambda \to \infty$. Разложив функцию $\psi(\lambda(1-z))$ по степеням z, $\left(\psi(\lambda(1-z)) = \sum_0^\infty \dfrac{(-\lambda z)^i}{i!}\psi^{(i)}(\lambda)\right)$ потом умножив обе стороны (2) на разложенный вид $\psi(\lambda(1-z))-z$, после сравнения

коэффициентов двух сторон получаем, что

$$\mu_n\psi(\lambda)+\mu_{n-1}\frac{(-\lambda)}{1!}\psi^{(1)}(\lambda)+\ldots+\mu_{k-i}\frac{(-\lambda)^i}{i!}\psi^{(i)}(\lambda)+\ldots+ \tag{5}$$

$$+\mu_1\frac{(-\lambda)^{n-1}}{(n-1)!}\psi^{(n-1)}(\lambda)-\mu_{n-1}=0.$$

Очевидно, что для $k<n$ $\dfrac{\mu_k}{\mu_n}<1$ и что при $\lambda\to\infty$

$$\frac{(-\lambda)^i}{i!}\psi^{(i)}(\lambda)=\int_0^\infty\frac{(\lambda x)^i}{i!}e^{-\lambda x}dF(x)\to 0 \qquad (i\leqq n).$$

Поэтому, если делить (5) на μ_{n-1}, то мы находим асимптотическое равенство

$$\frac{\mu_n}{\mu_{n-1}}\psi(\lambda)\sim 1, \quad \text{при} \quad \lambda\to\infty. \tag{6}$$

Легко видеть, что

$$\mu_1\psi(\lambda)\sim 1, \quad \text{при} \quad \lambda\to\infty.$$

откуда выводим, что при любом фиксированном $n\geqq 1$

$$\mu_n\sim[\psi(\lambda)]^{-n}, \quad \text{при} \quad \lambda\to\infty. \tag{7}$$

4. Предельная теорема для $G_n(x)$ (при $\lambda\to\infty$). В начале находим соотношение между преобразованиями $\Phi_n(s)$ $(n\geqq 1)$. Введем обозначения

$$r_i(u)=e^{-\lambda u}\sum_{j=i}^{\infty}{}'\frac{(\lambda u)^j}{j!}, \qquad (i\geqq 0)$$

$$G_i^{(n)}(x)=\begin{cases}G_n * G_{n-}1 * \ldots * G_{n-i+1}(x), & x\geqq 0\\ 0, & x<0\end{cases} \qquad (1\leqq i\leqq n)$$

и

$$G_0^{(n)}(x)=\begin{cases}1, & \text{при} \quad x\geqq 0\\ 0, & \text{при} \quad x<0.\end{cases}$$

Из рекуррентного соотношения (1) стандартным путем выводится интегральное уравнение

$$G_n(x)=\int_0^x\sum_{i=0}^{n-1}{}'\frac{(\lambda u)^i}{i!}e^{-\lambda u}G_i^{(n)}(x-u)dF(u)+\int_0^x r_n(u)G_n^{(n)}(x-u)dF(u). \tag{8}$$

Введя в рассмотрение преобразования

$$\Phi_i^{(n)}(s)=\int_0^\infty e^{-su}dG_i^{(n)}(u)=\begin{cases}\Phi_n(s)\ldots\Phi_{n-i+1}(s), & \text{если} \quad n\geqq i\geqq 1\\ 1, & \text{если} \quad i=0.\end{cases}$$

соотношение (8) для преобразований $\Phi_n(s)$ примет вид

$$(9) \qquad \Phi_n(s) = \sum_{i=0}^{n} \Phi_i^{(n)}(s) \frac{(-\lambda)^i}{i!} \frac{d^i}{ds^i} \psi(s+\lambda) + \Phi_n^{(n)}(s) \sum_{i=n+1}^{\infty} \frac{(-\lambda)^i}{i!} \frac{d^i}{ds^i} \psi(s+\lambda).$$

Как видно, в соотношение (9) для $\Phi_n(s)$ входят все преобразования $\Phi_i(s)$ с номером $1 \leqq i \leqq n$. Это обстоятельство сильно усложняет нахождение $\Phi_n(s)$ в явном виде. Даже в случае показательного распределения $F(x) = 1 - e^{-x}$, который наиболее часто поддается явным вычислениям, трудно это сделать. Обратим теперь внимание на следующее обстоятельство. Допустим, что время рабочего периода системы достигло уже значение T, и интересуемся его дальнейшей длительностью, которую обозначим через $\eta^{(n)}(T)$. Ясно, что на величну $\eta^{(n)}(T)$ большое влияние имеют события, в будущем поступающие в входящем потоке. Кроме того, чем более вероятно поступление за короткий срок времени события входящего потока, тем сильнее становится его влияние и достигнутое время T перестает играть роль в распределении величины $\eta^{(n)}(T)$. Этот факт подсказывает нам нижедоказанное предложение.

Теорема 2. При любом $n \geqq 1$

$$\lim_{\lambda \to \infty} \mathrm{P}\left\{\frac{\eta_i^{(n)}}{\mu_n} < t\right\} = 1 - e^{-t}.$$

Доказательство. Нам достаточно показать, что

$$\lim_{\lambda \to \infty} \Phi_n\left(\frac{s}{\mu_n}\right) = \frac{1}{1+s}.$$

Для этого приводим соотношение (9) к виду

$$\Phi_n(s)\left[1 - \sum_{i=1}^{n} \Phi_{i-1}^{(n-1)}(s) \frac{(-\lambda)^i}{i!} \frac{d^i}{ds^i} \psi(s+\lambda) - \Phi_{n-1}^{(n-1)}(s) \sum_{i=n+1}^{\infty} \frac{(-\lambda)^i}{i!} \frac{d^i}{ds^i} \psi(s+\lambda)\right] =$$

$$= \psi(s+\lambda).$$

Займемся теперь выражением, стоящим тут в скобках. Мы имеем ряд равенств,

$$1 - \sum_{i=1}^{n} \Phi_{i-1}^{(n-1)}(s) \frac{(-\lambda)^i}{i!} \frac{d^i}{ds^i} \psi(s+\lambda) - \Phi_{n-1}^{(n-1)}(s) \sum_{i=n+1}^{\infty} \frac{(-\lambda)^i}{i!} \frac{d^i}{ds^i} \psi(s+\lambda) =$$

$$= 1 - \sum_{i=1}^{n} \Phi_{i-1}^{(n-1)}(s) \frac{(-\lambda)^i}{i!} \frac{d^i}{ds^i} \psi(s+\lambda) - \Phi_{n-1}^{(n-1)}(s)\left[\psi(s) - \sum_{0}^{\infty} \frac{(-\lambda)^i}{i!} \frac{d^i}{ds^i} \psi(s+\lambda)\right] =$$

$$= 1 - \Phi_{n-1}^{(n-1)}(s)\psi(s) + \Phi_{n-1}^{(n-1)}(s)\psi(s+\lambda) + \sum_{i=1}^{n} \left(\Phi_{n-1}^{(n-1)}(s) - \Phi_{i-1}^{(n-1)}(s)\right) \frac{(-\lambda)^i}{i!} \frac{d^i}{ds^i} \Psi(s+\lambda) =$$

$$= 1 - \psi(s) - \psi(s)[\Phi_{n-1}^{(n-1)}(s) - 1] + \Phi_{n-1}^{(n-1)}(s)\psi(s+\lambda) -$$

$$- \sum_{i=1}^{n} \left(\Phi_{n-1}^{(n-1)}(s) - \Phi_{i-1}^{(n-1)}(s)\right) \frac{(-\lambda)^i}{i!} \frac{d^i}{ds^i} \psi(s+\lambda).$$

Теперь заметим, что из (7) следует для любого $k<n$,

$$\frac{M\eta^{(k)}}{\mu_n}=\frac{\mu_k}{\mu_n}\to 0 \quad \text{при} \quad \lambda\to\infty,$$

поэтому для всех $k<n$

$$Me^{-\frac{s\eta^{(k)}}{\mu_n}}=\Phi_k\left(\frac{s}{\mu_n}\right)\to 1 \quad \text{при} \quad \lambda\to\infty.$$

Далее, из (6) и (7) получаем, что

$$\lim_{\lambda\to\infty}\frac{1-\psi\left(\frac{s}{\mu_n}\right)}{\psi\left(\frac{s}{\mu_n}+\lambda\right)}=\lim_{\lambda\to\infty}\frac{1-\psi\left(\frac{s}{\mu_n}\right)}{\frac{s}{\mu_n}}\lim_{\lambda\to\infty}\frac{s}{\mu_n\psi\left(\frac{s}{\mu_n}+\lambda\right)}=\lim_{\lambda\to\infty}\frac{s}{\mu_{n-1}}=0$$

Аналогичным же образом находим, что пределы, при $i\geqq 2$,

$$\varphi_i=\lim_{\lambda\to\infty}\frac{1}{\psi\left(\frac{s}{\mu_n}+\lambda\right)}\left[\Phi_{i-1}^{(n-1)}\left(\frac{s}{\mu_n}\right)-1\right]=$$

$$=-\lim_{\lambda\to\infty}\left[\mu_{n-1}+\ldots+\mu_{n-i+1}\right]\frac{s}{\mu_n\psi\left(\frac{s}{\mu_n}+\lambda\right)}=$$

$$=-\lim_{\lambda\to\infty}\frac{\mu_{n-1}+\ldots+\mu_{n-i+1}}{\mu_{n-1}}=-1\,,$$

а для $i=1$, $\varphi_1=0$. Отсюда, в частности, получаем, что

$$\lim_{\lambda\to\infty}\frac{1}{\psi\left(\frac{s}{\mu_n}+\lambda\right)}\left[\Phi_{n-1}^{(n-1)}\left(\frac{s}{\mu_n}\right)-\Phi_{i-1}^{(n-1)}\left(\frac{s}{\mu_n}\right)\right]=0.$$

Рассмотрев соотношение

$$\Phi_n\left(\frac{s}{\mu_n}\right)=\left[\frac{1-\psi\left(\frac{s}{\mu_n}\right)}{\psi\left(\frac{s}{\mu_n}+\lambda\right)}-\frac{\psi\left(\frac{s}{\mu_n}\right)}{\psi\left(\frac{s}{\mu_n}+\lambda\right)}\left(\Phi_{n-1}^{(n-1)}\left(\frac{s}{\mu_n}\right)-1\right)+\right.$$

$$\left.+1-\frac{1}{\psi\left(\frac{s}{\mu_n}+\lambda\right)}\sum_{i=1}^{n}\left(\Phi_{n-1}^{(n-1)}\left(\frac{s}{\mu_n}\right)-\Phi_{i-1}^{(n-1)}\left(\frac{s}{\mu_n}\right)\right)\frac{(-\lambda)^i}{i!}\frac{d^i}{ds^i}\psi\left(\frac{s}{\mu_n}+\lambda\right)\right]^{-1}$$

и замечая, что $\frac{(-\lambda)^i}{i!}\frac{d^i}{ds^i}\psi\left(\frac{s}{\mu_n}+\lambda\right)\to 0$ при $\lambda\to\infty$, немедленно находим нуж-

ное нам равенство

$$\lim_{\lambda\to\infty}\Phi_n\left(\frac{s}{\mu_n}\right)=\frac{1}{1+s}.$$

На этом доказательство теоремы 2 завершено.

Замачание. В силу (7) нормирующую константу μ_n в теореме 2, вычисление которой в большинстве случаев является трудоемким, можно заменить на $[\psi(\lambda)]^{-n}$.

5. Примеры. Разберем два частных случая распределения $F(x)$.

а) $F(x)=1-e^{-x}$. Производящая функция $\mu(z)$ в этом случае принимает вид

$$\mu(z)=\frac{z}{\dfrac{1}{1+\lambda(1-z)}-z}=\frac{z\big(1+\lambda(1-z)\big)}{1-(1+\lambda)z+\lambda z^2},$$

Корни знаменателя $z_1=1$, $z_2=1/\lambda$, и $\mu(z)$ можем привести к виду

$$\mu(z)=\frac{1}{\lambda-1}z[1+\lambda(1-z)]\left(\frac{1}{z-1}-\frac{\lambda}{\lambda z-1}\right)=$$

$$=\frac{1}{\lambda-1}z[1+\lambda(1-z)]\left(-\sum_1^\infty z^k+\lambda\sum_0^\infty(\lambda z)^k\right).$$

Откуда находим, что

$$\mu_n=1+\lambda+\lambda^2+\ldots+\lambda^n.$$

Для $\lambda<1$, μ_n при $n\to\infty$ должен сходится к среднему значению рабочего периода той системы, в которой имеется сколько угодно мест для ожидания. Действительно, мы получаем

$$\lim_{n\to\infty}\mu_n=\frac{1}{1-\lambda},$$

что согласуется хорошо известным значением математического ожидания упомянутого рабочего периода.

б)
$$F(x)=\begin{cases}1 & \text{при } x\geqq 1\\ 0 & \text{при } x<1.\end{cases}$$

Теперь $\mu(z)$ представляется в виде

$$\mu(z)=\frac{z}{e^{-\lambda(1-z)}-z}=\frac{ze^{\lambda(1-z)}}{1-ze^{\lambda(1-z)}}$$

При $|ze^{\lambda(1-z)}|<1$

$$\mu(z)=ze^{\lambda(1-z)}\sum_{k=0}^\infty z^k e^{\lambda k(1-z)}=\sum_{k=0}^\infty z^{k+1}e^{\lambda(k+1)}\sum_{j=0}^\infty\frac{[-\lambda(k+1)z]^j}{j!}.$$

Сосчитая коеффициенты при z^n последней суммы, находим формулу

$$\mu_n=\sum_{j=0}^{n-1}e^{\lambda(n-j)}\frac{[-\lambda(n-j)]^j}{j!}=e^{\lambda n}\sum_{j=0}^{n-1}(-1)^j\frac{[\lambda(n-j)]^j}{j!}e^{-\lambda j}.$$

Асимптотика (7) для этого случая получается непосредственно из этой же формулы. Изучать поведение μ_n ростом n при $\lambda<1$ теперь немного сложнее чем в предыдущем случае. При $\lambda<1$ ряд $\mu(z)$ сходится в круге $|z|<1$, и чтобы найти асимптотическое поведение μ_n ростом n, можем воспользоватся теоремой типа Таубера. Для этого найдем порядок роста функции

$$\frac{z}{e^{-\lambda(1-z)}-z} \quad \text{при} \quad z\to 1-0.$$

Получаем

$$\frac{z}{e^{-\lambda(1-z)}-z}\sim\frac{1}{1-\lambda}(1-z)^{-1} \quad \text{при} \quad z\to 1-0.$$

Отсюда по Тауберовой теореме [5] (теор. 4. 3. стр. 192) вытекает, что

$$\lim_{n-\infty}\frac{1}{n}\sum_{1}^{n}\mu_k=\frac{1}{1-\lambda},$$

а в силу монотонности μ_n, это равносильно предельному соотношению

$$\lim_{n\to\infty}\mu_n=\frac{1}{1-\lambda} \quad \text{при} \quad \lambda<1$$

и согласованность с известными результатами, о которой шла речь в предыдущем примере, для этого случая также имеет место.

В заключение мне хочется выразить благодарность моим коллегам М. Арато и Й. Пергелу, с которыми я имел возможность обсуждать проблемы, возникающие при выполнении данной работы.

БИБЛИОГРАФИЯ

[1] TAKÁCS, L.: *Introduction to the theory of queues,* New York, Oxford University Press 1962.
[2] KENDALL, D. G.: Stochastic processes occurring in the theory of quenes and their analisys by the method of imbedded Markov chain, *Ann. Math. Statist.* **24** (1953) 338—354.
[3] FINCH, P. D.: On the busy period in the queueing system G(GI) 1. *J. Austral. Math. Soc.* **2** (1961) 217.
[4] KEILSON, J.: The ergodic queue length distribution for queueing with finite capacity, *J. Roy. Statist. Soc.* B **28** (1966) 190—201.
[5] WIDDER, D. V.: *The Laplace transform,* Princeton, 1941.

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР АКАДЕМИИ НАУК ВЕНГРИИ, БУДАПЕШТ

(Поступило 20-ого февраля 1967 г.)

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 455—456.

ZUR ZWEISTUFIGEN SATZSTRUKTUR-GRAMMATIK

von

R. PÉTER

Wie ich von L. KALMÁR erfahren habe, wurde unter den Vorschlägen bezüglich der Weiterentwicklung der algorithmischen Sprache ALGOL 60 auch eine solche Sprache aufgeworfen, welche durch unendlich viele Satzstruktur-Produktionen definiert wird, derart, daß diese Produktionen selber durch eine besondere Metasprache generiert werden.

Eine solche Sprache kann durch ein geordnetes Quintupel

$$(Z, M, P, V, K)$$

endlicher Mengen angegeben werden, wobei Z die Menge der Zeichen, M die Menge der Metazeichen, P die Menge der „Metaproduktionen", V die Menge der „Vorproduktionen" und K die Menge der „Kategorienamen" der Sprache ist. Zur Erklärung dieser Begriffe benutze ich die Benennungen „Kette" oder „Liste" für endliche Folgen gewisser Elemente, je nachdem diese Elemente ohne weiteres nacheinandergesetzt, oder durch Kommata getrennt werden; es werden aus den Elementen von Z Zeichenketten, und aus diesen Zeichenkettenlisten, ferner aus den Elementen von $Z \cup M$ „Mischketten", und aus diesen Mischkettenlisten gebildet. Dann hat (nach unwesentlichen Abänderungen) jede Metaproduktion die Form

$$m: \mu,$$

wobei m ein Metazeichen und μ eine Mischkette ist; die Anwendung einer solchen Metaproduktion auf eine rechtsseitige Mischkette bedeutet das Ersetzen darin eines Vorkommens von m durch μ; endlich viele nach einander ausgeführte Anwendungen von Metaproduktionen generieren „Entwicklungen" der Metazeichen. Ferner hat jede Vorproduktion die Form

$$\mu: \Lambda,$$

wobei μ eine Mischkette und Λ eine Mischkettenliste ist (dabei heißt μ die „linke Seite", Λ die „rechte Seite" der Vorproduktion).

Das geordnete Tripel

$$(Z, M, P)$$

ergibt eine kontextunabhängige Satzstruktur-Grammatik für die Metasprache, wobei sämtliche (der endlich vielen) Metazeichen als ausgezeichnet gelten. Durch diese Grammatik werden im allgemeinen unendlich viele „terminale" (d.h. keine Metazeichen enthaltende) Entwicklungen für je ein Metazeichen generiert; diese nenne ich kurz die „Werte" der betreffenden Metazeichen.

Nun erhält man aus den Vorproduktionen durch Einsetzen dieser Werte für die Metazeichen die (im allgemeinen unendlich vielen) Produktionen, welche die betrachtete Sprache erzeugen. Genauer: Die linken Seiten der so entstehenden Produktionen sind Zeichenketten — diese nenne ich „Kategorienamen" — und die rechten Seiten sind Zeichenkettenlisten — jene Glieder dieser Listen, welche in keiner der Produktionen als linke Seiten auftreten, nenne ich „terminale Begriffe". Durch Anwendungen der Produktionen kommt man zu „Entwicklungen" der Kategorienamen; sind in einer Entwicklung schon alle Glieder terminale Begriffe, so heißt diese eine „terminale Entwicklung" des betreffenden Kategorienamens. Die Menge sämtlicher terminaler Entwicklungen eines Kategorienamens ergibt die durch diesen Namen bezeichnete „Kategorie". Für eine Sprache sind nur endlich viele Kategorien von Bedeutung; die Namen dieser bilden die Menge *K*. Die „Sprache" selbst besteht aus ihren Kategorien.

Es erhebt sich die Frage, ob die Einführung einer solchen zweistufigen Satzstruktur-Grammatik eine prinzipielle Notwendigkeit ist. Da nur endlich viele Kategorien in Betracht kommen, könnte man denken, daß sich vielleicht jede derart generierte Sprache auch einstufig, durch endlich viele Produktionen generieren läßt.

Dies kann ich aber durch ein einfaches Beispiel widerlegen, wobei die Menge der terminalen Begriffe in der entsprechenden Wortemenge primitiv-rekursiv ist. Sogar mit endlich vielen terminalen Begriffen ergibt sich ein ähnliches Gegenbeispiel, falls auch in den Metaproduktionen nicht nur Mischketten, sondern auch Mischkettenlisten als rechte Seiten zugelassen werden.

Beschränkt man sich aber in der ursprünglichen Definition auf solche zweistufig generierte Sprachen, welche nur endlich viele terminale Begriffe enthalten (möglicherweise wird die Weiterentwicklung vom ALGOL 60 zu einer solchen Sprache führen), so kann ich beweisen, daß diese sich tatsächlich auch einstufigs durch endlich viele Produktionen generieren lassen.

Die Ausarbeitung der genannten Beweise (nebst den exakten Definitionen) reiche ich zur selben Zeitschrift ein.

EÖTVÖS L. UNIVERSITÄT, BUDAPEST

(Eingegangen: 14. März, 1967)

Zusatz bei der Korrektur: Die genaue Ausarbeitung ist am 29. März, 1967. eingegangen.

Printed in Hungary

A kiadásért felel az Akadémiai Kiadó igazgatója – Műszaki szerkesztő: Farkas Sándor
A kézirat nyomdába érkezett: 1967. IV. 12. – Terjedelem: 16,75 (A/5) ív, 17 ábra

67-5595 Szegedi Nyomda

Die *Studia Scientiarum Mathematicarum Hungarica* ist eine Halbjahrsschrift der Ungarischen Akademie der Wissenschaften. Sie veröffentlicht Originalbeiträge aus dem Bereiche der Mathematik in deutscher, englischer, französischer oder russischer Sprache. Es erscheint jährlich ein Band.

Adresse der Redaktion: Budapest V., Reáltanoda u. 13—15, Ungarn.
Technischer Redaktor: Gy. Katona.

Abonnementspreis pro Band (pro Jahr): 165.— Ft. Bestellbar bei Buch- und Zeitungs-Aussenhandelsunternehmen *Kultúra* (Budapest 62, P.O.B. 149), oder bei den Vertretungen im Ausland.
Austauschabmachungen können mit der Bibliothek des Mathematischen Instituts (Budapest V., Reáltanoda u. 13—15) getroffen werden.
Die zur Veröffentlichung bestimmten Manuskripte sind in zwei Exemplaren an die Redaktion zu schicken.

Studia Scientiarum Mathematicarum Hungarica est une revue biannuelle de l'Académie Hongroise des Sciences publiant des essais originaux, en francais, anglais, allemand ou russe, du domaine des mathématiques.

Rédaction: Budapest V. Reáltanoda u. 13—15, Hongrie.
Rédacteur technique: Gy. Katona

Le prix de l'abonnement: 165 Forints par an (volume). On s'abonne chez *Kultúra*. Société pour le Commerce de Livres et Journaux (Budapest 62, P.O.B. 149) ou chez ses représentantsàt l.étranger.
Pour établir des relations d'échange on est prié de s'adresser à la Biblioth que de l'Institut de Mathématique (Budapest V., Reáltanoda u. 13—15).
Onfest prié d'envoyer les articles destinés à la publication en deux exemplaires à l'adresse de la rédaction.

Studia Scientiarum Mathematicarum Hungarica — выходит два раза в год в издании Академии наук Венгрии. Журнал публикует оригинальные исследования в области математики на немецком, английском, французском и русском языках. Отдельные выпуски составляют ежегодно один том.

Адрес редакции: Budapest V., Reáltanoda u. 13—15, Венгрия.
Технический редактор: Gy. Katona.

Подписная цена на год (за один том): 165 форинтов. Подписка на журнал принимается Внешнеторговым предприятием „Культура" (Budapest 62, P. O. B. 149) или его представительствами за границей.
По поводу отношения обмена просим обращаться к Библиотеке Института Математики (Budapest V., Reáltanoda u. 13—15).
Работы, предназначенные для опубликования в журнале следует направлять по адресу редакции в двух экземплярах.

Ára: 72.— Ft

All the reviews of the Hungarian Academy of Sciences may be obtained among others from the following bookshops:

ALBANIA

Ndermarja Shtetnore e Botimeve
Tirana

AUSTRALIA

A. Keesing
Box 4886, GPO
Sidney

AUSTRIA

Globus Buchvertrieb
Salzgries 16
Wien I.

BELGIUM

Office International de Librairie
30, Avenue Marnix
Bruxelles 5
Du Monde Entier
5, Place St. Jean
Bruxelles

BULGARIA

Raznoiznos
1 Tzar Assen
Sofia

CANADA

Pannonia Books
2 Spadina Road
Toronto 4, Ont.

CHINA

Waiwen Shudian
Peking
P.O.B. Nr. 88.

CHECHOSLOVAKIA

Artia A. G.
Ve Smeckách 30
Praha II.
Postova Novinova Sluzba
Dovoz tisku
Vinohradska 46
Praha 2
Postova Novinova Sluzba
Dovoz tlace
Leningradska 14
Bratislava

DENMARK

Ejnar Munksgaard
Nörregade 6
Kopenhagen

FINLAND

Akateeminen Kirjakauppa
Keskuskatu 2
Helsinki

FRANCE

Office International de Documentation et Libraire
48, rue Gay Lussac
Paris 5

GERMAN DEMOCRATIC REPUBLIC

Deutscher Buch-export und Import
Leninstraße 16.
Leipzig C. I.
Zeitungsvertriebsamt
Clara Zetkin Straße 62.
Berlin N. W.

GERMAN FEDERAL REPUBLIC

Kunst und Wissen
Erich Bieber
Postfach 46.
7 Stuttgart S.

GREAT BRITAIN

Collet's' Subscription Dept.
44–45 Museum Streetf
London W.C.I.
Robert Maxwell and Co. Ltd.
Waynflete Bldg. The Plain
Oxford

HOLLAND

Swetz and Zeitlinger
Keizersgracht 471–487
Amsterdam C.
Martinus Nijhof
Lange Voorhout 9
The Hague

INDIA

Current Technical Literature
Co. Private Ltd.
Head Pffice:
India House OPP.
GPO Post Box 1374
Bombay I.

ITALY

Santo Vanasia
71 Via M. Macchi
Milano
Libreria Commissionaria Sansoni
Via La Marmora 45
Firenze

JAPAN

Nauka Ltd.
2 Kanada-Zimbocho 2-chome
Chiyoda-ku
Tokyo
Maruzen and Co. Ltd.

Tokyo

Far Eastern Booksellers
Kanada P.O. Box 72
Tokyo

KOREA

Chulpanmul
Korejskoje Obschestvo po Exportu i Importu Proizvedenij Pechati
Phenjan

NORWAY

Johan Grundt Tanum
Karl Johansgatan 43
Oslo

POLAND

Export und Import Unternehmen
RUCH
ul. Wilcza 46.
Warszawa

ROUMANIA

Cartimex
Str. Aristide Briand 14–18.
Bucuresti

SOVIET UNION

Mezhdunarodnaja Kniga
Moscow
G–200

SWEDEN

Almquist and Wiksell
Gamla Brogatan 26
Stockholm

USA

Stechert Hafner Inc.
31 East 10th Street
New York 3 N Y.
Walter J. Johnson
111 Fifth Avenue
New York 3 N. Y.

VIETNAM

Xunhasaba
Service d'Export et d'Import des Livres et Périodiques
19. Tran Quoc Toan
Hanoi

YUGOSLAVIA

Forum
Vojvode Misiva broj 1.
Novi Sad
Jugoslovenska Kniga
Terazije 27.
Beograd

31. X. 1967 Index: 26 748