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 V.
FASC. 1–2.
1970

AKADÉMIAI KIADÓ, BUDAPEST

Studia Scientiarum Mathematicarum Hungarica

A Magyar Tudományos Akadémia matematikai folyóirata

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

Technikai szerkesztő: Petruska 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.

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. Petruska

Subscription rate: $ 16.00 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

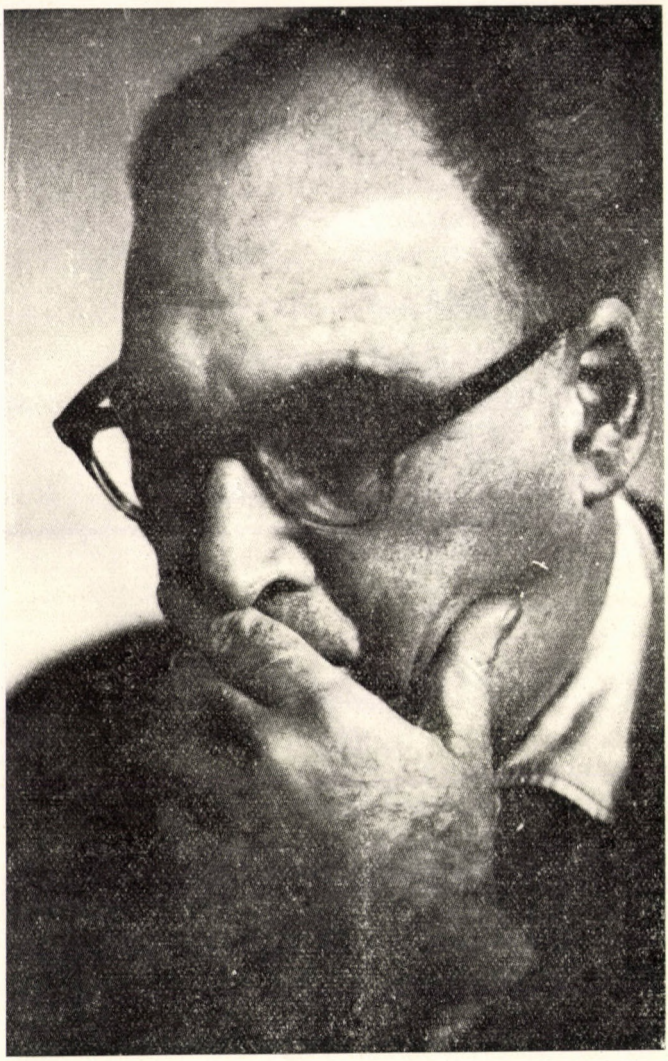
A. Rényi

Adiuvantibus

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

Tomus V

Akadémiai Kiadó, Budapest
1970



ALFRÉD RÉNYI
1921—1970

Alfréd Rényi, chief Editor of Studia Scientiarum Mathematicarum Hungarica, professor at the University Eötvös Loránd, member of the Hungarian Academy of Sciences, head of the Mathematical Institute, vice president of the Bolyai Mathematical Society, double winner of the Kossuth Prize, former vice president of the International Statistical Institute, editor of several journals and member of several scientific societies passed away at the age of 49 on 1 February 1970 after a short, but serious illness. He contributed a great deal to the development of many fields of mathematics, we are to remember his investigations in number theory, probability theory, mathematical statistics, graph theory and their applications.

His death is a great loss, that hit the Redaction of the Studia Scientiarum Mathematicarum Hungarica, our Academy and the whole Hungarian mathematical life. His papers and books earned him world-wide recognition in his early youth already. The style he used reflects his artistical mentality. All these and also his attractive personality created a probabilistic school in Hungary.

Alfréd Rényi was born on 20 march 1921 in Budapest. He studied at the University of Budapest, spending some time in Leningrad and Moscow as a postgraduate. He took his doctor's degree in 1947 and was appointed a professor in 1949 at the University of Debrecen and in 1952 at the University of Budapest which position he retained until his death. He became corresponding member of the Academy in 1949 and in 1956 a member. He was the head of Mathematical Institute from it's foundation in 1950. A list of his works will be published in the next number of our journal.

The Editorial Board

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 5—10

ON THE NUMBER OF ENDPOINTS OF A k-TREE

by

A. RÉNYI

§ 1. Definitions

A k-tree ($k=1, 2, 3, \ldots$) is the natural generalization to k dimensions of an ordinary (i.e. one-dimensional) tree (see [1], [2], [3], [4]). A k-tree can be considered either as a k-dimensional simplicial complex with certain properties, or as a graph. We shall take the second point of view.

A k-tree can be most conveniently defined inductively as follows: A k-tree of order $k+1$ is a complete $(k+1)$-graph. A k-tree $t_{(n+1)}^{(k)}$ of order $n+1$ ($n \geqq k+1$) is obtained by choosing an arbitrary k-tree $t_n^{(k)}$ of order n, and adding a new vertex, joining it to k such points of $t_n^{(k)}$ which form a complete graph in $t_n^{(k)}$, i.e. to the points of a $(k-1)$-cell of $t_n^{(k)}$. Thus a k-tree of order n contains n points (vertices), $n-k$ k-cells (i.e. complete subgraphs of order $k+1$) and $k(n-k)+1$ $(k-1)$-cells (i.e. complete subgraphs of order k).

A point of a k-tree of order n ($\geqq k+1$) is called an *endpoint* if it belongs to a single k-cell of the k-tree.

It is easy to see by induction that the number of endpoints of a k-tree of order $n \geqq k+1$ is at least 2 and at most $n-k$.

As a matter of fact two endpoints can not belong to the same $(k-1)$-cell if $n \geqq k+2$; thus if we take a k-tree $t_n^{(k)}$ of order $n \geqq k+2$ and form a k-tree $t_{n+1}^{(k)}$ of order $n+1$ by adding to $t_n^{(k)}$ a new point (joining it with the points of a $(k-1)$-cell of $t_n^{(k)}$) then the new point will be an endpoint, of $t_{n+1}^{(k)}$ and among the endpoints of $t_n^{(k)}$ at most one will not be an endpoint of $t_{n+1}^{(k)}$, thus the number of endpoints is either unchanged, or is increased by one; as a k-tree of order $k+2$ consists of two k-cells which have a $(k-1)$-cell in common and thus it contains exactly 2 endpoints, it follows that any k-tree of order $n \geqq k+2$ contains at least 2 and at most $n-k$ endpoints.

In this paper we consider *labeled*, more exactly: point-labeled k-trees, i.e. such k-trees of order n, the points of which are labeled by the numbers $1, 2, \ldots, n$. It has been shown (see [2] and [3]) that denoting by $T_k(n)$ the total number of labeled k-trees of order n, one has

$$T_k(n) = \binom{n}{k} [k(n-k)+1]^{n-k-2} \qquad (k = 1, 2, \ldots). \tag{1}$$

The aim of the present paper is to determine the number $T_k(n, r)$ of labeled k-trees of order n having r endpoints ($2 \leqq r \leqq n-k$). The corresponding problem or 1-trees has been solved in [5].

§ 2. Exact formulae

We prove first the following recursion formula

$$T_k(n+1, r)r = (n+1)[(k(n-k-r+1)+1)T_k(n, r-1)+rkT_k(n, r)] \tag{2}$$

where $2 \leqq r \leqq n+1-k$, $n \geqq k+2$.

To prove (2) let us mention that if we take any k-tree $t_n^{(k)}$ of order n having r endpoints, and join a new point to one of the $(k-1)$-cells of $t_n^{(k)}$ containing an endpoint of $t_n^{(k)}$, we get a k-tree $t_{n+1}^{(k)}$ of order $n+1$ with r endpoints, because one endpoint of $t_n^{(k)}$ disappears and one new endpoint is created. On the other hand if we take a k-tree $t_n^{(k)}$ of order n having $r-1$ endpoints and join a new point to one of the $(k-1)$-cells not containing any of the endpoints of $t_n^{(k)}$, we get a k-tree $t_{n+1}^{(k)}$ of order $n+1$ having r endpoints, because all the $r-1$ endpoints of $t_n^{(k)}$ will be endpoints of $t_{n+1}^{(k)}$, and the new point will also be an endpoint of $t_{n+1}^{(k)}$. If we do this with all k-trees of order n labeled by the numbers $1, 2, \ldots, n+1$ except j (where $j = 1, 2, \ldots, n+1$), we get all possible k-trees of order $n+1$ having r endpoints, each exactly r times; taking into account that each endpoint of a k-tree of order $n \geqq k+2$ belongs to k of its $k(n-k)+1$ $(k-1)$-cells, (2) follows.

Let us put now

$$t_k(n, s) = \frac{T_k(n, n-s)}{\binom{n}{s}}. \tag{3}$$

It follows that

$$t_k(n+1, s) = [k(s-k)+1]\cdot t_k(n, s)+kst_k(n, s-1) \quad \text{for} \quad k \leqq s \leqq n-1, \tag{4}$$

where $t_k(k+1, k) = 0$.
Thus, putting

$$G_k(z, s) = \sum_{n=s+2}^{+\infty} t_k(n, s)z^{n-s-2} \qquad (s = k, k+1, \ldots) \tag{5}$$

we obtain

$$G_k(z, s) = \frac{ksG_k(z, s-1)}{1-z[k(s-k)+1]}. \tag{6}$$

As however $T_k(n, n-k) = \binom{n}{k}$ if $n \geqq k+2$, we have

$$G_k(z, k) = \frac{1}{1-z}. \tag{7}$$

Thus it follows from (6) that

$$G_k(z, s) = \frac{s!\, k^{s-k}}{k! \prod_{j=0}^{s-k} (1-z(jk+1))}. \tag{8}$$

It follows in particular for $k=1$ that

$$G_1(z,s) = \frac{s!}{\prod_{h=1}^{s}(1-hz)}. \tag{9}$$

Now it is known (see [5]) that for $k=1$

$$T_1(n,n-s) = S(n-2,s)s!\binom{n}{s} \tag{10}$$

where $S(m,s)$ are the Stirling numbers of the second type, defined by

$$y^m = \sum_{s=1}^{m} S(m,s)y(y-1)\dots(y-s+1). \tag{11}$$

From (10) we can deduce (9) directly as follows.

It follows from (11) that

$$\sum_{m=s}^{\infty} \frac{S(m,s)z^m}{m!} = \frac{(e^z-1)^s}{s!} \tag{12}$$

and thus for $|z| < \frac{1}{s}$

$$\sum_{m=s}^{\infty} S(m,s)z^m = \sum_{m=s}^{\infty} \frac{S(m,s)z^m}{m!} \int_0^{\infty} y^m e^{-y}\,dy = \int_0^{\infty} \frac{(e^{zy}-1)^s}{s!} e^{-y}\,dy. \tag{13}$$

It follows by partial integration that

$$\sum_{m=s}^{\infty} S(m,s)z^m = \frac{z^s}{\prod_{h=1}^{s}(1-hz)} \tag{14}$$

i. e.

$$G_1(z,s) = s! \sum_{m=s}^{\infty} S(m,s)z^{m-s} = \frac{s!}{\prod_{h=1}^{s}(1-hz)} \tag{15}$$

in accordance with (9).

To get an explicit expression for $T_k(n,r)$ we need the following identity

$$\frac{1}{\prod_{j=0}^{m}[1-z(jk+1)]} = \frac{1}{m!\,k^m} \sum_{j=0}^{m} \binom{m}{j} \frac{(-1)^{m-j}(jk+1)^m}{1-z(jk+1)} \tag{16}$$

which can be proved e.g. by elementary function theory.

It follows from (8) and (16)

$$G_k(z,s) = \binom{s}{k} \sum_{j=0}^{s-k} \binom{s-k}{j} \frac{(-1)^{s-k-j}(jk+1)^{s-k}}{1-z(jk+1)} \tag{17}$$

and thus

$$t_k(n,s)=\frac{T_k(n,n-s)}{\binom{n}{s}}=\binom{s}{k}\sum_{j=0}^{s-k}\binom{s-k}{j}(-1)^{s-k-j}(jk+1)^{n-k-2}. \tag{18}$$

Thus we obtain finally

$$T_k(n,n-s)=\binom{n}{s}\binom{s}{k}\sum_{j=0}^{s-k}\binom{s-k}{j}(-1)^{s-k-j}(jk+1)^{n-k-2}. \tag{19}$$

By adding the values of $T_k(n,n-s)$ for $s=k,k+1,\ldots,n-2$ we must of course get the known formula (1) (see [2], [3]) for the total number $T_k(n)$ of k-trees of order n. As a matter of fact we get from (19)

$$T_k(n)=\sum_{s=k}^{n-2}T_k(n,n-s)= \tag{20}$$

$$=\binom{n}{k}\sum_{j=0}^{n-k-2}\binom{n-k}{j}(jk+1)^{n-k-2}(n-k-j-1)(-1)^{n-k-j}.$$

Taking into account that

$$\sum_{j=0}^{n-k-2}\binom{n-k}{j}(jk+1)^{n-k-2}(n-k-j-1)=$$

$$=\frac{1}{k}\left[(k(n-k-1)+1)H^{(n-k-2)}(0)-H^{(n-k-1)}(0)\right]$$

where

$$H(x)=e^x[(e^{kx}-1)^{n-k}+(n-k)e^{k(m-k-1)x}-e^{k(n-k)x}]$$

and we have

$$H^{(n-k-2)}(0)=(n-k)[k(n-k-1)+1]^{n-k-2}-[k(n-k)+1]^{n-k-2}$$

and

$$H^{(n-k-1)}(0)=(n-k)[k(n-k-1)+1]^{n-k-1}-[k(n-k)+1]^{n-k-1}$$

and thus

$$\frac{1}{k}\left[(k(n-k-1)+1)H^{(n-k-2)}(0)-H^{(n-k-1)}(0)\right]=[k(n-k)+1]^{n-k-2}$$

it follows that (1) holds.

Thus as a by-product we obtained another proof of the formula (1). Notice that for $k=1$ (1) reduces to CAYLEY's celebrated formula $T_1(n)=n^{n-2}$ for the total number of ordinary (one-dimensional) trees of order n. (For other proofs of this formula see [6] and [7].)

§ 3. The moments of the distribution of the number of endpoints

Let us first consider the mean value and the variance of the number of endpoints of a k-tree of order n. For the mean value

$$M_k(n) = \frac{1}{T_k(n)} \sum_{r=2}^{n-k} r T_k(n, r) \tag{21}$$

we get from the recursion formula (2) immediately

$$M_k(n) = \frac{n T_k(n-1)}{T_k(n)} [k(n-k-1)+1] \tag{22}$$

and thus

$$M_k(n) = \frac{n-k}{\left(1 + \frac{k}{k(n-k-1)+1}\right)^{n-k-2}} . \tag{23}$$

Thus we have

$$\lim_{n \to +\infty} \frac{M_k(n)}{n} = \frac{1}{e} \quad \text{for} \quad k = 1, 2, 3, \ldots . \tag{24}$$

Similarly we get from (2) for the variance

$$D_k^2(n) = \frac{1}{T_k(n)} \sum_{r=2}^{n-k} [r - M_k(n)]^2 T_k(n, r) \tag{25}$$

$$D_k^2(n) = \frac{(n-k-1)(n-k)[(n-k-2)k+1]^{n-k-3}}{[k(n-k)+1]^{n-k-2}} + \tag{26}$$

$$+ \frac{(n-k)[k(n-k-1)+1]^{n-k-3}}{[k(n-k)+1]^{n-k-2}} - \frac{(n-k)^2[k(n-k-1)+1]^{2n-2k-6}}{[k(n-k)+1]^{2n-2k-4}}$$

and thus

$$\lim_{n \to +\infty} \frac{D_k^2(n)}{n} = \frac{1}{e}\left(1 - \frac{2}{e}\right) \qquad (k = 1, 2, \ldots). \tag{2. 7}$$

It is remarkable that the asymptotic formulae for $M_k(n)$ and $D_k^2(n)$ do not depend on k, i.e. are the same as those obtained in [5] for $k=1$.

In [5] we have proved that if we choose at random one of the n^{n-2} labeled trees of order n, so that any one of these is chosen with the same probability, then denoting by v_n the number of endpoints of this random tree, the distribution of the random variable $\frac{v_n - M_1(n)}{D_1(n)}$ tends for $n \to +\infty$ to the standard normal distribution. By considering moments of every order one can prove that the same holds for every k, i.e. if we choose at random, with uniform distribution one of the $T_k(n)$ labeled k-trees of order n, and denote the number of its endpoints by $v_n^{(k)}$, then the distribution of $\frac{v_n^{(k)} - M_k(n)}{D_k(n)}$ tends for $n \to +\infty$ to the standard normal distribution.

REFERENCES

[1] HARARY, F. and PALMER, E. M.: On Acyclic simplicial complexes, *Mathematika* **15** (1968) 115—122.
[2] BEINEKE, L. W. and PIPPERT, R. E.: The number of labeled k-dimensional trees, *Journal of Combinatorial Theory* **6** (1969) 200—205.
[3] MOON, J. W.: The number of labeled k-trees, *Journal of Combinatorial Theory* **6** (1969) 196—199.
[4] PALMER, E. M.: On the number of labeled 2-trees, *Journal of Combinatorial Theory* **4** (1969) 206—207.
[5] RÉNYI, A.: Some remarks on the theory of trees, *Publ. Math. Inst. Hung. Acad. Sci.* **4** (1959) 73—85.
[6] MOON, J. W.: *Various proofs of Cayley's formula for counting trees*, Seminar on Graph Theory, ed. by F. Harary.
[7] RÉNYI, A.: On Cayley's polynomials for counting trees, *Proc. of the Calgary Int. Conference on Combinatorial Structures and their Applications* (in print).

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received August 1, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 11–15.

ОБ ОЦЕНКАХ ПАРАМЕТРОВ ПРОЦЕССОВ, УДОВЛЕТВОРЯЮЩИХ ЛИНЕЙНЫМ ДИФФЕРЕНЦИАЛЬНЫМ СТОХАСТИЧЕСКИМ УРАВНЕНИЯМ

M. ARATÓ

1. Рассмотрим многомерный стохастический процесс $\boldsymbol{\xi}^*(t) = (\xi_0(t), \xi_1(t), \ldots, \xi_{k-1}(t))$, удовлетворяющий уравнению

$$d\boldsymbol{\xi}(t) = A\boldsymbol{\xi}(t)dt + d\mathbf{w}(t) \tag{1. 1}$$

где $\mathbf{w}(t)$ является непрерывным мартингалом, $\mathsf{M}d\mathbf{w}(t) = 0$, $\mathsf{M}(d\mathbf{w}(t)d\mathbf{w}^*(t)) = B_w(t)dt$. Матрица $B_w(t)$ является симметричной и положительно определенной.

Предполагается, что элементы матрицы A являются неизвестными и оцениваются по реализации $\boldsymbol{\xi}(t)$ на отрезке $0 \leqq t \leqq T$.

Метод наименьших квадратов для оценки параметров $\{a_{pq}\} = A$ состоит в следующем. Рассмотрим функционал траектории $\xi(t)$

$$\int_0^T [B_w^{-1}(s)A\boldsymbol{\xi}(s), d\boldsymbol{\xi}(s)] - \frac{1}{2}\int_0^T [A\boldsymbol{\xi}(s), B_w^{-1}(s)A\boldsymbol{\xi}(s)]ds =$$

$$= \frac{1}{2}\int_0^T [B_w^{-1}(s)A\boldsymbol{\xi}(s), A\boldsymbol{\xi}(s)]ds + \int_0^T [B_w^{-1}(s)A\boldsymbol{\xi}(s), d\mathbf{w}(s)], \tag{1. 2}$$

и ищем ту матрицу $\hat{A}$, при которой (1. 2) является минимальным, т. е. ищем решения уравнений

$$\frac{\partial}{\partial a_{pq}}\left\{\int_0^T [B_w^{-1}(s)A\boldsymbol{\xi}(s), d\boldsymbol{\xi}(s)] - \frac{1}{2}\int_0^T [A\boldsymbol{\xi}(s), B_w^{-1}(s)A\boldsymbol{\xi}(s)]ds\right\} = 0; \tag{1. 3}$$

$$p, q = 0, 1, \ldots, k-1.$$

Здесь $[\mathbf{a}, \mathbf{b}]$ обозначает скалярное произведение векторов $\mathbf{a}$ и $\mathbf{b}$.

В том случае, когда $\mathbf{w}(t)$ является винеровским процессом $B_w(s) = B_w$, и если матрица A имеет характеристические значения λ_i с отрицательной действительной части, процесс $\boldsymbol{\xi}(t)$ является стационарным гауссовским марковским процессом. Условная функция правдоподобия (производное Радона-Никодима меры P_A относительно меры Винера W), при условии $\boldsymbol{\xi}(0) = \mathbf{x}$, имеет вид

$$\frac{dP_A}{dW}(\boldsymbol{\xi}(t)) = L = \exp\left\{\int_0^T [B_w^{-1}A\boldsymbol{\xi}(s), d\boldsymbol{\xi}(s)] - \frac{1}{2}\int_0^T [A\boldsymbol{\xi}(s), B_w^{-1}A\boldsymbol{\xi}(s)]ds\right\}, \tag{1. 4}$$

и метод условного наибольшего правдоподобия совпадает методом наименьших квадратов.

Пусть $B_w^{-1}(t)=\{b_{ij}^{-1}(t)\}$, тогда уравнение (1. 3) можно переписать в следующую форму

$$\int_0^T \xi_q(s)\sum_j b_{pj}^{-1}(s)d\xi_j(s)-\sum_{i,j}\hat{a}_{ji}\int_0^T b_{jp}^{-1}(s)\xi_q(s)\xi_i(s)ds=0;\quad p,q=0,1,\ldots,k-1,$$

или

$$\text{(1. 3')}\quad \int_0^T \xi_q(s)\sum_j b_{pj}^{-1}(s)\left(d\xi_j(s)-\sum_i \hat{a}_{ji}\xi_i(s)ds\right)=0;\quad p,q=0,1,\ldots,k-1.$$

Из уравнения (1. 1) следует, что

$$\text{(1. 5)}\quad \int_0^T \xi_q(s)\sum_j b_{pj}^{-1}(s)\left(d\xi_j(s)-\sum_i a_{ji}\xi(s)ds\right)=\int_0^T \xi_q(s)\sum_j b_{pj}^{-1}(s)dw_j(s);$$

$$p,q=0,1,\ldots,k-1.$$

Вычитая из (1. 5) уравнение (1. 3') получим, что

$$\frac{1}{T}\int_0^T \xi_q(s)\sum_j b_{pj}^{-1}(s)\sum_i \sqrt{T}(\hat{a}_{ji}-a_{ji})\xi_i(s)ds=\frac{1}{\sqrt{T}}\int_0^T \xi_q(s)\sum_j b_{pj}^{-1}(s)dw_j(s)=\eta_{pq}(T);$$

$$\text{(1. 6)}\qquad p,q=0,1,\ldots,k-1.$$

Моменты правой стороны (1. 6) легко сосчитать, а именно

$$\mathsf{M}\eta_{pq}(T)=0$$

$$\text{(1. 7)}\quad \mathsf{M}\eta_{pq}(T)\eta_{rs}(T)=\frac{1}{T}\int_0^T \mathsf{M}\xi_q(t)\xi_s(t)\sum_{j_1 j_2} b_{pj_1}^{-1}(t)b_{rj_2}^{-1}(t)b_{j_1 j_2}(t)dt=$$

$$=\frac{1}{T}\int_0^T b_{rp}^{-1}(t)\mathsf{M}\xi_q(t)\xi_s(t)dt,$$

при $p,q=0,1,\ldots,k-1$.

Во многих случаях выполняются следующие условия:

а) $$\frac{1}{T}\int_0^T b_{pj}^{-1}(t)\xi_q(t)\xi_i(t)dt\to b_{pjqi},\quad \text{при}\quad T\to\infty,$$

б) $\eta_{pq}(T)$ имеют асимптотически (при $T\to\infty$) нормальное распределение с матрицей вторых моментов $B=\{b_{pjqi}\}$.

Докажем следующую теорему, являющуюся обобщением известной теоремы Манн и Вальда в случае дискретного времени (см. Манн и Вальд [4]).

Теорема 1. *При условиях а), б) случайные величины* $\sqrt{T}(\hat{a}_{pq}-a_{pq})$ *являются асимптотически нормально распределенными (при* $T\to\infty$*) нулевым среднем и матрицей вторых моментов* B^{-1}.

Доказательство. По условию *а)* левая сторона (1. 6) имеет вид, при $T \to \infty$

$$B \cdot \sqrt{T}(\hat{\mathbf{a}} - \mathbf{a}).$$

Из условия *б)* следует, что правая сторона (1. 6) имеет нормальное распределение асимптотически, т. е.

$$\boldsymbol{\eta}(T) \to \boldsymbol{\eta}, \quad \text{где} \quad \mathsf{M}\boldsymbol{\eta}\boldsymbol{\eta}^* = B.$$

Тогда

$$B \cdot \sqrt{T}(\hat{\mathbf{a}} - \mathbf{a}) \sim \boldsymbol{\eta}, \quad \sqrt{T}(\hat{\mathbf{a}} - \mathbf{a}) \sim B^{-1} \cdot \boldsymbol{\eta}$$

и

$$\mathsf{M}\sqrt{T}(\hat{\mathbf{a}} - \mathbf{a})\sqrt{T}(\hat{\mathbf{a}} - \mathbf{a})^* \sim \mathsf{M}(B^{-1}\boldsymbol{\eta})(B^{-1}\boldsymbol{\eta})^* = B^{-1},$$

что и требовалось доказать.

Замечание 1. Если процесс $\mathbf{w}(t)$ имеет независимые компоненты, тогда система уравнений (1. 3) распадается на k отдельных систем (при $p = 0, 1, \ldots, k-1$), и при фиксированном p (1. 6) имеет следующий вид

$$\frac{1}{T}\int_0^T \xi_q(t) b_{pp}^{-1}(t) \sum_i \sqrt{T}(\hat{a}_{pi} - a_{pi}) \xi_i(t)\, dt = \frac{1}{\sqrt{T}}\int_0^T \xi_q(t) b_{pp}^{-1}(t)\, dw_p(t) = \eta_{pq}(T),$$

где

$$\mathsf{M}\eta_{pq}(T) = 0, \quad \mathsf{M}\eta_{pq}(T)\eta_{ps}(T) = \frac{1}{T}\int_0^T b_{pp}^{-1}(t)\mathsf{M}\xi_q(t)\xi_s(t)\, dt \to b_{pqs}.$$

Величины $\sqrt{T}(\hat{a}_{pq} - a_{pq})$ являются — при выполнении предположений *а)* и *б)* — асимптотически нормальными с матрицей ковариации $B_p^{-1} = \{b_{pqs}\}^{-1}$, $q, s = 0, 1, \ldots, k-1$.

Теорема 2. *Если* $\mathbf{w}(t)$ *является винеровским процессом и матрица* A *имеет характеристические числа с отрицательной вещественной части, тогда оценки наименьших квадратов асимптотически нормально распределены со матрицей ковариации* $B^{-1} = \{b_{pr}^{-1}\mathsf{M}\xi_q(0)\xi_s(0)\}^{-1}$.

Замечание 2. Когда винеровский процесс $\mathbf{w}(t)$ является процессом с независимыми компонентами и $\mathsf{M}(dw_p)^2 = \sigma_p^2 \cdot dt$, $B(o) = \{\mathsf{M}\xi_s(o)\xi_q(o)\}$ матрица вторых моментов оценок $\hat{a}_{pq}, q = 0, 1, \ldots, k-1$ (при фиксированном p) имеет вид

$$\sigma_p^2 \cdot B^{-1}(0).$$

Матрица $B(o)$ определяется из уравнения (см. Арато [1])

$$A \cdot B(0) + B(0)A^* = -B_w.$$

Доказательство теоремы 2. При условиях нашей теоремы гауссовский, стационарный, марковский процесс $\boldsymbol{\xi}(t)$ будет эргодическим (см. Розанов [8]) и поэтому выполняется условие *а)*. Кроме того, процесс $\boldsymbol{\xi}(t)$ обладает свойством сильного перемешивания и из

$$\mathsf{M}\eta_{pq}(T)\eta_{rs}(T) = \sigma_{rp}^{-1}\mathsf{M}\xi_q(0)\xi_s(0), \quad \text{где} \quad \mathsf{M}(dw_r dw_p) = \sigma_{rp} \cdot dt,$$

следует (см. Волконский—Розанов [2], Розанов [6]), что выполняется и условие *б)*. Тем самым теорема доказана.

Пример. В статье Писаренко [5] рассматривался случай одномерного стационарного гауссовского процесса $\xi(t)$, удовлетворяющего уравнению

$$d\xi^{(k-1)}(t)+\left[a_{k-1}\xi^{(k-1)}(t)+\ \dots\ +a_0\xi(t)\right]dt = dw(t), \quad \text{где} \quad \mathsf{M}(dw)^2=\sigma^2\cdot dt.$$

Из теоремы 2 следует, что оценки условного наибольшего правдоподобия являются асимптотически нормально распределенными с матрицей ковариации $\sigma^2\cdot B^{-1}(0)$. Легко показать, на основе замечания 2, что $B^{-1}(0)$ можно задать в явном виде, а именно (см. Арато [1]), если $B^{-1}(0)=\{b_{ij}^{-1}\}$,

$$b_{ij}^{-1}=\begin{cases} 0, & \text{при } i\not\equiv j \pmod 2,\\ \dfrac{2}{\sigma^2}\displaystyle\sum_l (-1)^l a_{i-l}a_{j+1+l}, & \text{при } i\equiv j \pmod 2,\end{cases}$$

где $a_i=0$, при $i<0$, или $i>k$ ($a_k=1$), и $b_{ij}^{-1}=b_{ji}^{-1}$.

2. Доверительные границы. В дальнейшем предположим, что выполняются условия *а)* и *б)* и, простоты ради, что процесс $\mathbf{w}(t)$ имеет независимые компоненты. Как следует из замечания 1 в этом случае можно — при фиксированном p — независимо рассматривать оценки $\hat{a}_{pq}(q=0,1,\dots,k-1)$ для разных p. Пусть $B_p=\{b_{pqr}\}$, тогда из теоремы 1 следует, что величина

$$T\left[(\hat{\mathbf{a}}_p-\mathbf{a}_p),\ B_p(\hat{\mathbf{a}}_p-\mathbf{a}_p)\right] \tag{2.1}$$

является асимптотически χ_k^2 распределенной величиной, при $T\to\infty$. Из условия *а)* и из одной известной леммы (см. Крамер [3] стр. 281) следует, что величина

$$T\left[(\hat{\mathbf{a}}_p-\mathbf{a}_p),\ \hat{B}_p(\hat{\mathbf{a}}_p-\mathbf{a}_p)\right] \tag{2.2}$$

также является асимптотически χ_k^2 распределенной, где

$$\hat{B}_p=\{\hat{b}_{pqr}\}=\left\{\frac{1}{T}\int_0^T b_{pp}^{-1}(t)\xi_q(t)\xi_r(t)\,dt\right\}.$$

Если асимптотическая нормальность имеет место равномерно, тогда для параметров a_{pq} можно построить доверительные границы. То, что в общем случае не имеет место равномерность, показывает пример стационарного процесса $\boldsymbol{\xi}(t)$, когда для равномерности требуется, чтобы характеристические числа матрицы A (обозначим их через λ_i, $i=0,1,\dots,k-1$) должны удовлетворять условию $Re\ \lambda_i\geqq\varepsilon>0$.

Ограничиваясь несмещенными оценками $\tilde{a}_p$ легко показать, в случае винеровского процесса $\mathbf{w}(t)$, что для матрицы вторых моментов $S=\mathsf{M}(\tilde{\mathbf{a}}_p-\mathbf{a}_p)(\tilde{\mathbf{a}}_p-\mathbf{a}_p)^*$ асимптотически выполняется условие

$$S\geqq\left\{\mathsf{M}\left(\frac{\partial\log L}{\partial a_{pr}}\cdot\frac{\partial\log L}{\partial a_{pq}}\right)\right\}^{-1}_{r,q\,=\,\overline{0,k-1}}, \quad \text{(см. (1.4))}, \tag{2.3}$$

и при $T \to \infty$

$$S \geqq B^{-1} = B^{-1}(0) \tag{2. 4}$$

в том смысле, что $S - B^{-1}(0)$ является положительно определенной матрицей.

Теорема 2 в этом случае означает, что оценки наименьших квадратов (условного наибольшего правдоподобия) являются, при $T \to \infty$, асимптотически эффективными, т. е. имеют меньший эллипсоид рассеивания по сравнению с любыми несмещенными оценками.

Если матрица A такая, что $\lambda_i \sim 0$, тогда $\eta_{pq}(T)$ является χ^2 распределенной величиной и доверительные границы нельзя строить по формуле (2. 2).

ЛИТЕРАТУРА

[1] Арато, М.: Точные формулы для плотностей мер элементарных гауссовских процессов, *Studia Sci. Math. Hungar.* **5** (1970) 17—27.

[2] Волконский, В. А. и Розанов, Ю. А.: Некоторые предельные теоремы для случайных функций, *Теор. Вероятност. и Применен.* **4** (1959) 186—207.

[3] Крамер, Г.: *Математические методы статистики,* Москва, (перевод с англ.).

[4] Mann, H. B. and Wald, A.: On the statistical treatment of linear stochastic difference equations, *Econometrica* **11** (1943) 173—220.

[5] Писаренко, В. Ф.: Об оценках параметров гауссовского стационарного процесса со спектральной плотностью. $|P(i\lambda)|^{-2}$. *Литовский Математический сборник II,* No **2**, (1963) 159—167.

[6] Розанов, Ю. А.: An application of the central limit theorem. *Proc. Fourth Berkeley Symposium*, Vol. 2. (1960) 445—454.

[7] Розанов, Ю. А.: Дополнение редактора к книге Э. Хеннан: *Анализ временных рядов,* Наука, Москва, 1964.

[8] Розанов, Ю. А.: *Стационарные случайные процессы,* Москва, 1963.

Вычислительный Центр Академии Наук Венгрии, Будапешт

(Поступила 6-ого января 1969 г.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 17–27.

ТОЧНЫЕ ФОРМУЛЫ ДЛЯ ПЛОТНОСТЕЙ МЕР ЭЛЕМЕНТАРНЫХ ГАУССОВСКИХ ПРОЦЕССОВ

M. ARATÓ

Целью настоящей работы является определение точных формул плотности мер многомерного гауссовского стационарного марковского процесса (называемого элементарным процессом) относительно соответствующей винеровской меры. При изучении литературы оказывается, что в разных местах даются то ошибочные, то совсем неудобные для практических целей формулы, или определяются условные, при данном начальном условии, плотности (см. [4]—[7], [9]—[11]).

Хотелось бы подчеркнуть полезность формулы Дуба [3] (см. ниже лемму 1), с помощью которой во многих случаях в явном виде дается начальное распределение случайных величин.

В работе рассматриваются производные относительно стандартной винеровской меры, имеющие самостоятельный интерес, особенно в математической статистике. В разных примерах изучаются всевозможные варианты плотностей, важные для практических целей. Для полноты доказательства приводится сжатое доказательство Прохорова [8] для многомерного случая, хотя здесь рассматривается более элементарный случай постоянного переноса. Надо заметить, что абсолютная непрерывность рассматриваемых мер подробно доказывается в литературе (см. цитированную литературу).

Так как нас будет интересовать только случай процесса с непрерывными траекториями, то в дальнейшем всегда предполагается, что $C_k[0, T]$ обозначает пространство всех непрерывных на $[0, T]$ функций $\mathbf{x}^*(t) = (x_0(t), x_1(t), \ldots, x_{k-1}(t))$ со значениями в k-мерном эвклидовом пространстве R_k. Через $\mathscr{A}$ обозначаем σ-алгебру борелевских подмножеств $C_k[0, T]$, порожденную множествами вида $\{\mathbf{x}(s),\ 0 \leqq s \leqq T : \mathbf{x}(t) \in B\}$, $t \in [0, T]$, где B борелевское множество в R_k. Вероятностная мера P ($\mathsf{P}(C_k) = 1$) определена на $\mathscr{A}$.

1. Рассматриваем многомерный стационарный марковский гауссовский процесс $\boldsymbol{\xi}^*(t) = (\xi_0(t), \ldots, \xi_{k-1}(t))$ (где * обозначает сопряженную матрицу), так называемый элементарный гауссовский процесс, удовлетворяющий стохастическому дифференциальному уравнению

$$d\boldsymbol{\xi}(t) = A\boldsymbol{\xi}(t)\,dt + d\mathbf{w}(t), \tag{1}$$

где характеристические числа матрицы A, т. е. решения λ_i уравнения $|A - \lambda E| = 0$, имеют отрицательные вещественные части $Re\ \lambda_i < 0$. Винеровский процесс $\mathbf{w}(t)$ является в общем случае $l (l \leqq k)$ мерным с параметрами $\mathsf{M}d\mathbf{w}(t) = 0$, $\mathsf{M}d\mathbf{w}(t)\,d\mathbf{w}^*(t) = B_w \cdot dt$, где B_w положительно определенная матрица. В случае $l < k$ процесс $\boldsymbol{\xi}(t)$ $(0 \leqq t \leqq T)$ состоит из k-мерного вектора $\boldsymbol{\xi}(0) \in R_k$ с начальным

распределением, заданным с плотностью $f_A(\mathbf{x}_0^*)$ (см., ниже), и из процесса $\boldsymbol{\xi}(t)$ $(0<t\leqq T)$ в l-мерном пространстве непрерывных функций C_l. Простоты ради предположим, что в этом случае (1) имеет вид

$$\frac{d\xi_0(t)}{dt} = \xi_1(t),$$

$$\frac{d\xi_1(t)}{dt} = \xi_2(t),$$

$$\vdots$$

$$\frac{d\xi_{k-l-1}(t)}{dt} = \xi_{k-l}(t),$$

(1′)
$$d\xi_{k-l}(t) = \big(a_{k-l,\,k-1}\,\xi_{k-1}(t) + \ldots + a_{k-l,\,0}\,\xi_0(t)\big)dt + dw_{k-l}(t),$$

$$d\xi_{k-l+1}(t) = \big(a_{k-l+1,\,k-1}\,\xi_{k-1}(t) + \ldots + a_{k-l+1,\,0}\,\xi_0(t)\big)dt + dw_{k-l+1}(t),$$

$$\vdots$$

$$d\xi_{k-1}(t) = \big(a_{k-1,\,k-1}\,\xi_{k-1}(t) + \ldots + a_{k-1,\,0}\,\xi_0(t)\big)dt + dw_{k-1}(t),$$

и пусть $C = B_w^{-1}\tilde{A}$, где $\tilde{A} = \{a_{ij}\}$, $(i = k-l, k-l+1, \ldots, k-1;\ j = 0, 1, \ldots, k-1)$. В этом случае $C_l[0, T]$ состоит из l-мерных непрерывных функций $(x_{k-l}(t), \ldots, x_{k-1}(t))$. Если W_0^l обозначает условную меру (при условии $\mathbf{w}(0) = \mathbf{0}$) винера в пространстве $C_l[0, T]$, то имеет место следующая теорема.

Теорема 1. *Если процесс* $\boldsymbol{\xi}(t)$ *является элементарным гауссовским, удовлетворяющим уравнению* (1′), *и* P_A *обозначает меру, соответствующую этому процессу в пространстве* $C_l[0, T]$, *то* (*при* $\mathbf{x}(0) = \mathbf{0}$)

(2)
$$\frac{d\mathsf{P}_A}{dW_0^l}\big((\mathbf{x}(t))\big) = \exp\left\{\int_0^T (C\mathbf{x}(s), d\mathbf{x}(s)) - \frac{1}{2}\int_0^T (\tilde{A}\mathbf{x}(s), C\mathbf{x}(s))ds\right\},$$

где $\mathbf{x}^*(t) = (x_0(t), \ldots, x_{k-1}(t)) \in C_l[0, T]$ и $(\mathbf{a}, \mathbf{b})$ обозначает скалярное произведение векторов $\mathbf{a}$ и $\mathbf{b}$.

Пусть L^k — мера Лебега в k-мерном пространстве и $L^k \times W_{\mathbf{x}}^l$ — произведение мер в пространстве $R_k \times C_l[0, T]$, где $W_{\mathbf{x}}^l$ условная мера винера при условии $\mathbf{x}^*(0) = \mathbf{x} = (x_0(0), x_1(0), \ldots, x_{k-1}(0))$. Если $f_A(\mathbf{x})$ обозначает плотность вероятности случайного вектора $\boldsymbol{\xi}^*(0) = (\xi_0(0), \ldots, \xi_{k-1}(0))$, то имеем следующее важное следствие.

Следствие 1. При условии теоремы 1

(3)
$$\frac{d\mathsf{P}_A}{d(L^k \times W_{\mathbf{x}}^l)}(\mathbf{x}(t)) = f_A(\mathbf{x}^*(0)) \exp\left\{\int_0^T (C\mathbf{x}(s), d\mathbf{x}(s)) - \frac{1}{2}\int_0^T (\tilde{A}\mathbf{x}(s), C\mathbf{x}(s))ds\right\}.$$

Следствие 2. Если рассматриваются меры P_{A_1} и P_{A_2} с разными матрицами A_1 и A_2 в (1), но с тем же винеровским процессом $\mathbf{w}(t)$, то

$$\frac{d\mathsf{P}_{A_1}}{d\mathsf{P}_{A_2}}(\mathbf{x}(t)) = \frac{d\mathsf{P}_{A_1}}{d(L^k \times W_{\mathbf{x}}^l)}(\mathbf{x}(t)) \frac{d(L^k \times W_{\mathbf{x}}^l)}{d\mathsf{P}_{A_2}}(\mathbf{x}(t)) = \frac{f_{A_1}(\mathbf{x}^*(0))}{f_{A_2}(\mathbf{x}^*(0))} \cdot$$

(3. *a*)

$$\cdot \exp\left\{\int_0^T [(C_1 - C_2)\mathbf{x}(s), d\mathbf{x}(s)] - \frac{1}{2}\int_0^T [(\tilde{A}_1 \mathbf{x}(s), C_1 \mathbf{x}(s)) - (A_2 \mathbf{x}(s), C_2 \mathbf{x}(s)]ds\right\}.$$

Матрица ковариации процесса $\boldsymbol{\xi}(t)$ имеет вид

$$\mathsf{M}\boldsymbol{\xi}(t)\boldsymbol{\xi}^*(s) = e^{A|t-s|}B, \tag{4}$$

где для матрицы B имеет место следующая лемма.

Лемма 1. *Матрица B удовлетворяет матричному уравнению*

$$AB + BA^* = -\tilde{B}_w, \tag{5}$$

где $\tilde{B}_w$ является $k \times k$ матрицей вида $\begin{pmatrix} 0 & 0 \\ 0 & B_w \end{pmatrix}$.

Доказательство леммы легко получить из уравнения, (1) умножая его на $\boldsymbol{\xi}^*(t)$, а потом умножая транспонированное уравнение на $\boldsymbol{\xi}(t+dt)$ и в обоих случаях беря математическое ожидание.

Замечание 1. Если A имеет вид (и в то же время $\tilde{B}_w$)

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & 0 & \dots & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 & \dots & -a_{k-1} \end{pmatrix}, \quad \tilde{B}_w = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \sigma^2 \end{pmatrix}, \tag{6}$$

т. е. одномерный процесс $\xi(t)$ дифференцируем $k-1$ раз и удовлетворяет уравнению

$$d\xi^{(k-1)}(t) + (a_0 \xi(t) + a_1 \xi'(t) + \dots + a_{k-1}\xi^{(k-1)}(t))dt = dw(t), \tag{7}$$

где $\mathsf{M}(dw)^2 = \sigma^2 \cdot dt$, то легко показать, что

$$B^{-1} = \frac{2}{\sigma^2}\begin{pmatrix} a_0 a_1 & 0 & a_0 a_3 & 0 & \dots \\ 0 & a_1 a_2 - a_0 a_3 & 0 & & \dots \\ \vdots & & & & \end{pmatrix}, \quad B^{-1} = \{b_{ij}^{-1}\},$$

(8)

$$b_{ij}^{-1} = \frac{2}{\sigma^2} \cdot \begin{cases} 0, & \text{при } i \equiv j+1 \pmod 2, \\ \sum_{l=0} (-1)^l a_{i-l} a_{j+1+l}, & \text{при } i \equiv j \pmod 2, \ (i = 0, 1, \dots, k-1), \end{cases}$$

где $i<j$, $a_i = 0$ (при $i<0$ или $i>k$), $a_k = 1$, и $b_{ij}^{-1} = b_{ji}^{-1}$.

2. Перед доказательством основной теоремы рассматриваем важные примеры, вытекающие из теоремы 1 и ее следствия 1.

Пример 1. Для стационарного гауссовского процесса, удовлетворяющего уравнению (7), с постоянными коэффициентами получается

$$\frac{d\mathsf{P}_A}{d(L^k\times W_{\mathbf{x}}^l)}(\mathbf{x}(t)) = (2\pi)^{-\frac{k}{2}}|B|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=0}^{k-1}\left(\sum_{l=-i}^{i}(-1)^l a_{i-l}a_{i+l}\right)\cdot\right.$$

$$\cdot\int_0^T\cdot[x^{(i)}(t)]^2dt+\frac{1}{2\sigma^2}\sum_{i,j=0}^{k-1}\left(\sum_{l=0}(-1)^l a_{i-l}a_{j+1+l}\right)\cdot \tag{9}$$

$$\left.\cdot[x^{(i)}(T)x^{(j)}(T)+(-1)^{j-i}x^{(i)}(0)x^{(j)}(0)]+\frac{a_{k-1}a_k}{2}T\right\},$$

где матрица B^{-1} имеет вид (8), где $c_{ij}=\sum_l(-1)^l a_{i-l}a_{j+1+l}$, при $i>j$ и $c_{ij}=c_{ji}$.

Доказательство формулы (9). По формуле (3) имеем

$$\frac{d\mathsf{P}_A}{d(L^k\times W_{\mathbf{x}}^l)}(\mathbf{x}(t)) = (2\pi)^{-\frac{k}{2}}|B|^{-\frac{1}{2}}\exp\left\{-\frac{1}{\sigma^2}\sum_{i,j=0}^{k-1}{}'\left(\sum_{l=0}(-1)^l a_{i-l}a_{j+1+l}\right)\cdot\right.$$

$$\left.\cdot x^{(i)}(0)x^{(j)}(0)+\frac{1}{\sigma^2}\int_0^T\sum_{i=0}^{k-1}a_i x^{(i)}(t)dx^{(k-1)}(t)-\frac{1}{2\sigma^2}\int_0^T\sum_{i,j=0}^{k-1}a_i a_j x^{(i)}(t)x^{(j)}(t)dt\right\},$$

где в первой сумме $'$ обозначает, что сумма распространяется на те j, при которых $i\equiv j \pmod 2$. Используя известные соотношения

$$\int_0^T x^{(k-1)}(t)dx^{(k-1)}(t)=\frac{1}{2}\left[(x^{(k-1)}(T))^2-(x^{(k-1)}(0))^2-\sigma^2\cdot T\right],$$

$$\int_0^T x^{(i)}(t)dx^{(k-1)}(t)=x^{(i)}(t)x^{(k-1)}(t)\Big|_0^T-\int_0^T x^{(i+1)}(t)x^{(k-1)}(t)dt,\quad i<k-1,$$

и

$$\int_0^T x^{(i)}(t)x^{(i+l)}(t)dt=[x^{(i)}(t)x^{(i+l-1)}(t)]_0^T-[x^{(i+1)}(t)x^{(i+l-2)}(t)]_0^T\pm\ldots+\eta_l, \tag{10}$$

где

$$\eta_l=\begin{cases}(-1)^{\frac{l}{2}}\int_0^T\left[x^{\left(i+\frac{l}{2}\right)}(t)\right]^2dt, & \text{при четном } l,\\ \frac{(-1)^{\frac{l-1}{2}}}{2}\left[x^{\left(i+\frac{l-1}{2}\right)}\right]_0^T, & \text{при нечетном } l.\end{cases}$$

Таким образом, мы получаем для плотности

$$\frac{d\mathsf{P}_A}{d(L^k\times W_{\mathbf{x}}^l)}(\mathbf{x}(t)) = (2\pi)^{-\frac{k}{2}}|B|^{-\frac{1}{2}}\exp\left\{-\frac{1}{\sigma^2}\sum_{i,j=0}^{k-1}{}'\left(\sum_{l=0}(-1)^l a_{i-l}a_{j+1+l}\right)\cdot\right.$$

$$\cdot x^{(i)}(0)x^{(j)}(0) - \frac{1}{\sigma^2}\sum_{i=0}^{k-1} a_i\left[x^{(i)}(t)x^{(k-1)}(t)\right]_0^T - \frac{1}{2\sigma^2}\left[(x^{k-1})^2\right]_0^T +$$

$$+\frac{1}{\sigma^2}\sum_{i=0}^{k-1} a_i\sum_{l=0}^{\prime\prime}(-1)^l\left[x^{(i+1+l)}(t)x^{(k-2-l)}\right]_0^T + \frac{a_k a_{k-1}}{2}T +$$

$$\left.-\frac{1}{2\sigma^2}\sum_{i=0}^{k-1} a_i^2\int_0^T\left[x^{(i)}(t)\right]^2 dt - \frac{1}{\sigma^2}\sum_{i=0}^{k-1}\sum_{j=0}^{k-1-i} a_i a_{i+j}\sum_{l=0}^{j/2\,\prime\prime}(-1)^l\left[x^{(i+l)}(t)x^{(i+j-l-1)}(t)\right]_0^T\right\},$$

где $''$ (при суммировании) обозначает, что последний член имеется в виду в той форме, как это указано в формуле (10), т. е. имеет вид $\int_0^T (x^{(i+j/2)}(t))^2 dt$ при четном j.

Поменяв порядок суммирования и используя, что $a_k = 1$ приходим к формуле (9), так как

$$\sum_{i=0}^{k-1} a_i\left[x^{(i)}(t)x^{(k-1)}(t)\right]_0^T + \sum_{i=0}^{k-1} a_i\sum_{l=0}^{\prime\prime}(-1)^l\left[x^{(i+1+l)}(t)x^{(k-2-l)}(t)\right]_0^T +$$

$$+\sum_{i=0}^{k-1}\sum_{j=0}^{k-1-i} a_i a_{i+j}\sum_{l=0}^{\prime\prime}(-1)^l\left[x^{(i+l)}(t)x^{(i+j-l-1)}(t)\right]_0^T =$$

$$=\sum_{i=0}^{k-2} a_i a_k\left[x^{(i)}(t)x^{(k-1)}(t)\right]_0^T + \sum_{\substack{i=1\\ j>i}}^{k-2}\sum_{l}(-1)^l a_{i-l-1}a_k\left[x^{(i)}(t)x^{(j)}(t)\right]_0^T +$$

$$+\frac{1}{2}\left[(x^{(k-1)})^2\right]_0^T + \sum_{\substack{i=0\\ j>i}}^{k-2}\sum_{l=0}^{\min(i,k-1-(j+1))}(-1)^l a_{i-l}a_{j+1+l}\left[x^{(i)}(t)x^{(j)}(t)\right]_0^T +$$

$$+\sum_{i=0}^{k-1}\sum_{\substack{l=-i\\ l\neq 0}}^{i}(-1)^l a_{i-l}a_{i+l}\int_0^T\left[x^{(i)}(t)\right]^2 dt = \frac{1}{2}\sum_{i,j=0}^{k-1}\sum_{l=0}(-1)^l a_{i-l}a_{j+1+l}\left[x^{(i)}(t)x^{(j)}(t)\right]_0^T -$$

$$-\frac{1}{2}\sum_{i=0}^{k-1}\left(\sum_{\substack{l=-i\\ l\neq 0}}^{i}(-1)^l a_{i-l}a_{i+l}\right)\int_0^T\left[x^{(i)}(t)\right]^2 dt.$$

При $k=1$ получается известная формула (см. Прохоров [8], Striebel [11], $a_0=\lambda>0$)

$$\frac{d\mathsf{P}_\lambda}{d(L^1\times W_x^1)}(x(t)) = \sqrt{\frac{\lambda}{\pi}}\frac{1}{\sigma}\exp\left\{-\frac{\lambda^2}{2\sigma^2}\int_0^T x^2(t)dt - \frac{\lambda}{2\sigma^2}\left[x^2(T)+x^2(0)\right] + \frac{\lambda T}{2}\right\}.$$

При $k=2$ получаем

$$\frac{d\mathsf{P}_A}{d(L^2\times W_x^1)}(\mathbf{x}(t)) = \frac{a_1\sqrt{a_0}}{\pi\sigma^2}\exp\left\{-\frac{a_0^2}{2\sigma^2}\int_0^T x^2(t)\,dt - \frac{a_1^2-2a_0}{2\sigma^2}\int_0^T [x^{(1)}(t)]^2\,dt + \right.$$

$$+\frac{a_1 T}{2} - \frac{a_0 a_1}{2\sigma^2}[x^2(T)+x^2(0)] - \frac{a_1}{2\sigma^2}[(x^{(1)}(T))^2+(x^{(1)}(0))^2] -$$

$$\left. -\frac{a_0}{\sigma^2}[x(T)x^{(1)}(T) - x(0)x^{(1)}(0)]\right\}.$$

При $k=3$ получаем

$$\frac{d\mathsf{P}_A}{d(L^3\times W_x^1)}(\mathbf{x}(t)) = \frac{\sqrt{a_0}(a_2a_1-a_0)}{(\pi\sigma^2)^{3/2}}\exp\left\{-\frac{a_0^2}{2\sigma^2}\int_0^T x^2(t)\,dt - \frac{a_1^2-2a_0a_2}{2\sigma^2}\cdot\right.$$

$$\cdot\int_0^T [x^{(1)}(t)]^2\,dt - \frac{a_2^2-2a_1}{2\sigma^2}\int_0^T [x^{(2)}(t)]^2\,dt + \frac{a_2 T}{2} - \frac{a_0a_1}{2\sigma^2}[x^2(T)+x^2(0)] -$$

$$-\frac{a_1a_2-a_0}{2\sigma^2}[(x^{(1)}(T))^2+(x^{(1)}(0))^2] - \frac{a_2}{2\sigma^2}[(x^{(2)}(T))^2+(x^{(2)}(0))^2] - \frac{a_0a_2}{\sigma^2}\cdot$$

$$\cdot[x(T)x^{(1)}(T)-x(0)x^{(1)}(0)] - \frac{a_0}{\sigma^2}[x(T)x^{(2)}(T)+x(0)x^{(2)}(0)] -$$

$$\left. -\frac{a_1}{\sigma^2}[x^{(1)}(T)x^{(2)}(T)-x^{(1)}(0)x^{(2)}(0)]\right\}.$$

По этим формулам и с помощью формулы (3. а) легко определить плотность, соответствующую элементарным процессам с разными матрицами A. На этом не будем останавливаться.

Пример 2. Рассмотрим двумерный процесс, где различается три случая: *а)* корни характеристического полинома матрицы A вещественные и разные, *б)* корни комплексно сопряженные, *в)* имеется двойной корень. Во всех этих случаях предположим, что винеровский процесс $\mathbf{w}^*(t)=(w_1(t), w_2(t))$ имеет независимые компоненты.

а) Если

$$d\xi_1 = -\lambda_1\xi_1(t)\,dt + dw_1,$$

$$d\xi_2 = -\lambda_2\xi_2(t)\,dt + dw_2,$$

где $\mathsf{M}(dw_i)^2 = \sigma_i^2\cdot dt$, то

$$\frac{d\mathsf{P}_{\lambda_1,\lambda_2}}{d(L^2\times W_x^2)}(\mathbf{x}(t)) =$$

$$= \prod_{i=1}^{2}\sqrt{\frac{\lambda_i}{\pi}}\cdot\frac{1}{\sigma_i}\exp\left\{-\frac{\lambda_i^2}{2\sigma_i^2}\int_0^T x_i^2(t)\,dt - \frac{\lambda_i}{2\sigma_i^2}[x_i^2(T)+x_i^2(0)] + \frac{\lambda_i T}{2}\right\}.$$

δ) Если

$$A = \begin{pmatrix} -\lambda & -\omega \\ \omega & -\lambda \end{pmatrix} \quad \text{и} \quad \mathsf{M}(dw_i)^2 = \sigma^2 \cdot dt,$$

то

$$f_A(\mathbf{x}(0)) = \frac{\lambda}{\pi\sigma^2} \exp\left\{-\frac{\lambda}{\sigma^2} x_1^2(0) - \frac{\lambda}{\sigma^2} x_2^2(0)\right\},$$

и

$$\frac{d\mathsf{P}_A}{d(L^2 \times W_x^2)}(\mathbf{x}(t)) = \frac{\lambda}{\pi\sigma^2} \exp\left\{-\frac{\lambda^2+\omega^2}{2\sigma^2} \int_0^T [x_1^2(t)+x_2^2(t)]\,dt - \right.$$

$$\left. -\frac{\lambda}{2\sigma^2}[x_1^2(T)+x_2^2(T)+x_1^2(0)+x_2^2(0)] + \lambda T + \frac{\omega}{\sigma^2} \int_0^T [x_1(t)\,dx_2(t) - x_2(t)\,dx_1(t)]\right\}.$$

При обозначениях $x(t) = x_1(t) + ix_2(t)$, $|x(t)|^2 = x_1^2(t) + x_2^2(t)$, $x(t) = |x(t)|\, e^{i\theta(t)}$ плотность перепишется с помощью следующего соотношения

$$(*) \qquad \int_0^T [x_1(t)\,dx_2(t) - x_2(t)\,dx_1(t)] = \int_0^T |x(t)|^2\,d\theta.$$

Чтобы доказать $(*)$ заметим, что

$$\sum_j [x(t_j)\overline{x(t_{j-1})} - x(t_{j-1})\overline{x(t_j)}] =$$

$$= -2i \sum_j [x_2(t_j)(x_1(t_j) - x_1(t_{j-1})) - x_1(t_j)(x_2(t_j) - x_2(t_{j-1}))],$$

где левая сторона выражается следующим образом

$$\sum_j |x(t_j)|\,|x(t_{j-1})|[e^{i(\theta(t_j)-\theta(t_{j-1}))} - e^{i(\theta(t_{j-1})-\theta(t_j))}] =$$

$$= \sum_j |x(t_j)|\,|x(t_{j-1})|\,2i \sin(\theta(t_j) - \theta(t_{j-1})) \sim 2i \sum_j |x(t_j)|^2(\theta(t_j) - \theta(t_{j-1})).$$

Таким образом

$$\frac{d\mathsf{P}_A}{d(L^2 \times W_x^2)}(x(t)) = \frac{\lambda}{\pi\sigma^2} \exp\left\{-\frac{\lambda^2+\omega^2}{2\sigma^2} \int_0^T |x(t)|^2\,dt + \right.$$

$$\left. + \frac{\omega}{\sigma^2} \int_0^T |x(t)|^2\,d\theta + \lambda T - \frac{\lambda}{2\sigma^2}[|x(T)|^2 + |x(0)|^2]\right\}.$$

в) Если P_A обозначает меру, соответствующую процессу

$$d\xi_1 = -\lambda\xi_1(t)\,dt + \xi_2(t)\,dt + dw_1(t),$$

$$d\xi_2 = -\lambda\xi_2(t)\,dt + dw_2(t),$$

где $\mathsf{M}(dw_1)^2 = \sigma^2 \cdot dt$, $\mathsf{M}(dw_2)^2 = dt$, т. е.

$$A = \begin{pmatrix} -\lambda & 1 \\ 0 & -\lambda \end{pmatrix}, \quad B_w = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix},$$

то

$$B = \mathsf{M}\xi(t)\xi^*(t) = \begin{pmatrix} \frac{\sigma^2}{2\lambda} + \frac{1}{\lambda(2\lambda)^2} & \frac{1}{(2\lambda)^2} \\ \frac{1}{(2\lambda)^2} & \frac{1}{2\lambda} \end{pmatrix}, \quad |B|^{-\frac{1}{2}} = \frac{(2\lambda)^2}{\sqrt{1+\sigma^2(2\lambda)^2}}.$$

Таким образом

$$\frac{d\mathsf{P}_A}{d(L^2 \times W_x^2)}(\mathbf{x}(t)) = f_A(\mathbf{x}^*(0)) \exp\left\{ -\frac{\lambda}{\sigma^2} \int_0^T x_1(t)dx_1(t) + \frac{1}{\sigma^2} \int_0^T x_2(t)dx_1(t) + \right.$$

$$\left. - \lambda \int_0^T x_2(t)dx_2(t) - \frac{1}{2} \int_0^T \left[\frac{\lambda^2}{\sigma^2} x_1^2(t) + \frac{1}{\sigma^2} x_2^2(t) - \frac{2\lambda}{\sigma^2} x_1(t)x_2(t) + \left(\frac{1}{\sigma^2} + \lambda^2 \right) x_2^2(t) \right] dt, \right.$$

и используя уже известные формулы получаем

$$\frac{d\mathsf{P}_A}{d(L^2 \times W_x^2)}(\mathbf{x}(t)) = \frac{(2\lambda)^2}{2\pi\sqrt{1+\sigma^2(2\lambda)^2}} \exp\left\{ -\frac{1}{2} \int_0^T \left[\frac{\lambda^2}{\sigma^2} x_1^2(t) - \frac{2\lambda}{\sigma^2} x_1(t)x_2(t) + \right. \right.$$

$$\left. + \left(\lambda^2 + \frac{1}{\sigma^2} \right) x_2^2(t) \right] dt + T\lambda + \frac{1}{\sigma^2} \int_0^T x_2(t)dx_1(t) - \frac{\lambda}{2\sigma^2} x_1^2(T) - \frac{\lambda}{2} x_2^2(T) +$$

$$- \left[\frac{\lambda}{2\sigma^2} + \frac{(2\lambda)^3}{1+\sigma^2(2\lambda)^2} \right] x_1^2(0) - \left[\frac{\lambda}{2} + \lambda \left(1 + \frac{1}{1+\sigma^2(2\lambda)^2} \right) \right] x_2^2(0) +$$

$$\left. + \frac{(2\lambda)^2}{1+\sigma^2(2\lambda)^2} x_1(0)x_2(0) \right\}.$$

Заметим, что в экспоненте присутствуют интегралы, не зависящие от параметра λ, что сокращается при вычислении плотности мер, соответствующих разным матрицам A.

г) Если процесс удовлетворяет уравнению

$$d\xi_1 = a_{11}\xi_1(t)dt + a_{12}\xi_2(t)dt + dw_1 + dw_2,$$

$$d\xi_2 = a_{21}\xi_1(t)dt + a_{22}\xi_2(t)dt + dw_2,$$

где $w_1(t)$ и $w_2(t)$ независимые, то $(\mathsf{M}(dw_i)^2 = dt;\ i = 1, 2)$

$$B_w = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad B_w^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix},$$

(и $\mathbf{x}(0)=\mathbf{o}$)

$$\frac{dP_A}{dW_0^2}(\mathbf{x}(t)) = \exp\left\{-\frac{a_{11}^2-2a_{11}a_{21}+2a_{21}^2}{2}\int_0^T x_1^2(t)\,dt-\frac{a_{12}^2-2a_{12}a_{22}+2a_{22}^2}{2}\cdot\right.$$

$$\cdot\int_0^T x_2^2(t)\,dt+\frac{2a_{12}a_{11}-2a_{12}a_{21}-2a_{11}a_{22}+4a_{21}a_{22}}{2}\int_0^T x_1(t)x_2(t)\,dt+$$

$$+(a_{12}-a_{22})\int_0^T x_2(t)\,dx_1(t)+(2a_{21}-a_{11})\int_0^T x_1(t)\,dx_2(t)+$$

$$\left.+\frac{a_{11}-a_{21}}{2}[x_1^2(T)-x_1^2(0)]+\frac{2a_{22}-a_{12}}{2}[x_2^2(T)-x_2^2(0)]-\frac{2(a_{22}+a_{11})-a_{12}-2a_{21}}{2}T\right\}.$$

3. Доказательство основной теоремы. В дальнейшем предположим, что $\mathbf{w}(t)$ k-мерный. Если $\mathbf{y}(t)\in C_k[0,T]$ и $d_n=(t_0^{d_n}<t_1^{d_n}<\ldots<t_m^{d_n})$ какое-нибудь разбиение отрезка $[0,T]$, каждое из которых является продолжением предыдущего и для которых $\varrho(d_n)=\max\limits_i(t_{i+1}^{d_n}-t_i^{d_n})\to 0$ при $n\to\infty$, то с помощью эйлерова приближения

$$\mathbf{Y}(0)=\mathbf{y}(0),$$

$$\mathbf{Y}(t)=\mathbf{Y}(t_{i-1}^{d_n})+A\mathbf{Y}(t_{i-1}^{d_n})(t-t_{i-1}^{d_n})+\mathbf{x}(t)-\mathbf{x}(t_{i-1}^{d_n}),$$

$$(t_{i-1}<t\leqq t_i),$$

где

$$\mathbf{y}(t)=\int_0^t A\mathbf{y}(s)\,ds+\mathbf{x}(t),\quad(\mathbf{x}(0)=0),$$

получается непрерывное отображение пространства C_k в себя

$$\pi(d_n)\colon\ \mathbf{y}(t)\to\mathbf{Y}(t).$$

Из свойств эйлеровых приближений и теоремы 1. 10 работы [8] следует, что $\mathbf{Y}(t)\to\mathbf{y}(t)$ равномерно на каждом компакте пространства C_k, и обозначая меру процесса $\mathbf{Y}(t)$ через $(\mathsf{P}_A)^{\pi(d_n)}$

$$(\mathsf{P}_A)^{\pi(d_n)}\Rightarrow\mathsf{P}_A, \tag{1}$$

если $\varrho(d_n)\to 0$, в смысле слабой сходимости.

Имеет место следующее утверждение $(\mathsf{P}_A)^{\pi(d_n)}\ll W_o^k$ и

$$p_n=\frac{d(\mathsf{P}_A)^{\pi(d_n)}}{dW_0^k}(\mathbf{x}(t))=\exp\left\{\sum_{j=1}^m(C\mathbf{x}_{j-1},\Delta\mathbf{x}_j)-\frac{1}{2}\sum_{j=1}^m[A\mathbf{x}_{j-1},C\mathbf{x}_{j-1}]\Delta t_j^{d_n}\right\}. \tag{2}$$

Доказательство (2) легко получается, если рассмотреть продолжение $d_{n'}$ разбиения d_n, тогда конечные распределения меры $(\mathsf{P}_A)^{\pi(d_n)}$ в точках $t_0^{d_{n'}},\ldots,t_{im'}^{d_n}$

абсолютно непрерывны относительно меры W_0^k в тех же точках и их плотность равна

$$\exp\frac{1}{2}\left\{-\sum_{j=1}^{m}\sum_{i>i_{j-1}}^{i_j}\frac{(B_w^{-1}\Delta\mathbf{x}_i^{d_{n'}}-C\mathbf{x}_{j-1}^{d_n}\Delta t_i^{d_{n'}},\ \Delta\mathbf{x}_i^{d_{n'}}-A\mathbf{x}_{j-1}^{d_n}\Delta t_i^{d_{n'}})}{\Delta t_i^{d_{n'}}}+\right.$$

$$\left.+\sum_{i=1}^{i_m}\frac{(B_w^{-1}\Delta\mathbf{x}_i^{d_{n'}},\ \Delta\mathbf{x}_i^{d_{n'}})}{\Delta t_i^{d_{n'}}}\right\},$$

что переходит в (2), по теореме сходимости мартингалов и из того факта, что интеграл функции $\frac{d(\mathsf{P}_A)^{\pi(d_n)}}{dW_o^k}$ равен 1 по всему пространству. Первая сумма в формуле (2) сходится W_0^k-среднем квадратичном к

$$\int_0^T (C\mathbf{x}(t), d\mathbf{x}(t)),$$

а вторая для всех $\mathbf{x}(t)\in C_k$ к

$$-\frac{1}{2}\int_0^T (A\mathbf{x}(t), C\mathbf{x}(t))dt,$$

поэтому можно выбрать такую подпоследовательность d_{n_r}, что сходимость имеет место почти всюду по W_0^k.

Так как $\ln p_n$ имеет предел и по мере P_A, легко показать, что последовательность p_n равномерно интегрируема. Но тогда

$$(\mathsf{P}_A)^{\pi(d_n)}(B)=\int_B p_n dW_o^k \to \int_B p\,dW_o^k,$$

где

$$p(\mathbf{x}(t))=\exp\left\{\int_0^T (C\mathbf{x}(t), d\mathbf{x}(t))-\frac{1}{2}\int_0^T (A\mathbf{x}(t), C\mathbf{x}(t))dt\right\}.$$

С другой стороны мы уже показали, что

$$(\mathsf{P}_A)^{(\pi d_n)}\Rightarrow \mathsf{P}_A,$$

таким образом

$$\mathsf{P}_A(B)=\int_B p\,dW_0^k$$

что и требовалось доказать.

Надо заметить, что если вместо „процесса" $\mathbf{Y}(t)$ мы рассматриваем процесс $\tilde{y}(t)$

$$\tilde{\mathbf{y}}(t)=\mathbf{y}(t_{i-1})+\frac{(t-t_{i-1})}{t_i-t_{i-1}}(\mathbf{y}(t_i)-\mathbf{y}(t_{i-1})),\quad (t_{i-1}<t\leqq t_i),$$

то нельзя определить так легко плотность вероятности относительно W_0^k, как это сделано в формуле (2).

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] Гихман, И. И., Скороход, А. В.: О плотностях вероятностных мер в функциональных пространствах, *Успехи Мат. Наук* **21** (6) (1966) 83—152.
[2] Гирсанов, И. В.: О преобразовании одного класса случайных процессов с помощью абсолютно непрерывной замены меры, *Теор. Вероятност.* и *Применен.* **5** (1960) 314—330.
[3] Doob, J. L.: The elementary gaussian processes, *Ann. Math. Statist.* **15** (1944) 229—281.
[4] Hajek, J.: On linear statistical problems in stochastic processes, *Czechoslovak Math. J.* **12** (1962) 404—444.
[5] Михалевич, В. С., Скороход, А. В.: О статистике некоторых процессов, *Труды VI Всесоюзного сов. по теории вер. и мат. стат.* (1960) 229—232.
[6] Писаренко, В. Ф.: К задаче обнаружения случайного сигнала на фоне шума, *Радиотехн. и Электрон.* **6** (1961) 515—528.
[7] Писаренко, В. Ф.: Об оценках параметров гауссовского стационарного процесса со спектральной плотностью, *Литовск. Мат. Сб.* **2** (2) (1962) 159—167.
[8] Прохоров, Ю. В.: Сходимость случайных процессов и предельные теоремы теории вероятностей. Теор. Вероятност. и Применен. **1** (1956) 177—238.
[9] Розанов, Ю. А.: Дополнение к книге Хэннан Э.: *Анализ временных рядов,* Москва, 1964.
[10] Скороход, А. В.: *Исследование по теории случайных процессов,* Киев.
[11] Striebel, Ch.: Densities for stochastic processes, *Ann. Math. Statist.* **30** (1959) 559—567.
[12] Халмош, П.: Теория меры, Москва, И. Л., 1953.

Вычислительный Центр Академии Наук Венгрии, Будапешт

(Поступила 6-ого января 1969 г.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 29—36.

ON A PROBLEM OF STATISTICAL GROUP THEORY

by

K. BOGNÁR

The subject of this paper is connected, to a certain extent, with the investigations of P. ERDŐS and A. RÉNYI having their origin in number theoretical problems and leading to the group theoretical results given in [1]; on the other hand it is connected with a paper ([2]) of R. J. MIECH, in which the author gives a generalization of Theorem 1 of [1].

The problem will be discussed in the present paper by a new method which is based on the application of group characters and is especially suitable for investigating such problems where the structure of the group is of importance. The application of group characters simplifies the computations and provides an opportunity for generalizations as well.

Before giving precise results, let us introduce some notations.

NOTATIONS. Let G_n be a finite Abelian group of order n. The group operation will be written as addition. The elements of G_n will be denoted by a, b, g (or a_i, b_i, g_i).

The number of the elements in a set H will be denoted by $N(H)$.

Let a probability space (see [3]) $V_k = (\Omega_k, \mathscr{A}_k, \mathsf{P})$ be defined as follows:

$$\Omega_k = G_n^{(1)} \times G_n^{(2)} \times \cdots \times G_n^{(k)}$$

is the Cartesian product of k copies of G_n. (An element of Ω_k, i.e. a k-tuple of group elements, will be denoted by $\omega(k)$.)

$\mathscr{A}_k$ is the algebra of all subsets of Ω_k and P the probability measure on $\mathscr{A}_k$, whose value at each point of Ω_k is equal to $\frac{1}{n^k}$.

Let the set $E_k^{(s)}$ $(s \geqq 1)$ be defined as follows:

$$(1) \qquad E_k^{(s)} = \{(\varepsilon^{(s)}) = (\varepsilon_1^{(s)}, \ldots, \varepsilon_k^{(s)}) : \varepsilon_i^{(s)} = 0, 1, 2, \ldots, s; \qquad i = 1, 2, \ldots, k\},$$

(for $s = 1$ we shall omit the index, i. e. $E_k^{(1)} \equiv E_k$; $(\varepsilon^{(1)}) \equiv (\varepsilon)$, $\varepsilon_i^{(1)} \equiv \varepsilon_i$).

Let us define for every element g of G_n the random variable $\nu_g^{(s)}(k, \omega(k))$ (briefly $\nu_g^{(s)}$) $(\nu_g^{(1)} \equiv \nu_g)$ which gives in the point $\omega(k) = (a_1, a_2, \ldots, a_k)$ the number of representations of the group element g in the form

$$(2) \qquad g = \sum_{i=1}^{k} \varepsilon_i^{(s)} a_i \qquad \text{(briefly: } g = (\varepsilon^{(s)}, \omega)\text{)}$$

whenever $(\varepsilon^{(s)}) \in E_k^{(s)}$.

M(...) denotes the expectation of the random variable in the bracket.
Let further be for fixed $s \geqq 1$ and $d \geqq 2$

$$\mu_d^{(s)}(n, k) = \mathsf{M}\left(\sum_{g \in G_n} \left[v_g^{(s)}(k, \omega(k))\right]^d\right) \tag{3}$$

and

$$m_d^{(s)}(n, k) = \mathsf{M}\left(\sum_{g \in G_n} \left[v_g^{(s)}(k, \omega(k)) - \frac{(s+1)^k}{n}\right]^d\right) \tag{4}$$

respectively ($s = 1, 2, 3, \ldots$; $d = 2, 3, \ldots$). ($\mu_d^{(1)} \equiv \mu_d$; $m_d^{(1)} \equiv m_d$).
(5) If G_n is expressed as the direct sum of cyclic groups of prime power order, then let r denote the number of summands having orders that are powers of 2.

Theorem 1 of ERDŐS and RÉNYI, proved in [1], can be formulated with our notations as follows: If

$$k \geqq \frac{2 \log n + 2 \log \frac{1}{\varepsilon} + \log \frac{1}{\delta}}{\log 2}, \tag{6}$$

where ε and δ are arbitrary positive numbers, then

$$\mathsf{P}\left(\max_{g \in G_n} \left| v_g(k, \omega(k)) - \frac{2^k}{n} \right| < \varepsilon \frac{2^k}{n}\right) > 1 - \delta. \tag{7}$$

To prove this result they determined the value of $m_2(n, k)$ which turned out to be independent of the special structure of the group.

R. J. MIECH ([2]), investigating the problem whether the factor 2 of $\log n$ in (6) can be replaced by a smaller one, got the following result: If

$$k \geqq \left(\max\left\{\frac{3}{2} \log n; \log n + r \log 2\right\} + 4 \log \frac{1}{\varepsilon} + \log \frac{1}{\delta}\right) \frac{1}{\log 2} + 8, \tag{8}$$

then (7) holds. His proof was based on the (upper) estimation of the value of $m_4(n, k)$ which turned out to depend on the structure of the group G_n, more exactly only on the number r of summands (in the direct sum decomposition of G_n) having orders that are powers of 2.

In the present paper the exact value of $m_4(n, k)$ will be given as well as the value of $m_2^{(2)}(n, k)$, by aid of which we get a generalization of Theorem 1 of ERDŐS and RÉNYI.

In the computations the characters of a finite Abelian group will play important role, hence we summarize some of their simple properties to make references easier (see e.g. [4]).

A character $\chi(g)$ of G_n is a complex valued (not identically zero) function defined for $g \in G_n$, for which

(C1) $\chi(g_1 + g_2) = \chi(g_1)\chi(g_2) \qquad (g_1 \in G_n, g_2 \in G_n).$

(C2) $\chi(e) = 1$ for each character χ of G_n, if e is the unit element of G_n.
(C3) $|\chi(g)| = 1$ for each $g \in G_n$.

(C4) The characters of a group G_n form a group with respect to the operation of multiplication of characters, isomorphic to the group G_n with the unit element $\chi_0 \equiv 1$. The group of the characters of G_n will be denoted by H_n. The inverse of a character χ is $\bar{\chi}$, where $\overline{\chi(g)}$ is the complex conjugate of $\chi(g)$.

Let us introduce the following notation:

$$\text{(C5)} \qquad \frac{1}{n} \sum_{g \in G_n} \chi(g) = \delta_\chi .$$

Then for $\chi \in H_n$

$$\text{(C6)} \qquad \delta_\chi = \begin{cases} 1, & \text{if} \quad \chi = \chi_0 \\ 0, & \text{if} \quad \chi \neq \chi_0 , \end{cases}$$

$$\text{(C7)} \qquad \sum_{\chi \in H_n} \chi(g) = \begin{cases} n, & \text{if} \quad g = e, \\ 0, & \text{if} \quad g \neq e. \end{cases}$$

If $\chi_1 \in H_n$, $\chi_2 \in H_n$, then

$$\text{(C8)} \qquad \delta_{\chi_1 \bar{\chi}_2} = \begin{cases} 1, & \text{if} \quad \chi_1 = \chi_2 , \\ 0, & \text{if} \quad \chi_1 \neq \chi_2 . \end{cases}$$

If $g_1 \in G_n$, $g_2 \in G_n$, then

$$\text{(C9)} \qquad \sum_{\chi \in H_n} \chi(g_1) \overline{\chi(g_2)} = \begin{cases} n, & \text{if} \quad g_1 = g_2 , \\ 0, & \text{if} \quad g_1 \neq g_2 . \end{cases}$$

The method of characters is based on the following lemmas:

LEMMA 1. *If $\chi \in H_n$ is an arbitrary character of G_n, $\omega(k) = (a_1, a_2, \ldots, a_k) \in \Omega_k$ and*

$$\text{(9)} \qquad A_k^{(s)}(\chi) \overset{\text{def}}{=} \prod_{j=1}^{k} \left(1 + \chi(a_j) + \chi^2(a_j) + \cdots + \chi^s(a_j)\right), \quad s \geqq 1; \quad \left(A_k^{(1)}(\chi) \equiv A_k(\chi)\right),$$

then

$$\text{(10)} \qquad A_k^{(s)}(\chi) = \sum_{g \in G_n} v_g^{(s)} \chi(g).$$

PROOF. Performing multiplication in (9) and taking into account that (by (C1)) $\chi^t(a_j) = \chi(ta_j)$, $(t = 1, 2, \ldots)$, we have

$$A_k^{(s)}(\chi) = \sum_{(\varepsilon^{(s)}) \in E_k^{(s)}} \chi\left((\varepsilon^{(s)}, \omega)\right),$$

hence by the definition of $v_g^{(s)}$, we get (10).

LEMMA 2. *It is supposed that the elements of H_n are numbered from 0 to $n-1$ ($\chi_0 \equiv 1$), then we have*

$$\text{(11)} \qquad \mu_d^{(s)}(n, k) \cdot n^{d-1} =$$

$$= \mathsf{M}\left(\sum_{l_1=0}^{n-1} \sum_{l_2=0}^{n-1} \cdots \sum_{l_{d-1}=0}^{n-1} A_k^{(s)}(\chi_{l_1}) A_k^{(s)}(\bar{\chi}_{l_1} \chi_{l_2}) \cdots A_k^{(s)}(\bar{\chi}_{l_{d-2}} \chi_{l_{d-1}}) A_k^{(s)}(\bar{\chi}_{l_{d-1}}) \right).$$

$$(d = 2, 3, \ldots)$$

PROOF. According to (10), changing the orders of summation on the right hand side of (11) with respect to the elements of H_n and G_n resp., we have

$$\mathsf{M}\left(\sum_{l_1=0}^{n-1}\sum_{l_2=0}^{n-1}\cdots\sum_{l_{d-1}=0}^{n-1}A_k^{(s)}(\chi_{l_1})A_k^{(s)}(\bar{\chi}_{l_1}\chi_{l_2})\cdots A_k^{(s)}(\bar{\chi}_{l_{d-1}})\right)=$$

$$=\mathsf{M}\left(\sum_{g_1\in G_n}\cdots\sum_{g_d\in G_n}v_{g_1}^{(s)}\cdots v_{g_d}^{(s)}\sum_{l_1=0}^{n-1}\chi_{l_1}(g_1)\bar{\chi}_{l_1}(g_2)\cdots\sum_{l_{d-1}=0}^{n-1}\chi_{l_{d-1}}(g_{d-1})\bar{\chi}_{l_{d-1}}(g_d)\right),$$

and the latter, applying (C9), turns to be equal to

$$n^{d-1}\,\mathsf{M}\left(\sum_{g\in G_n}[v_g^{(s)}]^d\right),$$

which was to be proved.

Now we are going to compute, by aid of these lemmas, the value of $\mu_2(n,k)$, $\mu_3(n,k)$, $m_4(n,k)$ and $m_2^{(2)}(n,k)$ resp.

REMARK. The results concerning the value of $\mu_2(n,k)$ and $\mu_3(n,k)$ can be found in [1], but in order to show how the method of characters works, we deduce them here too.

LEMMA 3. (Computation of $\mu_2(n,k)$).

$$\mu_2(n,k)=\frac{4^k}{n}+\frac{(n-1)2^k}{n}. \tag{12}$$

PROOF. Making use of (11) for $s=1$, we have

$$\mu_2(n,k)=\frac{1}{n}\,\mathsf{M}\left(\sum_{l=0}^{n-1}A_k(\chi_l)A_k(\bar{\chi}_l)\right). \tag{13}$$

Replacing $A_k(\chi)$ from (9), performing multiplication, we have by the additivity of expectation

$$\mu_2(n,k)=\frac{1}{n}\sum_{l=0}^{n-1}\mathsf{M}\left(\prod_{j=1}^{k}\{2+\chi_l(a_j)+\bar{\chi}_l(a_j)\}\right). \tag{14}$$

Since the a_j's were chosen independently and each can be equal to any element of G_n with probability $\frac{1}{n}$, the factors of the product in (14) are independent and equally distributed random variables, hence

$$\mu_2(n,k)=\frac{1}{n}\sum_{l=0}^{n-1}[2+2\,\mathrm{Re}\,\mathsf{M}(\chi_l(a_j))]^k. \tag{15}$$

Since further

$$\mathsf{M}(\chi_l(a_j))=\frac{1}{n}\sum_{g\in G_n}\chi_l(g),$$

thus by (C5) and (C6), we have

$$\mathsf{M}(\chi_l(a_j))=\delta_{\chi_l}=\begin{cases}1, & \text{if } l=0,\\ 0, & \text{if } l\neq 0\end{cases}$$

which gives immediately (12).

LEMMA 4. (Computation of $\mu_3(n, k)$).

$$\mu_3(n,k) = 2^k + \frac{3\cdot 2^k(2^k-1)}{n} + \frac{2^k(2^k-1)(2^k-2)}{n^2}. \tag{16}$$

PROOF. By the same argument as was used when proving (12), according to (C5), we have

$$\mu_3(n,k) = \frac{1}{n^2}\sum_{l_1=0}^{n-1}\sum_{l_2=0}^{n-1}[2(1+\delta_{\chi_{l_1}}+\delta_{\bar{\chi}_{l_1}\chi_{l_2}}+\delta_{\chi_{l_2}})]^k. \tag{17}$$

Applying (C6) and (C8), we get (16).

LEMMA 5. (Computation of $m_4(n, k)$). *By the notations* (3), (4) *and* (5), *we have*

$$\mu_4(n,k) = 2^k + \frac{7\cdot 2^k(2^k-1)}{n} + \frac{6\cdot 2^k(2^k-1)(2^k-2)+3\cdot 6^k-6\cdot 4^k+3\cdot 2^k}{n^2} + \tag{18}$$

$$+\frac{2^k(2^k-1)(2^k-2)(2^k-3)+(2^r-1)8^k-3\cdot 2^r\cdot 6^k+3(2^r+1)4^k-(2^r+2)2^k}{n^3}$$

and

$$m_4(n,k) = 2^k + \frac{3\cdot 2^{2k}-7\cdot 2^k}{n} + \frac{3\cdot 6^k-12\cdot 2^{2k}+15\cdot 2^k}{n^2} + \tag{19}$$

$$+\frac{-3\cdot 2^r\cdot 6^k+(2^r-1)2^{3k}+(6+3\cdot 2^r)2^{2k}-(8+2^r)2^k}{n^3}.$$

PROOF. From (11), repeating the same arguments applied when proving (12), we have:

$$\mu_4(n,k)\cdot n^3 = \tag{20}$$

$$= \sum_{l_1=0}^{n-1}\sum_{l_2=0}^{n-1}\sum_{l_3=0}^{n-1}[2(1+\delta_{\chi_{l_1}}+\delta_{\chi_{l_2}}+\delta_{\chi_{l_3}}+\delta_{\chi_{l_1}\bar{\chi}_{l_2}}+\delta_{\bar{\chi}_{l_2}\chi_{l_3}}+\delta_{\chi_{l_1}\bar{\chi}_{l_3}}+\delta_{\chi_{l_1}\bar{\chi}_{l_2}\chi_{l_3}})]^k.$$

REMARK. It can be seen from (20), that the value of $\mu_4(n, k)$ depends on the special structure of the group through the term $\delta_{\chi_{l_1}\bar{\chi}_{l_2}\chi_{l_3}}$ which depends just on the group elements of order 2.

By a simple, but cumbersome calculation, applying (C1)—(C9), from (20) we get (18).

Since further

$$\sum_{g\in G_n} \nu_g(k, \omega(k)) = 2^k, \tag{21}$$

we have

$$m_4(n,k) = \mu_4(n,k) - 4\frac{2^k}{n}\mu_3(n,k) + 6\cdot\frac{2^{2k}}{n^2}\mu_2(n,k) - 3\cdot\frac{2^{4k}}{n^3},$$

thus by (12), (16) and (18) we have (19).

LEMMA 6. *We have*

$$m_2^{(2)}(n,k) = \frac{(2^r-1)5^k+(n-2^r)3^k}{n}. \tag{22}$$

PROOF. By (11) we have

$$\mu_2^{(s)}(n,k) = \frac{1}{n}\sum_{l=0}^{n-1}\left[(s+1)+2\sum_{i=1}^{s}(s+1-i)\delta_{\chi_l^i}\right]^k, \tag{23}$$

i.e. for $s=2$, we have

$$\mu_2^{(2)}(n,k) = \frac{1}{n}\sum_{l=0}^{n-1}(3+4\delta_{\chi_l}+2\delta_{\chi_l^2})^k. \tag{24}$$

Let us introduce the following notation:

$$\Gamma_2 = \{t: t \neq 0, \quad \chi_t^2 = \chi_0, \quad \chi_t \in H_n\}.$$

Thus we have

$$\mu_2^{(2}(n,k) = \frac{1}{n}\left\{9^k+\sum_{l\in\Gamma_2}(3+4\delta_{\chi_l}+2\delta_{\chi_l^2})^k+\sum_{\substack{l\notin\Gamma_2\\ l\neq 0}}(3+4\delta_{\chi_l}+2\delta_{\chi_l^2})^k\right\}.$$

It is easy to see that by (5) and (C4)

$$N(\Gamma_2) = 2^r-1,$$

hence it follows

$$\mu_2^{(2)}(n,k) = \frac{9^k+(2^r-1)5^k+(n-2^r)3^k}{n}. \tag{25}$$

Since further

$$\sum_{g\in G_n} v_g^{(s)}(k,\omega(k)) = (s+1)^k, \tag{26}$$

we have

$$m_2^{(s)}(n,k) = \mu_2^{(s)}(n,k) - \frac{(s+1)^{2k}}{n}, \tag{27}$$

which proves (22).

By aid of lemma 6, we can prove the following generalization of Theorem 1 of ERDŐS and RÉNYI:

THEOREM. *Using of the notations introduced above, let us suppose that* $r \leqq \frac{\log n}{2\log 2}$; *let further* ε *and* δ *be arbitrary fixed positive numbers. Then for any integer*

$$k \geqq \max(k_1, k_2), \tag{28}$$

where

$$k_1 \geqq \frac{r\log 2+\log n+2\log\frac{1}{\varepsilon}+\log\frac{2}{\delta}}{\log\frac{9}{5}},$$

$$k_2 \geqq \frac{2\log n+2\log\frac{1}{\varepsilon}+\log\frac{2}{\delta}}{\log 3},$$

we have

$$\mathsf{P}\left(\max_{g\in G_n}\left|\nu_g^{(2)}(k,\omega(k))-\frac{3^k}{n}\right|<\varepsilon\frac{3^k}{n}\right)>1-\delta. \tag{29}$$

In other words every $g\in G_n$ has approximately the same number of representations in the form (2) *(for $s=2$) with probability near to* 1.

REMARK. If $r=0$ (e.g. let G_n be the group of residue classes mod n, where n is an odd number), then (29) can be stated under the condition

$$k\geqq\frac{2\log n+2\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}}{\log 3}. \tag{30}$$

It can be easily seen, that if $r\leqq\frac{\log n}{2\log 2}$, then

$$\max(k_1,k_2)=\begin{cases}k_1, & \text{if} \quad C_1\log n-C_2<r\leqq\frac{\log n}{2\log 2},\\ k_2, & \text{if} \quad 1\leqq r\leqq C_1\log n-C_2,\end{cases} \tag{31}$$

where

$$C_1=\frac{\log 1{,}08}{\log 2\cdot\log 3};\quad C_2=\frac{\left(2\log\frac{1}{\varepsilon}+\log\frac{2}{\delta}\right)\log\frac{5}{3}}{\log 2\cdot\log 3}. \tag{32}$$

PROOF OF THE THEOREM. Following the considerations used by ERDŐS and RÉNYI proving their Theorem 1, we have by (22):

$$\mathsf{P}\left(\max_{g\in G_n}\left|\nu_g^{(2)}-\frac{3^k}{n}\right|\geqq\varepsilon\frac{3^k}{n}\right)\leqq\mathsf{P}\left(\sum_{g\in G_n}\left(\nu_g^{(2)}-\frac{3^k}{n}\right)^2\geqq\varepsilon^2\frac{3^{2k}}{n^2}\right)\leqq \tag{33}$$

$$\leqq\frac{m_2^{(2)}(n,k)\cdot n^2}{\varepsilon^2\cdot 3^{2k}}<\frac{2^r\cdot n}{\varepsilon^2\left(\frac{9}{5}\right)^k}+\frac{n^2}{\varepsilon^2\cdot 3^k}.$$

On the right hand side of the latter inequality both terms are $\leqq\frac{\delta}{2}$ if (28) holds, which proves (29).

SOME REMARKS. We do not know the exact probability distribution of the random variable $\nu_g^{(s)}(k,\omega(k))$; it seems to be rather difficult to find it.

It is easy to see, that if g_1 and g_2 are different elements of G_n, then the distributions of ν_{g_1} and ν_{g_2}, in general, are different. Let us consider e.g. the group of residue classes mod 4 and let be $k=3$. Then, denoting the group elements by 0, 1, 2, 3, the distribution of ν_1 and ν_2 are different, namely

$$\mathsf{P}(\nu_1=0)=\frac{14}{64};\quad\mathsf{P}(\nu_2=0)=\frac{13}{64}.$$

It would be interesting to decide the necessary and sufficient condition that for arbitrary group elements $g_1 \in G_n$, $g_2 \in G_n$, ν_{g_1} and ν_{g_2} should be equally distributed. A trivial sufficient condition is the following: if there exists an automorphism T of G_n for which $Tg_1 = g_2$, then ν_{g_1} and ν_{g_2} are equally distributed.

It is easy to give an example such that ν_{g_1} and ν_{g_2} are equally distributed for some k, and there does not exists such an automorphism T for which $Tg_1 = g_2$. Let us consider the previous example and let be $k = 2$. Then ν_1 and ν_2 are equally distributed, namely

$$\mathsf{P}(\nu_1 = 0) = \mathsf{P}(\nu_2 = 0) = \frac{7}{16}; \quad \mathsf{P}(\nu_1 = 1) = \mathsf{P}(\nu_2 = 1) = \frac{6}{16}; \quad \mathsf{P}(\nu_1 = 2) = \mathsf{P}(\nu_2 = 2) = \frac{3}{16},$$

but there does not exists such an automorphism which maps 1 into 2. We can not, however, give an example having the above property for *all* k.

We mention finally that the following relation holds between $\nu_g^{(s)}(k, \omega(k))$ and the quantity introduced in (9):

$$\nu_g^{(s)} = \frac{1}{n} \sum_{\chi \in H_n} A_k^{(s)}(\chi) \overline{\chi(g)}, \tag{34}$$

which is a trivial consequence of lemma 1.

REFERENCES

[1] ERDŐS, P. and RÉNYI, A.: Probabilistic methods in group theory, *J. Analyse Math.* **14** (1965) 127—138.
[2] MIECH, R. J.: On a conjecture of Erdős and Rényi, *Illinois J. Math.* **11** (1967) 114—127.
[3] KOLMOGOROV, A. N.: *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin, 1933.
[4] HALL, M.: *The theory of groups*, New York, 1959.

Eötvös Loránd University, Budapest

(Received January 14, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 37—40.

ON THE TWO-STAGE PROGRAMMING UNDER UNCERTAINTY

by

G. KÉRI

The problem of two-stage programming under uncertainty formulated first by DANTZIG and MADANSKY is the following:

$$\min z(\mathbf{x}) = \mathbf{c}'\mathbf{x} + \mathsf{E}\{\min \mathbf{q}'\mathbf{y} \mid T\mathbf{x} + M\mathbf{y} = \xi, \mathbf{y} \geqq \mathbf{0}\}$$

$$\text{subject to } A\mathbf{x} = \mathbf{b} \tag{1}$$

$$\mathbf{x} \geqq \mathbf{0},$$

where A, T and M are $m \times n$, $m' \times n$ and $m' \times n'$ matrices respectively, $\mathbf{b}$ is an m-component vector, ξ is a random m'-component vector, the range of which is Ξ, $\mathbf{x}$ is an n-component vector and $\mathbf{y}$ is an n'-component vector.

The second stage of the problem is:

$$\min \mathbf{q}'\mathbf{y}$$

$$\text{subject to } M\mathbf{y} = \xi - T\mathbf{x} \tag{2}$$

$$\mathbf{y} \geqq \mathbf{0}.$$

A vector $\mathbf{x}$ is said to be feasible if $A\mathbf{x} = \mathbf{b}$, $\mathbf{x} \geqq \mathbf{0}$ and for every $\xi \in \Xi$ there exists an optimal solution $\mathbf{y}$ to (2). Let us denote by K_1 the set of those $\mathbf{x}$ for which $A\mathbf{x} = \mathbf{b}$, $\mathbf{x} \geqq \mathbf{0}$ and by K_2 the set of those $\mathbf{x}$ for which it is true that for every $\xi \in \Xi$ there exists an optimal solution to (2). Let further $K = K_1 \cap K_2$ be the set of feasible solutions.

I only deal with the case of Ξ is the whole m'-dimensional Euclidian space denoted by $R^{m'}$. This is for example the situation if the distribution of ξ is normal. If the range of ξ is the whole space and ξ is varying through $R^{m'}$ by fixed $\mathbf{x}$ then $\xi - T\mathbf{x}$ can be arbitrary vector of $R^{m'}$ too. In this way in the case of $\Xi = R^{m'}$ K_2 is either the empty set or the whole n-dimensional Euclidian space. In the second case $K = K_1 = \{\mathbf{x} | A\mathbf{x} = \mathbf{b}, \mathbf{x} \geqq \mathbf{0}\}$. In what follows I shall deal with necessary and sufficient condition for K_2 to be the whole R^n space, viz. the linear programming problem

$$\min \mathbf{q}'\mathbf{y}$$

$$\text{subject to } M\mathbf{y} = \mathbf{t} \tag{3}$$

$$\mathbf{y} \geqq \mathbf{0}$$

depending on the parameter $\mathbf{t}$ has an optimal solution for every $\mathbf{t}$. Because in what follows the vectors $\mathbf{b}$ and $\mathbf{x}$ will not be used, the dimension of $\mathbf{t}$ and $\mathbf{y}$ will be denoted instead of m' and n' by m and n, respectively.

DEFINITION. The convex cone hull of a set H in the space R^n is

$$H^< = \left\{\mathbf{x} | \mathbf{x} = \sum_{i=1}^{s} \lambda_i \mathbf{x}_i, \quad s = 1, 2, 3, \ldots, \quad \mathbf{x}_i \in H, \quad \lambda_i \geqq 0 \qquad (i = 1, 2, \ldots, s)\right\}.$$

In particular for a finite $H = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$,

$$H^< = \left\{\mathbf{x} | \mathbf{x} = \sum_{i=1}^{N} \lambda_i \mathbf{x}_i, \quad \lambda_i \geqq 0 \qquad (i = 1, 2, \ldots, N)\right\}$$

PROPOSITION I *(triviality): The vector equation* $M\mathbf{y} = \mathbf{t}$ *has a nonnegative solution* $\mathbf{y}$ *for every* $\mathbf{t} \in R^n$ *if and only if the convex cone hull of the column-vectors of* M *is the whole* R^m *space.*

DEFINITION. The polar cone of a set H in the space R^n is

$$H^* = \{\mathbf{x} | \mathbf{x}'\mathbf{y} \leqq 0 \quad \text{for every} \quad \mathbf{y} \in H\}$$

PROPOSITION II. *The problem* (3) *has an optimal solution for every* $\mathbf{t} \in R^m$ *if and only if the convex cone hull of the column-vectors of* M *is the whole* R^m *space, and the vector* $-\mathbf{q}$ *is contained by the polar cone of the cone*

$$\{\mathbf{y} | M\mathbf{y} = \mathbf{0}, \mathbf{y} \geqq \mathbf{0}\}.$$

To prove this proposition we need two lemmas from which, taking into account even the proposition I, the statement evidently follows.

LEMMA 1. If the problem (3) has a feasible solution for every $\mathbf{t} \in R^m$ and for a $\mathbf{t}$ has an optimal one too then (3) has an optimal solution for every $\mathbf{t} \in R^m$.

PROOF. To the contrary, let us assume that when $\mathbf{t} = \mathbf{t}_1$ there is a finite minimum and when $\mathbf{t} = \mathbf{t}_2$ there is not. Let $\mathbf{t}_0 \in R^m$ such that $\mathbf{t}_1$ is an interior point of the interval $(\mathbf{t}_0, \mathbf{t}_2)$ i.e. $\mathbf{t}_1 = (1-\lambda)\mathbf{t}_0 + \lambda\mathbf{t}_2$ where $0 < \lambda < 1$. According to the assumption there exists $\mathbf{y}_0 \in R^n$ for which $M\mathbf{y}_0 = \mathbf{t}_0$, $\mathbf{y}_0 \geqq \mathbf{0}$ further for any natural number N there exists $\mathbf{y}_{2,N} \in R^n$ for which $M\mathbf{y}_{2,N} = \mathbf{t}_2$, $\mathbf{y}_{2,N} \geqq \mathbf{0}$ and $\mathbf{q}'\mathbf{y}_{2,N} < -N$. Obviously $\mathbf{y}_{1,N} = (1-\lambda)\mathbf{y}_0 + \lambda\mathbf{y}_{2,N}$ is a feasible solution to (3) when $\mathbf{t} = \mathbf{t}_1$ for which

$$\mathbf{q}\mathbf{y}_{1,N} = (1-\lambda)\mathbf{q}'\mathbf{y}_0 + \lambda\mathbf{q}'\mathbf{y}_{2,N} < (1-\lambda)\mathbf{q}'\mathbf{y}_0 - \lambda N$$

and so $\lim_{N \to \infty} \mathbf{q}'\mathbf{y}_{1,N} = -\infty$. It means that the problem (3) has no optimal solution for $\mathbf{t} = \mathbf{t}_1$ which contradicts to the assumptions.

LEMMA 2. For $\mathbf{t} = \mathbf{0}$ (3) has an optimal solution if and only if $-\mathbf{q}$ is contained by the polar cone of the set

$$\{\mathbf{y} | M\mathbf{y} = \mathbf{0}, \mathbf{y} \geqq \mathbf{0}\}.$$

PROOF. If $\mathbf{q}'\mathbf{y} \geqq \mathbf{0}$ for every $\mathbf{y}$ for which $M\mathbf{y} = \mathbf{0}$ and $\mathbf{y} \geqq \mathbf{0}$ then $\mathbf{0}$ is an optimal solution to the problem

$$\begin{aligned} &\min \mathbf{q}'\mathbf{y} \\ &\text{subject to } M\mathbf{y} = \mathbf{0} \\ &\mathbf{y} \geqq \mathbf{0} \end{aligned} \tag{4}$$

If on the other hand there exists a feasible solution $\mathbf{y}_0$ to (4) for which $\mathbf{q}'\mathbf{y}_0 < 0$, then, since $\lambda\mathbf{y}_0$ is also a feasible solution to (4) for every positive number λ, the minimum of the objective function of (4) is not finite.

PROPOSITION III. *Let $n = m+1$. Denote by $\boldsymbol{\mu}_i$ the i-th column-vector of the matrix M and by B_i the matrix obtained from M by omitting the i-th column. Let finally $\beta_i = (-1)^{i+1} \det B_i$. In this case the vector equation $M\mathbf{y} = \mathbf{t}$ has a nonnegative solution for every $\mathbf{t} \in R^n$ if and only if either all β_i are negative or all β_i are positive.*

PROOF. NECESSITY. According to the first proposition it is a necessary condition that the convex cone hull of the vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{m+1}$ be the whole R^n space. For this it is necessery that having chosen any m vectors from $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{m+1}$ these vectors be linearly independent i.e. the numbers β_i should be different from 0. In the other case there is a semispace in R^m containing the origin on its boundary which contains every $\boldsymbol{\mu}_i$, so it is impossible that the convex cone hull of the vectors $\boldsymbol{\mu}_i$ be the whole space. Since the convex cone hull of m vectors from the space R^m cannot be the whole space too, it is also necessary that not any of the $\boldsymbol{\mu}_i$'s be nonnegative linear combination of the others i.e. the orientation of two vector-series both of m different vectors from $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{m+1}$ must be opposite if these series differ one from another only at one place. This fact exactly means that the numbers β_i must be either all positive or all negative.

SUFFICIENCY. Let us assume that either all β_i are positive or all are negative. Then every component of the unique solution of the equation $B_1\mathbf{u}_1 = \boldsymbol{\mu}_1$ as opposite of the quotient of two β_i is negative, accordingly $\boldsymbol{\mu}_1 = \sum_{j=2}^{m+1} \gamma_j \boldsymbol{\mu}_j$ where $\gamma_j < 0$ for every j. Now let $\mathbf{v}$ be an arbitrary vector of the m-dimensional Euclidian space. Owing to $\beta_i \neq 0$

$$\mathbf{v} = \sum_{j=2}^{m+1} \delta_j \boldsymbol{\mu}_j \tag{5}$$

where among the numbers δ_j some of them can be negative. From (5) we get if we consider also $\boldsymbol{\mu}_1 = \sum_{j=2}^{m+1} \gamma_j \boldsymbol{\mu}_j$ that

$$\mathbf{v} = \lambda\boldsymbol{\mu}_1 + \sum_{j=2}^{m+1} (\delta_j - \lambda\gamma_j)\boldsymbol{\mu}_j. \tag{6}$$

If λ is great enough then by (6) $\mathbf{v}$ is made as a nonnegative linear combination of the $\boldsymbol{\mu}_i$'s. Since $\mathbf{v}$ could be arbitrary vector, accordingly the first proposition the vector equation $M\mathbf{y} = \mathbf{t}$ has a nonnegative solution for every $\mathbf{t} \in R^m$.

PROPOSITION IV. *In the case of $n = m+1$, (3) has an optimal solution for every $\mathbf{t} \in R^m$ if and only if one of the two systems of inequalities is satisfied:*

$$\beta_1 > 0, \beta_2 > 0, \ldots, \beta_{m+1} > 0, \quad q_1\beta_1 + q_2\beta_2 + \ldots + q_{m+1}\beta_{m+1} \geqq 0$$

or

$$\beta_1 < 0, \beta_2 < 0, \ldots, \beta_{m+1} < 0, \quad q_1\beta_1 + q_2\beta_2 + \ldots + q_{m+1}\beta_{m+1} \leqq 0.$$

The statement of this proposition follows from the statements of the third proposition, the first lemma and the next third lemma.

LEMMA 3. In case of $n = m+1$ by the assumption that either all β_i are positive or all are negative, (4) has an optimal solution if and only if either the sign of the sum $\sum_{j=1}^{m+1} q_j\beta_j$ is equal to the sign of the numbers β_i or the value of this sum is zero.

PROOF. If (4) has an optimal solution then it has a basis satisfying the optimality criterion, since every β_i is different from 0. In our case the optimality criterion for a basis

$$B_i = \{\mathbf{\mu}_1, \mathbf{\mu}_2, \ldots, \mathbf{\mu}_{i-1}, \mathbf{\mu}_{i+1}, \ldots, \mathbf{\mu}_m, \mathbf{\mu}_{m+1}\}$$

is the following

$$\mathbf{q}'_{B_i} B_i^{-1} \mathbf{\mu}_i \leqq q_i \tag{7}$$

where

$$\mathbf{q}'_{B_i} = [q_1, q_2, \ldots, q_{i-1}, q_{i+1}, \ldots, q_m, q_{m+1}].$$

As $\mathbf{u}_i = B_i^{-1}\mathbf{\mu}_i$ is the solution of the equation $B_i\mathbf{u}_i = \mathbf{\mu}_i$,

$$\mathbf{u}'_i = \left[-\frac{\beta_1}{\beta_i}, -\frac{\beta_2}{\beta_i}, \ldots, -\frac{\beta_{i-1}}{\beta_i}, -\frac{\beta_{i+1}}{\beta_i}, \ldots, -\frac{\beta_{m+1}}{\beta_i}\right]$$

By converting (7) we get the inequality $\sum_{j=1}^{m+1} q_j\beta_j \geqq 0$ or $\sum_{j=1}^{m+1} q_j\beta_j \leqq 0$ depending on whether $\beta_i > 0$ or $\beta_i < 0$. Therefore in our case the optimality criterion is the same for every basis namely $\sum_{j=1}^{m+1} q_j\beta_j \geqq 0$ (all β_j are positive) or $\sum_{j=1}^{m+1} q_j\beta_j \leqq 0$ (all β_j are negative) is a necessary and sufficient condition for the existence of an optimal solution to (4).

REFERENCES

[1] DANTZIG, G. B. and MADANSKY, A.: On the Solution of Two-stage Linear Programs under Uncertainty. *The RAND Corporation paper*, P—2039, 28 July, 1960.

[2] WETS, R.: Programming under Uncertainty; The equivalent convex program, *SIAM J. Appl. Math.* **14** (1966) 89—105.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received January 14, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 41—49

ÜBER DIE ENTFERNUNG DER IRRFAHRTSWEGE

von

P. BÁRTFAI

In meiner Arbeit [1] habe ich bewiesen, daß das sogenannte Geysir-Problem im Falle $C_n = o(\log n)$ gelöst werden kann. Ausführlicher seien $\xi_1, \xi_2, \ldots$ unabhängige Zufallsveränderliche mit derselben Verteilungsfunktion $F(x)$. Das verallgemeinerte Geysir-Problem besteht darin, ob $F(x)$ durch eine einzige unendliche Realisation des Prozesses $\xi_1 + \chi_1, \xi_1 + \xi_2 + \chi_2, \ldots, \xi_1 + \cdots + \xi_n + \chi_n, \ldots$ bestimmt ist, wo χ_n eine beschränkte Zufallsveränderliche, ein sogenanntes Fehlerglied ist. (Die Unabhängigkeit der Zufallsveränderlichen χ_n von der Folge $\xi_1, \xi_2, \ldots$ ist nicht vorausgesetzt.)

Nun werden wir untersuchen, ob die Größenordnung $C_n = o(\log n)$ verbessert werden kann. Das führt uns zum Studium der Entfernung der Irrfahrtswege. Es sei $\eta_1, \eta_2, \ldots$ eine Folge der unabhängigen Zufallsveränderlichen mit derselben Verteilungsfunktion $G(x)$ und wir setzen $S_n = \sum_{k=1}^{n} \xi_k$, $T_n = \sum_{k=1}^{n} \eta_k$. Wenn die Abhängigkeit zwischen der zwei Folgen $\xi_1, \xi_2, \ldots$ und $\eta_1, \eta_2, \ldots$ derart gegeben werden kann, daß

$$\overline{\lim_{n\to\infty}} \frac{|S_n - T_n|}{f(n)} < K_1 \tag{1}$$

mit positiver Wahrscheinlichkeit gilt, so kann das verallgemeinerte Geysir-Problem im Falle $C_n = O(f(n))$ offenbar nicht gelöst werden. Statt (1) werden wir eine stärkere Relation erfordern:

$$\lim_{n\to\infty} \frac{S_n - T_n}{f(n)} = 0 \tag{2}$$

mit Wahrscheinlichkeit 1. Da die Folge $S_n - T_n$ nicht unbedingt eine $0-1$ Folge ist (d.h. das Null- oder Eins-Gesetz nicht unbedingt erfüllt ist), wollen wir wirklich mehr erreichen, wenn wir die Gültigkeit der Relationen (1) oder (2) mit Wahrscheinlichkeit 1 erfordern. Dieses Problem läßt sich leichter lösen und führt zu einem neuen Themakreis. Die Lösung gibt eine obere Schätzung für die maximale Größenordnung von C_n.

Vor allem beschäftigen wir uns mit der Untersuchung des Korrelationskoeffizients $r_n = \mathrm{R}(S_n, T_n)$. Unser Ziel ist zu erreichen, daß $1 - r_n$ je schneller zur Null streben soll. Zuerst beweisen wir einen Satz darüber, wie groß der Korrelationskoeffizient der zwei Zufallsveränderlichen sein kann, wenn die marginalen Verteilungen gegeben sind.

Es sei $F(x)$ bzw. $G(x)$ die Verteilungsfunktion von ξ bzw. η. Wir bezeichnen

die Umkehrfunktion der $F(x)$ mit $F^{-1}(y)$, wenn $F(x)$ eine streng monotone und stetige Funktion ist, sonst wenden wir die Definition

$$F_n^{-1}(y) = \sup \{x : F(x) \leqq y\} \tag{3}$$

an. $G_n^{-1}(y)$ wird ähnlicherweise definiert.

SATZ 1.

$$|\mathsf{M}(\xi\eta)| \leqq \int_0^1 F^{-1}(y) G^{-1}(y)\, dy \tag{3a}$$

und die Gleichung besteht, wenn

$$\eta = G^{-1}(\tilde{F}(\xi)), \tag{4}$$

wobei

$$\tilde{F}(\xi) = F(\xi) + \alpha\big(F(\xi+0) - F(\xi)\big) \tag{4a}$$

und α eine von ξ unabhängige, in dem Intervall [0, 1] *gleichverteilte Zufallsveränderliche ist.*

Daraus folgt schon, wenn ξ und η standardisiert sind, so ist

$$\varrho = \int_0^1 F^{-1}(y) G^{-1}(y)\, dy \tag{5}$$

das Maximum des Korrelationskoeffizients $\mathsf{R}(\xi, \eta) = \mathsf{M}(\xi\eta)$ im Falle der gegebenen marginalen Verteilungen. Dieses Maximum kann durch die Transformation (4) erreicht werden, diese Transformation wird Quantiltransformation genannt.

Der Beweisentwurf stammt von J. BASS [2].

Man kann die Ausdrücke $1-\varrho$ und $\sqrt{1-\varrho}$ als eine Maßzahl der Entfernung der standardisierten Verteilungsfunktionen $F(x)$ und $G(x)$ interpretieren. Da

$$\sqrt{1-\varrho} = \left[\frac{1}{2}\int_0^1 (F^{-1}(y) - G^{-1}(y))^2\, dy\right]^{\frac{1}{2}}$$

ist, kann man leicht sehen, daß $\sqrt{1-\varrho}$ zugleich auch ein Metrik ist.

Bezeichnen wir die n-te standardisierte Faltungspotenz von $F(x)$ mit $F_n(x)$. Wir werden die vorher genannte Entfernung der $F_n(x)$ und der normalen Verteilungsfunktion untersuchen.

SATZ 2. *Existiert die momenterzeugende Funktion von $F(x)$, so gilt für beliebiges $r > 0$ und $\varepsilon > 0$*

$$\int_0^1 (F_n^{-1}(y) - \Phi^{-1}(y))^r\, dy \leqq \frac{K_2}{n^{\frac{r}{2}-\varepsilon}}. \tag{6}$$

Vielleicht kann man in diesem Satz die Größenordnung $O(n^{-\frac{r}{2}})$ mit anderen Methoden erreichen.

Auf diesem Satz beruht der

SATZ 3. *Existieren die momenterzeugenden Funktionen der Zufallsveränderlichen* ξ_1 *und* η_1 *so kann die stochastische Abhängigkeit zwischen den Folgen* $\xi_1, \xi_2, \ldots$ *und* $\eta_1, \eta_2, \ldots$ *derart gegeben werden, daß*

$$\lim_{n\to\infty} \frac{S_n - T_n}{n^{\frac{1}{4}+\delta}} = 0 \tag{7}$$

für beliebiges $\delta > 0$ *mit Wahrscheinlichkeit 1 gilt.*

FOLGERUNG. *Das verallgemeinerte Geysir-Problem kann im Falle* $C_n = O(n^{\frac{1}{4}+\delta})$ *nicht gelöst werden, d.h. eine einzige Realisation des Prozesses* $\xi_1 + \chi_1, \xi_1 + \xi_2 + \chi_2, \ldots, \xi_1 + \xi_2 + \cdots + \xi_n + \chi_n, \ldots$ *bestimmt die Verteilungsfunktion* $F(x)$ *nicht. Die Verteilungsfunktion von* ξ_1 *kann eine beliebige Verteilung sein, die eine momenterzeugende Funktion hat.*

Obgleich das Geysir-Problem laut des Satzes 3 nur im Falle $C_n = O(n^{\frac{1}{4}+\delta})$ unlösbar ist, kann die Größenordnung $C_n = O(\log n)$ — meiner Meinung nach — wahrscheinlich nicht verbessert werden.

Beweise:

BEWEIS VON SATZ 1. Wir werden nun den Beweisentwurf von J. BASS [2] ausführlicher schildern.

Es sei $H(x, y)$ eine beliebige zweidimensionale Verteilungsfunktion mit den Eigenschaften $H(x, \infty) = F(x)$, $H(\infty, y) = G(y)$. Wegen der Monotonität ist $H(x, y) \leqq F(x)$ und $H(x, y) \leqq G(y)$, also

$$H(x, y) \leqq \min\big(F(x), G(y)\big) = H_1(x, y).$$

$\mathsf{M}(\xi\eta)$ kann durch partielle Integration zur folgenden Gestalt gebracht werden:

$$\mathsf{M}(\xi\eta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, d^2 H(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H_0(x, y)\, dx\, dy,$$

wobei

$$H_0(x, y) = \begin{cases} H(x, y) & \text{für } x < 0,\ y < 0, \\ H(x, y) - F(x) & \text{für } x < 0,\ y \geqq 0, \\ H(x, y) - G(y) & \text{für } x \geqq 0,\ y < 0, \\ H(x, y) - F(x) - G(y) + 1 & \text{für } x \geqq 0,\ y \geqq 0 \end{cases}$$

gesetzt wurde. Daraus ergibt sich unmittelbar die Ungleichung

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, d^2 H(x, y) \leqq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, d^2 H_1(x, y). \tag{8}$$

Nun werden wir beweisen, daß $H_1(x, y)$ eben die gemeinsame Verteilungsfunktion von ξ, und $\eta^* = G^{-1}\big(\tilde{F}(\xi)\big)$ ist. Es ist bekannt, wenn β eine gleichverteilte Zufallsveränderliche auf dem Intervall [0, 1] ist, so gilt

$$\mathsf{P}\big(F^{-1}(\beta) < x\big) = F(x),$$

$$\mathsf{P}\big(G^{-1}(\beta) < x\big) = G(x),$$

daraus folgt

$$\mathsf{P}\left(F^{-1}(\beta)<x,\ G^{-1}(\beta)<y\right)=\min\left(F(x),G(y)\right)=H_1(x,y). \tag{9}$$

β sei $\tilde{F}(\xi)$ ($\tilde{F}(\xi)$ wurde so definiert um eine gleichverteilte Zufallsveränderliche zu sein), so ist $\mathsf{P}[F^{-1}(\tilde{F}(\xi))=\xi]=1$, also

$$\mathsf{P}\left(\xi<x,\ G^{-1}(\tilde{F}(\xi))<y\right)=H_1(x,y).$$

Um (3a) zu beweisen, ist es noch notwendig die rechte Seite von (8) umzugestalten. Aber aus (9) folgt

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\,d^2H_1(x,y)=\mathsf{M}\left(F^{-1}(\beta)G^{-1}(\beta)\right)=\int_0^1 F^{-1}(y)G^{-1}(y)\,dy,$$

womit Satz 1 bewiesen ist.

Beweis von Satz 2. Wir werden den Satz nur im Falle $r=2$ beweisen, im Falle $r>2$ kann man den Beweis ähnlich durchführen. Aus der Berry—Esseenschen Ungleichung folgt

$$\int_{\delta}^{1-\delta}\left(F_n^{-1}(y)-\Phi^{-1}(y)\right)^2 dy=O\left(\frac{1}{n}\right). \tag{10}$$

Es ist genügend die Gleichung

$$\int_{1-\delta}^{1}\left(F_n^{-1}(y)-\Phi^{-1}(y)\right)^2 dy=O\left(\frac{1}{n^{1-\varepsilon}}\right)$$

zu beweisen, da die Schätzung auf dem Intervall $(0,\delta)$ ähnlich durchgeführt werden kann. Führen wir die Bezeichnungen $G_n(x)=1-F_n(x)$, $\Psi(x)=1-\Phi(x)$ ein, so ist

$$\int_{1-\delta}^{1}\left(F_n^{-1}(y)-\Phi^{-1}(y)\right)^2 dy=\int_0^{\delta}\left(G_n^{-1}(y)-\Psi^{-1}(y)\right)^2 dy,$$

wobei (entsprechend (3))

$$G_n^{-1}(y)=\sup\{x:G_n(x)\geqq y\}$$

ist.

Bezeichnen wir die momenterzeugende Funktion von $F(x)$ mit $R(t)$, so ist

$$R^n\left(\frac{t}{\sqrt{n}}\right)=\int_{-\infty}^{\infty}e^{tu}\,dF_n(u)\geqq\int_x^{\infty}e^{tu}\,dF_n(u)\geqq e^{tx}G_n(x).$$

Daraus folgt genügend großes n

$$G_n(x)\leqq e^{-tx}R^n\left(\frac{t}{\sqrt{n}}\right)\leqq 2e^{-x},$$

es gilt nämlich

$$R^n\left(\frac{t}{\sqrt{n}}\right)\to e^{\frac{t^2}{2}}<2$$

im Falle $t=1$. So gilt für genügend großes n

$$\int_0^{2/n} (G_n^{-1}(y))^2\,dy \leqq \int_0^{2/n} \log^2 \frac{y}{2}\,dy \leqq K_3 \frac{\log^2 n}{n} < K_3 \frac{1}{n^{1-\varepsilon}},$$

daraus folgt

$$\int_0^{2/n} (G_n^{-1}(y) - \Psi^{-1}(y))^2\,dy < \frac{K_4}{n^{1-\varepsilon}}. \tag{11}$$

HILFSSATZ. *Existiert die momenterzeugende Funktion von* $F(x)$, *so gilt für* $0 \leqq x \leqq \log n$

$$|G_n(x) - \Psi(x)| \leqq \frac{K_5}{n^{\frac{1-\varepsilon}{2}}} \Psi(x).$$

BEWEIS des Hilfssatzes. Laut des Grenzverteilungssatzes über die großen Abweichungen ist

$$G_n(x) = H_n(x)\left(1 + O\left(\frac{x}{\sqrt{n}}\right)\right),$$

wobei

$$H_n(x) = \Psi(x) e^{\frac{x^3}{\sqrt{n}}\lambda\left(\frac{x}{\sqrt{n}}\right)}$$

gesetz wurde. Wir können die folgenden Schätzungen leicht durchführen:

$$|G_n(x) - H_n(x)| \leqq K_6 H_n(x) \frac{x}{\sqrt{n}} \leqq K_6 \Psi(x) \frac{1}{n^{\frac{1-\varepsilon}{2}}},$$

$$|H_n(x) - \Psi(x)| = \Psi(x)\left|e^{\frac{x^3}{\sqrt{n}}\lambda\left(\frac{x}{\sqrt{n}}\right)} - 1\right| \leqq \Psi(x) K_7 \frac{x^3}{\sqrt{n}} \leqq \Psi(x) \frac{K_7}{n^{\frac{1-\varepsilon}{2}}},$$

und damit ist der Hilfssatz bewiesen.

Nun werden wir den Beweis von Satz 2 forsetzen. Führen wir die folgenden Bezeichnungen ein:

$$d = \frac{K_5}{n^{\frac{1}{2}-\frac{\varepsilon}{2}}},$$

$$\Psi_1(x) = (1-d)\Psi(x),$$

$$\Psi_2(x) = (1+d)\Psi(x),$$

so folgt aus dem Hilfssatz (vgl. die Abbildung)

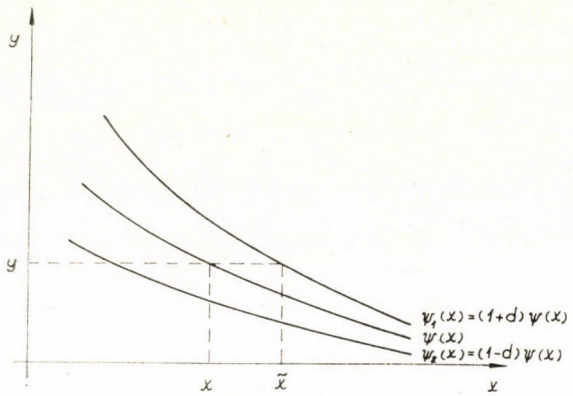
$$|G_n^{-1}(y) - \Psi^{-1}(y)| \leqq \Psi_2^{-1}(y) - \Psi_1^{-1}(y) < -(\Psi_2(\tilde{x}) - \Psi_1(\tilde{x}))\frac{1}{\Psi'(\tilde{x})} = \tag{12}$$

$$= -\frac{2d}{1-d}\frac{\Psi(\tilde{x})}{\Psi'(\tilde{x})} \qquad \left(\frac{2}{n} \leqq y \leqq \delta\right),$$

wobei $\tilde{x}$ durch die Gleichung

$$y = (1+d)\Psi(\tilde{x})$$

bestimmt wird. Man darf den Hilfssatz hier anwenden, denn aus der Relation $y \geqq \frac{2}{n}$ folgt



$\tilde{x} < \log n$, es gilt nämlich

$$\frac{2}{n} \leqq (1+d)\Psi(\tilde{x}) \leqq (1+d)\frac{1}{\sqrt{2\pi}\tilde{x}} e^{-\frac{\tilde{x}^2}{2}} \leqq K_8 e^{-\frac{\tilde{x}^2}{2}}$$

und daraus folgt für genügend großes n

$$\tilde{x} \leqq \sqrt{2 \log K_8 \frac{n}{2}} < \log n.$$

Aus (12) ergibt sich

$$(13) \qquad \int_{2/n}^{\delta} (G_n^{-1}(y) - \Psi^{-1}(y))^2 \, dy \leqq -\left(\frac{2d}{1-d}\right)^2 \int_0^{\infty} \left(\frac{\Psi(\tilde{x})}{\Psi'(\tilde{x})}\right)^2 (1+d)\Psi'(\tilde{x}) \, d\tilde{x} \leqq$$

$$\leqq -d^2 \frac{4(1+d)}{(1-d)^2} \int_0^{\infty} \frac{\Psi^2(\tilde{x})}{\Psi'(\tilde{x})} \, d\tilde{x} \leqq \frac{K_9}{n^{1-\varepsilon}}.$$

Aus den Schätzungen (10), (11) und (13) folgt der Satz 2 (im Falle $r=2$) unmittelbar.

BEWEIS VON SATZ 3. Es sei $n_k = [k^{2+4\delta}]$ und führen wir die folgenden Bezeichnungen ein:

$$U_k = \sum_{i=n_k+1}^{n_{k+1}} \xi_i,$$

$$V_k = \sum_{i=n_k+1}^{n_{k+1}} \eta_i,$$

$$W_k = U_k - V_k$$

so gilt

$$\Delta_k = n_{k+1} - n_k \leqq K_{10} k^{1+4\delta}.$$

Es seien die Zufallsveränderlichen $\xi_1, \xi_2, \ldots$ auf der Wahrscheinlichkeitsalgebra $(\Omega, \mathcal{F}, \mathsf{P})$ gegeben. Zu diesen Zufallsveränderlichen werden die unabhängigen Zufallsveränderlichen $\eta_1, \eta_2, \ldots$ mit derselben Verteilung den Forderungen des Satzes 3 gemäß konstruiert. Wir bezeichnen die Verteilungsfunktion η_1 mit $G(x)$.

Zuerst bestimmen wir V_k mit der Quantiltransformation aus U_k:

$$V_k = G^{(\Delta_k)-1}\left(\tilde{F}^{(\Delta_k)}(U_k)\right),$$

wobei $F^{(\Delta_k)}$ bzw. $G^{(\Delta_k)}$ die Δ_k-te Faltungspotenz von F bzw. G bedeutet. $\tilde{F}^{(\Delta_k)}(U_k)$ wurde mit (4a) definiert. (Wenn $F^{(\Delta_k)}$ nicht stetig ist, so erfordert diese Transformation die Erweiterung von Ω, aber sie wird ohnehin notwendig in den Folgenden. Wir werden dieses Verfahren nicht detaillieren.) Dann müssen wir noch die Darstellung $V_k = \eta_{n_k+1} + \cdots + \eta_{n_{k+1}}$ geben. Da die gemeinsame Verteilungsfunktion von $\eta_{n_k+1}, \ldots, \eta_{n_{k+1}}$ bekannt ist,

$$\mathsf{P}(\eta_{n_k+1} < x_{n_k+1}, \ldots, \eta_{n_{k+1}} < x_{n_{k+1}}) = \prod_{j=n_k+1}^{n_{k+1}} G(x_j), \tag{14}$$

kann man auch die bedingte Verteilungsfunktion

$$\mathsf{P}(\eta_{n_k+1} < x_{n_k+1}, \ldots, \eta_{n_{k+1}} < x_{n_{k+1}} | V_k = y) \tag{15}$$

als bekannt betrachten. Ist $V_k = y$, so wählen wir die $\eta_{n_k+1}, \ldots, \eta_{n_{k+1}}$ derart, daß (15) ihre gemeinsame (bedingte) Verteilungsfunktion sei. In solcher Weise wird die gemeinsame Verteilungsfunktion dieser Zufallsveränderlichen (ohne Bedingung) genau (14) ergeben, sie werden also unabhängig mit derselben Verteilungsfunktion $G(x)$ sein. Führen wir diese Konstruktion für alle k unabhängig voneinander durch, so haben wir die ganze Folge $\eta_1, \eta_2, \ldots$ konstruiert.

Aus den Sätzen 1 und 2 ergibt sich

$$\begin{aligned} \mathsf{D}^2(W_k) = \mathsf{D}^2(U_k - V_k) &= 2\Delta_k\left(1 - \mathsf{R}(U_k, V_k)\right) = \\ &= \Delta_k \int_0^1 \left(F_{\Delta_k}^{-1}(y) - G_{\Delta_k}^{-1}(y)\right)^2 dy \leqq \\ &\leqq 2\Delta_k \left\{ \int_0^1 \left(F_{\Delta_k}^{-1}(y) - \Phi^{-1}(y)\right)^2 dy + \int_0^1 \left(G_{\Delta_k}^{-1}(y) - \Phi^{-1}(y)\right)^2 dy \right\} \leqq \\ &\leqq K_{11} \Delta_k \frac{1}{\Delta_k^{1-\varepsilon_1}} \leqq K_{12} k^{\varepsilon_2}, \end{aligned} \tag{16}$$

wobei $\varepsilon_2 > 0$ beliebig klein sein kann und K_{12} nur von ε_2 abhängt. Wir werden beweisen, daß

$$\frac{1}{n_k^{1/4+\delta}} \sum_{j=1}^{k} W_j \to 0 \qquad (k \to \infty). \tag{17}$$

Statt (17) ist es notwendig zu beweisen, daß

$$\frac{1}{n^{\alpha}} \sum_{j=1}^{n} W_j \to 0 \qquad (n \to \infty), \tag{18}$$

wobei $\alpha = \frac{1}{2} + 3\delta$ gesetzt wurde.

Da die Variablen W_j unabhängig sind, können wir die KOLMOGOROFFschen Ungleichung bei der Schätzung von

$$P_k = \mathsf{P}\left(\max_{2^k < m \leqq 2^{k+1}} \sum_{j=1}^{m} W_j \geqq \varepsilon 2^{k\alpha} \right)$$

anwenden:

$$P_k \leqq \frac{\sum_{j=1}^{2^{k+1}} \mathsf{D}^2(W_j)}{\varepsilon^2 2^{2k\alpha}} \leqq K_{13} \frac{2^{(k+1)(1+\varepsilon_2)}}{\varepsilon^2 2^{2k\alpha}} \leqq K_{14} \frac{1}{\varepsilon^2} \frac{1}{2^{(6\delta-\varepsilon_2)k}}.$$

Ist $\varepsilon_2 < 6\delta$, so ist $\sum_k P_k$ konvergent, hieraus folgt (17) wegen des BOREL—CANTELLIschen Lemmas.

Wir haben bewiesen, daß

$$\lim_{k \to \infty} \frac{S_{n_k} - T_{n_k}}{n_k^{1/4+\delta}} = 0 \tag{19}$$

mit Wahrscheinlichkeit 1 gilt. Wir müssen noch die Abweichung zwischen S_m und S_{n_k} bzw. T_m und T_{n_k} untersuchen. Es ist bekannt (s. z. B. [3] S. 337), daß die Relation

$$\mathsf{P}\Big(\max_{n_k < m \leqq n_{k+1}} S_m - S_{n_k} \geqq x \Big) =$$

$$= \mathsf{P}\left(\max_{n_k < m \leqq n_{k+1}} \sum_{l=n_k+1}^{m} \xi_l \geqq x \right) \leqq \frac{4}{3} \mathsf{P}\big(U_k \geqq x - 2\sqrt{\Delta_k}\big) = c_k$$

besteht. Wählen wir $x = k^{\frac{1}{2}+3\delta}$, so ist

$$c_k \leqq \frac{4}{3} \mathsf{P}\left(\frac{U_k}{\sqrt{\Delta_k}} \geqq K_{15} k^{\delta} - 2 \right)$$

und wegen des Satzes über die großen Abweichungen folgt die Konvergenz der Reihe $\sum_1^{\infty} c_k$. Mit der Hilfe des BOREL—CANTELLIschen Satzes ergibt sich, daß $\max_{n_k < m \leqq n_{k+1}} S_m - S_{n_k} \geqq k^{\frac{1}{2}+3\delta}$ mit Wahrscheinlichkeit 1 nur für endlich viele k stattfindet. Man kann die untere Schätzung ähnlich durchführen, so ist

$$\max_{n_k < m \leqq n_{k+1}} |S_m - S_{n_k}| \leqq k^{1/2+3\delta}$$

mit Ausnahme der endlich vielen k. Hieraus folgt

$$\frac{|S_m - S_{n_k}|}{m^{1/4+\delta}} \leqq \frac{|S_m - S_{n_k}|}{n_k^{1/4+\delta}} \leqq \frac{k^{1/2+3\delta}}{k^{1/2+3\delta+4\delta^2}} \to 0. \tag{20}$$

(19), (20) und eine ähnliche Relation statt S_m, S_{n_k} mit T_m, T_{n_k} geben den Beweis des Satzes 3.

LITERATURVERZEICHNIS

[1] BÁRTFAI, P.: Die Bestimmung der zu einem wiederkehrenden Prozess gehörenden Verteilungsfunktion aus den mit Fehler behafteten Daten, *Studia Sci. Math. Hungar.* **1** (1966) 161—168.

[2] BASS, J.: Sur la compatibilité des fonctions de répartition, C. R. Acad. Sci. Paris **240** (1965) 839—841.

[3] RÉNYI, A.: *Wahrscheinlichkeitsrechnung*, Verlag der Wissenschaften, Berlin 1962.

Mathematisches Institut der Ungarischen Akademie der Wissenschaften, Budapest

(Eingegangen: 28. Juni 1968.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 51—62

COMPLETE ORTHOGONAL SYSTEMS OF EIGENFUNCTIONS OF THREE TRIANGULAR MEMBRANES

by

E. MAKAI

1. There exists a vast literature about solutions of the two-dimensional wave equation

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\lambda u = 0 \tag{1. 1}$$

for different boundary conditions, yet complete orthogonal systems of eigenfunctions and the corresponding eigenvalues of the first and second boundary condition problem are explicitly known but for a very few domains. These domains are:

(*a*) the rectangle,

(*b*) the isosceles rectangular triangle,

(*c*) the circle or more generally, the ellipse,

(*c'*) annular domains bounded by two confocal ellipses or domains bounded by arcs of two confocal ellipses and by arcs of two hyperbolas confocal to the given ellipses, finally limiting cases thereof.

Of course the explicit representation in cases (*c*) and (*c'*) must be taken in a very vague sense, since the eigenvalues are given as zeros of certain transcendental functions and the Mathieu functions involved in the representation of the eigenfunctions of say, the ellipses are by no way transcendentals capable of an easy treatment.

In this paper we wish to treat two more domains where complete orthogonal systems of eigenfunctions and the corresponding eigenvalues can be explicitly given for the first and second eigenvalue problems, as well. These are the 30°, 60°, 90° triangle and the equilateral triangle. Their eigenfunctions share with those of cases (*a*) and (*b*) the property that they can be represented as finite sums of certain trigonometric functions.

Let us first consider the 30°, 60°, 90° triangle $T: y \geqq 0,\ x \leqq h,\ y \leqq x/\sqrt{3}$. We shall treat in connection with the wave equation (1. 1) and the triangle T four different types of boundary conditions, namely

$$u=0 \quad \text{on} \quad y=0, \quad x=h, \quad y=x/\sqrt{3}, \tag{1. 2}$$

$$\partial u/\partial \mathbf{n}=0 \quad \text{on} \quad y=0, \quad x=h, \quad y=x/\sqrt{3}, \tag{1. 3}$$

$$\partial u/\partial \mathbf{n}=0 \quad \text{on} \quad y=0, \quad u=0 \quad \text{on} \quad x=h \quad \text{and on} \quad y=x/\sqrt{3}, \tag{1. 4}$$

$$u=0 \quad \text{on} \quad y=0, \quad \partial u/\partial \mathbf{n}=0 \quad \text{on} \quad x=h \quad \text{and on} \quad y=x/\sqrt{3}, \tag{1. 5}$$

$\partial/\partial \mathbf{n}$ meaning the normal derivative. For the sake of a concise representation we

introduce the vectors

$$\mathbf{v}_j^+ = \{a_j, b_j\} \qquad (j=1, 2, 3) \tag{1. 6}$$

with the components

$$a_1 = (m+n)\frac{\pi}{h}, \quad a_2 = -m\frac{\pi}{h}, \quad a_3 = -n\frac{\pi}{h}, \tag{1. 7}$$

$$b_1 = \frac{(m-n)}{\sqrt{3}}\frac{\pi}{h}, \quad b_2 = \frac{m+2n}{\sqrt{3}}\frac{\pi}{h}, \quad b_3 = -\frac{n+2m}{\sqrt{3}}\frac{\pi}{h}, \tag{1. 8}$$

where m and n denote integers. With the help of these vectors we define the functions

$$\begin{cases} u_{mn}^- = \sum \sin a_j x \sin b_j y, & u_{mn}^+ = \sum \cos a_j x \cos b_j y \\ v_{mn}^- = \sum \sin a_j x \cos b_j y, & v_{mn}^+ = \sum \cos a_j x \sin b_j y \end{cases} \tag{1. 9}$$

the summations being extended over $j=1, 2, 3$ and state

THEOREM 1. *Consider the domain* $T: y \geqq 0,\ x \leqq h,\ y \leqq x/\sqrt{3}$, *the membrane equation* (1. 1) *and the boundary conditions* (1. 2), (1. 3), (1. 4) *and* (1. 5). *The four sets of functions* u_{mn}^- $(m>n>0)$, u_{mn}^+ $(m \geqq n \geqq 0)$, v_{mn}^- $(m \geqq n>0)$ *and* v_{mn}^+ $(m>n \geqq 0)$, *respectively, are complete orthogonal systems of eigenfunctions of these four boundary condition problems.*

The eigenvalues belonging to any of the four functions $u_{mn}^{\pm}$, $v_{mn}^{\pm}$ *are given by*

$$\lambda_{mn} = \frac{4\pi^2}{3h^2}(m^2+mn+n^2). \tag{1. 10}$$

A simple consequence of this is

THEOREM 2. *Consider the equilateral triangular domain* $\Delta: x \leqq h,\ |y| \leqq x/\sqrt{3}$ *and the membrane equation* (1. 1). *A complete orthogonal system of eigenfunctions of the first boundary condition problem is given by the sets* u_{mn}^- $(m>n>0)$ *and* v_{mn}^- $(m \geqq n>0)$. *Similarly a complete orthogonal system of eigenfunctions of the second boundary problem is given by* u_{mn}^+ $(m \geqq n \geqq 0)$ *and* v_{mn}^+ $(m>n \geqq 0)$.

Again, the eigenvalues belonging to any of these eigenfunctions with indices m and n are given by the formula (1. 10).

We note that the smallest eigenvalue λ_{21} and the corresponding eigenfunction of the first boundary condition problem for T were found — in a different form — by G. PÓLYA [7]. Earlier, B. R. SETH [9] had given an infinite, but incomplete system of eigenfunctions and eigenvalues of the same problem. His list of eigenvalues contains λ_{21} without having really shown that it is the smallest positive eigenvalue, but fails to contain e.g. λ_{42}.

More was achieved in connection with the second boundary condition problem of the triangle T. Indeed it was shown by S. K. LAKSHMANA RAO [3, 4] — and this was the starting point of the present paper — that the functions u_{mn}^+ for any integer m and n are solutions of this problem. This set of functions is, however, not linearly independent and therefore not orthogonal: by virtue of the relations $u_{mn}^+ = u_{nm}^+ =$

$= u^+_{-m,-n} = u^+_{m+n,-n}$ any function u^+_{mn} (m, n integers) can be expressed by a function $u^+_{m_0 n_0}$ (m_0, n_0 integers, $m_0 \geqq n_0 \geqq 0$). Moreover the problem of the completeness of the system of functions u^+_{mn} was not treated by LAKSHMANA RAO.

As to the first boundary condition problem of the equilateral triangle, the system v^-_{mn} (m, n any integers) of solutions and the corresponding set of eigenvalues was found already by G. LAMÉ [5, p. 131—137]. This system contains the first eigenfunction v^-_{11}. Some of the eigenfunctions u^-_{mn} of the same problem (those for which $m \equiv n \mod 3$) were given by SETH [9].

Finally there does not seem to be any indication in the literature that the second boundary condition problem of the equilateral triangle has been ever treated.

An exception is paper [3], whose author remarks that the functions u^+_{mn} are eigenfunctions of this problem, too.

We remark that the asymptotic law (8. 1) for the distribution of eigenvalues and the conjecture (10. 1) of PÓLYA — both proved already for some other particular cases — turn out to be valid also for the domains T and Δ.

2. Let us start with a formal property of the vectors $\mathbf{v}^+_j$ defined by (1. 6) and the associated vectors $\mathbf{v}^-_j = \{a_j, -b_j\}$. It is easy to verify that one has

$$\mathbf{v}^\pm_j \mathbf{v}^\pm_j = \lambda_{mn} \qquad (j = 1, 2, 3) \tag{2. 1}$$

and

$$\sum_{j=1}^{3} \mathbf{v}^\pm_j = 0. \tag{2. 2}$$

This means that the vectors $\mathbf{v}^+_1$, $\mathbf{v}^+_2$, $\mathbf{v}^+_3$ [$\mathbf{v}^-_1$, $\mathbf{v}^-_2$, $\mathbf{v}^-_3$] can be arranged so as to form an equilateral triangle, or, alternatively, two of these vectors include an angle of 120°. Taking the origin as the starting point of these vectors, it follows from the definition that the vectors $\mathbf{v}^+_1$, $\mathbf{v}^+_2$, $\mathbf{v}^+_3$ [$\mathbf{v}^-_1$, $\mathbf{v}^-_2$, $\mathbf{v}^-_3$] form an anti-clockwise [clockwise] system.

Hence we can infer an important symmetry property of the functions $u^\pm_{mn}$, $v^\pm_{mn}$, namely that any of these functions exhibit a three-fold rotational symmetry about the origin. More explicitly let $P(x, y)$ and $P'(x', y')$ be two points of the xy plane, $\mathbf{r} = \overrightarrow{OP}$, $\mathbf{r}' = \overrightarrow{OP'}$ and the vector $\mathbf{r}'$ should be the result of rotating $\mathbf{r}$ about the origin by $+120°$. Then we state that the values of any of the functions $u^\pm_{mn}$, $v^\pm_{mn}$ at the places P and P' are the same.

Indeed

$$2u^\pm_{mn}(P) = \sum [\pm \cos (a_j x + b_j y) + \cos (a_j x - b_j y)] = \tag{2. 3}$$

$$= \sum [\pm \cos \mathbf{v}^+_j \mathbf{r} + \cos \mathbf{v}^-_j \mathbf{r}]$$

and similarly

$$2v^\pm_{mn}(P) = \sum [\sin \mathbf{v}^+_j \mathbf{r} \mp \sin \mathbf{v}^-_j \mathbf{r}], \tag{2. 4}$$

$$2u^\pm_{mn}(P') = \sum [\pm \cos \mathbf{v}^+_j \mathbf{r}' + \cos \mathbf{v}^-_j \mathbf{r}'], \tag{2. 5}$$

$$2v^\pm_{mn}(P') = \sum [\sin \mathbf{v}^+_j \mathbf{r}' \mp \sin \mathbf{v}^-_j \mathbf{r}']. \tag{2. 6}$$

Now by the definition of the scalar product $\mathbf{v}^+_j \mathbf{r} = \mathbf{v}^+_{j+1} \mathbf{r}'$ ($\mathbf{v}^+_4 = \mathbf{v}^+_1$) and $\mathbf{v}^-_j \mathbf{r} = \mathbf{v}^-_{j-1} \mathbf{r}'$ ($\mathbf{v}^-_0 = \mathbf{v}^-_3$). Hence $\sum \cos \mathbf{v}^\pm_j \mathbf{r} = \sum \cos \mathbf{v}^\pm_{j\pm 1} \mathbf{r}' = \sum \cos \mathbf{v}^\pm_j \mathbf{r}'$. Similarly $\sum \sin \mathbf{v}^\pm_j \mathbf{r} = \sum \sin \mathbf{v}^\pm_j \mathbf{r}'$ and finally

$$u^\pm_{mn}(P) = u^\pm_{mn}(P'), \quad v^\pm_{mn}(P) = v^\pm_{mn}(P'). \tag{2. 7}$$

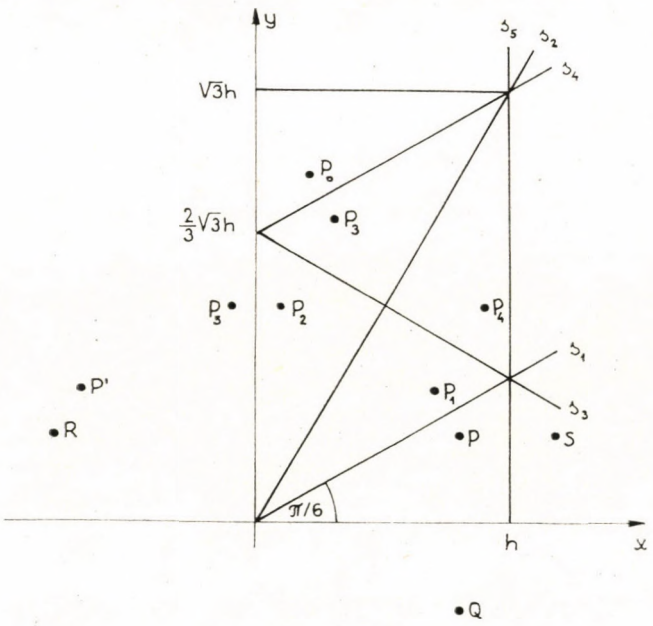
3. There are three other basic symmetry properties of the functions $u_{mn}^{\pm}$, $v_{mn}^{\pm}$ easy to verify by the aid of the definition of these functions. Let the Cartesian coordinates of the points P, Q, R, S be (x, y), $(x, -y)$, $(-x, y)$, $(2h-x, y)$, respectively. With these notations, these three symmetry properties are

$$u_{mn}^{\pm}(P) = \pm u_{mn}^{\pm}(Q), \quad v_{mn}^{\pm}(P) = \mp v_{mn}^{\pm}(Q), \tag{3. 1}$$

$$u_{mn}^{\pm}(P) = \pm u_{mn}^{\pm}(R), \quad v_{mn}^{\pm}(P) = \pm v_{mn}^{\pm}(R), \tag{3. 2}$$

$$u_{mn}^{\pm}(P) = \pm u_{mn}^{\pm}(S), \quad v_{mn}^{\pm}(P) = \pm v_{mn}^{\pm}(S). \tag{3. 3}$$

These properties, together with (2. 7) suffice to deduce all symmetry properties of our functions.



Indeed let us consider the rectangle $0 \leqq x \leqq h$, $0 \leqq y \leqq \sqrt{3}h$ which can be dissected into six congruent triangles, one of which is our triangle T.

Note that any two adjacent triangles are reflections of each other in their common sides. The points P_1, P_2, P_3, P_4, P_0 are the reflections of a point P when reflected successively in the oblique straight lines of the figure. (If $\mathcal{M}_s A = B$ symbolizes the fact that the point or point set B is the reflection of A in the straight line s, then $\mathcal{M}_{s_1} P = P_1$, $\mathcal{M}_{s_2} P_1 = P_2$, $\mathcal{M}_{s_3} P_1 = P_4$, $\mathcal{M}_{s_3} P_2 = P_3$, $\mathcal{M}_{s_4} P_3 = P_0$, $\mathcal{M}_x P = Q$, $\mathcal{M}_y P = R$, $\mathcal{M}_y P_2 = P_5$, $\mathcal{M}_{s_5} P = S$. The point P does not lie necessarily in the triangle T.)

Let us now consider the points P, Q, P_1, P_2, P'. Their radius vector is the same and we shall denote their polar angles by φ, $-\varphi$, $\pi/3-\varphi$, $\pi/3+\varphi$, $2\pi/3+\varphi$,

respectively. Then we have by (2. 7) and (3. 1)

$$u_{mn}^{\pm}(P) = u_{mn}^{\pm}(P') = \pm u_{mn}^{\pm}(P_1), \quad v_{mn}^{\pm}(P) = v_{mn}^{\pm}(P') = \pm v_{mn}^{\pm}(P_1), \tag{3. 4}$$

since P' and P_1 are reflections of each other in the y axis.

From now we shall deal only with the symmetry properties of the functions $u_{mn}^{\pm}$. We remark that since P_1 and P are reflections of each other in the straight line s_1 and so is the point pair Q and P_2, therefore we have by (3. 1) and (3. 4)

$$u_{mn}^{\pm}(P) = \pm u_{mn}^{\pm}(Q) = u_{mn}^{\pm}(P_2). \tag{3. 5}$$

From (3. 4) and (3. 5) we have $u_{mn}^{\pm}(P_1) = \pm u_{mn}^{\pm}(P_2)$ meaning that the straight line s_2 is an axis of symmetry [antisymmetry] of the functions u_{mn}^{+} $[u_{mn}^{-}]$. Hence

$$u_{mn}^{\pm}(P_3) = \pm u_{mn}^{\pm}(P_4). \tag{3. 6}$$

Again P_4 is the reflection of S in s_1, so we have by (3. 5) and (3. 3)

$$u_{mn}^{\pm}(P_4) = \pm u_{mn}^{\pm}(S) = u_{mn}^{\pm}(P). \tag{3. 7}$$

Combined with (3. 5) this leads to $u_{mn}^{\pm}(P_4) = \pm u_{mn}^{\pm}(P_1)$, meaning that s_3 is an axis of symmetry [antisymmetry] of the functions u_{mn}^{+} $[u_{mn}^{-}]$. Let now P_5 be the reflection of P_2 in the y axis. Then we have by (3. 2) $u_{mn}^{\pm}(P_2) = \pm u_{mn}^{\pm}(P_5)$. On the other hand P_0 and P_5 are reflections of each other in s_3. So we have finally by using (3. 5)

$$u_{mn}^{\pm}(P_0) = \pm u_{mn}^{\pm}(P_2) = u_{mn}^{\pm}(P). \tag{3. 8}$$

We collect our results in the formula

$$u_{mn}^{\pm}(P_j) = (-1)^j u_{mn}^{\pm}(P) \qquad (j = 0, 1, 2, 3, 4) \tag{3. 9}$$

and remark that in a similar way we can deduce

$$v_{mn}^{\pm}(P_j) = k_j v_{mn}^{\pm}(P); \ |k_j| = 1 \qquad (j = 0, \ldots, 4). \tag{3. 10}$$

4. With the help of (2. 1) one sees immediately that the functions $u_{mn}^{\pm}$ $[v_{mn}^{\pm}]$ satisfy the wave equation (1. 1) with the eigenvalue (1. 10). Indeed, by introducing the functions $\chi(z)$ and $\psi(z)$, any of which should mean either $\cos z$ or $\sin z$, our functions $u_{mn}^{\pm}$ and $v_{mn}^{\pm}$ can be represented in the uniform way $\sum \chi(a_j x)\psi(b_j y)$ and we have using (1. 6) and (2. 1)

$$\Delta \sum \chi(a_j x)\psi(b_j y) = \sum (-a_j^2 - b_j^2)\chi(a_j x)\psi(b_j y) = -\lambda_{mn} \sum \chi(a_j x)\psi(b_j y).$$

Remarking that if a straight line s is an axis of antisymmetry [symmetry] of an analytic function, then this function [the normal derivative of this function] vanishes on s, we are able to see from (3. 1), (3. 3) and (3. 4) that $u_{mn}^{\pm}$, $v_{mn}^{\pm}$ fulfil the prescribed boundary conditions.

It is somewhat more laborious to show the orthogonality and linear independence of these functions. We have to show that under the restrictions of Theorem 1 the necessary and sufficient conditions for the non-vanishing of the integral

$$I_{mn,m'n'} = \iint_T \sum_j \chi(a_j x)\psi(b_j y) \sum_k \chi(a'_k x)\psi(b'_k y)\, dx\, dy \tag{4. 1}$$

are

$$m = n, \quad m' = n'. \tag{4. 2}$$

In (4. 1) a_k' and b_k' depend on the same way on the quantities m' and n' as a_k and b_k do, respectively, on m and n; the integers m' and n' are subject to the restrictions of Theorem 1. It is not trivial that (4. 2) is sufficient for $I_{mn,m'n'} \neq 0$ since we do not still know that none of the functions mentioned in Theorem 1 is of 0 norm. So we show first the necessity of conditions (4. 2).

We remark that by (3. 9) and (3. 10) we have

$$u_{mn}^{\pm}(P_j)u_{m'n'}^{\pm}(P_j) = u_{mn}^{\pm}(P)u_{m'n'}^{\pm}(P), \quad v_{mn}^{\pm}(P_j)v_{m'n'}^{\pm}(P_j) = v_{mn}^{\pm}(P)v_{m'n'}^{\pm}(P)$$

for $j=0, 1, 2, 3, 4$, so the value of the right hand side of (4. 1) remains unchanged if we integrate over any triangular domain of the figure instead over T. Hence if we integrate over $0<x<h$, $0<y<\sqrt{3}h$ we get $6I_{mn,m'n'}$ or by introducing the new variables $\xi=\pi x/h$, $\eta=\pi y/(h\sqrt{3})$ we have

$$6I_{mn,m'n'} = \frac{h}{\pi}\cdot\frac{h\sqrt{3}}{\pi}\sum_{j,k=1}^{3} P_{jk}$$

with

$$P_{jk} = \int_0^{\pi}\int_0^{\pi} \chi\left(a_j\frac{h}{\pi}\xi\right)\chi\left(a_k'\frac{h}{\pi}\xi\right)\psi\left(b_j\frac{h\sqrt{3}}{\pi}\eta\right)\psi\left(b_k'\frac{h\sqrt{3}}{\pi}\eta\right)d\xi\,d\eta.$$

Now a necessary condition for the non-vanishing of $I_{mn,m'n'}$ is, that at least one of the nine quantities P_{jk} should not vanish. Again, necessary conditions for the non-vanishing of any of the quantities P_{jk} are just the conditions (4. 2). For seeing this one has to conclude in a somewhat different way, according to as $j=k$ or $j\neq k$. The reasonings are quite elementary and are based on the orthogonality properties of the functions $\sin \nu x$, $\cos \nu x$ ($\nu = 0, \pm 1, \pm 2, \ldots$) in the interval $(0, \pi)$ further on the fact that $a_j h/\pi$ and $b_j h\sqrt{3}/\pi$ are integers. We discuss only two cases on the lines of which each of the remaining cases can be treated.

Case (i): $j=k=1$. If P_{11} does not vanish, then necessarily

$$\left|a_1\frac{h}{\pi}\right| = \left|a_1'\frac{h}{\pi}\right|, \quad \left|b_1\frac{h\sqrt{3}}{\pi}\right| = \left|b_1'\frac{h\sqrt{3}}{\pi}\right|$$

since the quantities between the signs $|\ \ |$ are integers. Hence by (1. 7), (1. 8) and by $m\geqq n\geqq 0$, $m'\geqq n'\geqq 0$

$$m+n = m'+n', \quad m-n = m'-n',$$

equivalent to (4. 2).

Case (ii): $j=2$, $k=3$. Necessary conditions for the non-vanishing of P_{23} are

$$\left|a_2\frac{h}{\pi}\right| = \left|a_3'\frac{h}{\pi}\right|, \quad \left|b_2\frac{h\sqrt{3}}{\pi}\right| = \left|b_3'\frac{h\sqrt{3}}{\pi}\right|,$$

hence

$$m=n'; \qquad m+2n = n'+2m'.$$

Subtracting these we have $m'=n$. Now we have

$$m=n'\leqq m'=n\leqq m$$

meaning that all of these four quantities are equal, hence (4. 2) is again fulfilled.*

Now we proceed to proving that (4. 2) is sufficient for $I_{mn,m'n'}>0$. This could be done directly by calculating the norms of $u_{mn}^{\pm}$, $v_{mn}^{\pm}$. To avoid this tedious work we shall multiply these functions by suitable trial functions, calculate the integrals of these products on the rectangle $0\leqq x\leqq h$, $0\leqq y\leqq\sqrt{3}h$ and from the non-vanishing of these integrals infer that our functions do not vanish identically on the rectangle, hence — by their symmetry properties (3. 9) and (3. 10) — also on T.

First we shall consider the case $m\geqq n>0$ from which by (1. 7) and (1. 8)

$$|a_1|>|a_2|\geqq|a_3|>0, \quad 0\leqq|b_1|<|b_2|\leqq|b_3|,$$

write the functions $u_{mn}^{\pm}$, $v_{mn}^{\pm}$ in the unified form $\sum\chi(a_jx)\psi(b_jy)$ and take the trial function $\chi(a_1x)\psi(b_1y)$. Then we have by the orthogonality property of the sine and cosine functions

$$\int_0^{\sqrt{3}h}\int_0^{h}\chi(a_1x)\psi(b_1y)\sum\chi(a_jx)\psi(b_jy)\,dx\,dy=\int_0^{\sqrt{3}h}\int_0^{h}\{\chi(a_1x)\psi(b_1y)\}^2\,dx\,dy.$$

If $b_1\neq0$, i.e. $m\neq n$, the last integral differs from 0. Hence we infer that $u_{mn}^{\pm}$, $v_{mn}^{\pm}$ do not vanish identically, if $m>n>0$. If, however, $m=n$, and $\psi(z)=\cos z$, the integral still does not vanish and so $u_{mm}^{+}\not\equiv0$, $v_{mm}^{-}\not\equiv0$ if $m>0$.

Next consider the case $m>n\geqq0$ from which

$$|a_1|\geqq|a_2|>|a_3|\geqq0, \quad 0<|b_1|\leqq|b_2|<|b_3|$$

and use the trial function $\chi(a_3x)\psi(b_3y)$:

$$\int_0^{\sqrt{3}h}\int_0^{h}\chi(a_3x)\psi(b_3y)\sum\chi(a_jx)\psi(b_jy)\,dx\,dy=\int_0^{\sqrt{3}h}\int_0^{h}\{\chi(a_3x)\psi(b_3y)\}^2\,dx\,dy.$$

Let now be $\chi(z)=\cos z$. We infer by (1. 9) that

$$u_{mn}^{+}\not\equiv0, \quad v_{mn}^{+}\not\equiv0, \quad \text{if} \quad m>n\geqq0.$$

u_{00}^{+} is the only function in Theorem 1 which is not included into any of the above categories. Yet we have $u_{00}^{+}=3\neq0$.

5. The completeness of the systems u_{mn}^{+} $(m\geqq n\geqq0)$ and u_{mn}^{-} $(m>n>0)$ can be shown in the following manner: if $f(x,y)=f(P)$ is an L^2-integrable function defined on T, and if for each pair of integers m, n satisfying $m\geqq n\geqq0$ one has

$$\text{(5. 1)} \qquad \iint_T fu_{mn}^{+}\,dx\,dy=0 \quad \text{or} \quad \iint_T fu_{mn}^{-}\,dx\,dy=0$$

only for $f=0$ a.e., then our systems are complete. (Note that u_{mm}^{-} and u_{m0}^{-} vanish identically.)

We shall extend the definition of $f(x,y)$ to the domain $0\leqq x\leqq h$, $0\leqq y<\sqrt{3}h$ shown in the figure in such a way that it should have the same symmetry properties

* It can be shown by an argument of symmetry that each of the cases $P_{jj}\neq0$ can be reduced to case (i) and each of the cases $P_{jk}\neq0$ $(j\neq k)$ is equivalent to case (ii), yet we prefer not to detail it.

as the functions $u_{mn}^{\pm}$. We define so two functions f^+ and f^- satisfying the equalities

$$f^{\pm}(P_j) = (\pm 1)^j f(P). \tag{5. 2}$$

It is known that complete orthogonal systems of solutions of the membrane problem for the rectangle $0 \leqq x \leqq h$, $0 \leqq y \leqq \sqrt{3}h$ are given in the case of the first boundary condition problem by the functions

$$U_{\mu\nu}^{-} = \sin \mu \frac{\pi}{h} x \sin \nu \frac{\pi}{\sqrt{3}h} y \qquad (\mu, \nu = 1, 2, \ldots)$$

and in the case of the second boundary condition problem by the functions

$$U_{\mu\nu}^{+} = \cos \mu \frac{\pi}{h} x \cos \nu \frac{\pi}{\sqrt{3}h} y \qquad (\mu, \nu = 0, 1, \ldots).$$

So any L^2 integrable function F on this rectangle can be expanded in a series of the functions $U_{\mu\nu}^{-}$ or $U_{\mu\nu}^{+}$ converging in the mean to F. We shall choose $F = f^{\pm}$ and calculate the integrals

$$C_{\mu\nu} = \int_0^{\sqrt{3}h} \int_0^{h} f^{\pm} U_{\mu\nu}^{\pm} \, dx \, dy,$$

quantities proportional to the Fourier coefficients $c_{\mu\nu}^{\pm}$ of $f^{\pm}$ defined by the expansion $f^{\pm} \sim \sum c_{\mu\nu}^{\pm} U_{\mu\nu}^{\pm}$.

We remark that $f^{\pm}(P_0) = f^{\pm}(P)$, or $f^{\pm}(x, y) = f^{\pm}(h - x, \sqrt{3}h - y)$. On the other hand $U_{\mu\nu}^{\pm}(x, y) = (-1)^{\mu+\nu} U_{\mu\nu}^{\pm}(h - x, \sqrt{3}h - y)$, hence

$$C_{\mu\nu} = [1 + (-1)^{\mu+\nu}] \iint_{T \cup T_1 \cup T_2} f^{\pm} U_{\mu\nu}^{\pm} \, dx \, dy,$$

where $T_1 = \mathcal{M}_{s_1} T$, $T_2 = \mathcal{M}_{s_2} T_1$. So we have $C_{\mu\nu} = 0$ if $\mu + \nu$ is odd.

Consider now the case $\mu + \nu$ even. By virtue of (5. 2) we can express the last integral in the form

$$C_{\mu\nu} = 2 \iint_T f^{\pm}(x, y) \, w_{\mu\nu}^{\pm}(x, y) \, dx \, dy$$

where $w_{\mu\nu}^{+}$ has the following meaning. Let $\mathbf{t}_3^{\pm}$ be the vectors $\{\mu\pi/h, \pm\nu\pi/(\sqrt{3}h)\}$ and $\mathbf{r}$, $\mathbf{r}_1$, $\mathbf{r}_2$ the radius vectors of the points P, P_1, P_2, respectively. (These vectors have the same length $|\mathbf{r}|$ and include with the positive x axis the angles φ, $\pi/3 - \varphi$, $\pi/3 + \varphi$, respectively.) Then

$$U_{\mu\nu}^{\pm}(x, y) = U_{\mu\nu}^{\pm}(P) = \frac{1}{2} \{\mp \cos \mathbf{t}_3^{+} \mathbf{r} + \cos \mathbf{t}_3^{-} \mathbf{r}\}$$

and

$$w_{\mu\nu}^{\pm}(x, y) = U_{\mu\nu}^{\pm}(P) \pm U_{\mu\nu}^{\pm}(P_1) + U_{\mu\nu}^{\pm}(P_2) =$$

$$= \frac{1}{2} (\cos \mathbf{t}_3^{-} \mathbf{r} + \cos \mathbf{t}_3^{+} \mathbf{r}_1 + \cos \mathbf{t}_3^{-} \mathbf{r}_2) \pm \frac{1}{2} (\cos \mathbf{t}_3^{+} \mathbf{r} + \cos \mathbf{t}_3^{-} \mathbf{r}_1 + \cos \mathbf{t}_3^{+} \mathbf{r}_2).$$

Let now $\pm\tau$ be the angle between the vectors $\mathbf{t}_3^{\pm}$ and the positive x axis . We define now four new vectors $\mathbf{t}_1^+$, $\mathbf{t}_2^+$, $\mathbf{t}_1^-$, $\mathbf{t}_2^-$ all of which have the same length $|\mathbf{t}^{\pm}|$ and include with the positive x axis the angles

$$\tau+\frac{\pi}{3}, \quad \tau-\frac{\pi}{3}, \quad -\tau+\frac{\pi}{3}, \quad -\tau-\frac{\pi}{3},$$

respectively. By virtue of the relations

$$\mathbf{t}_3^+\mathbf{r}_1=\mathbf{t}_1^-\mathbf{r}, \quad \mathbf{t}_3^+\mathbf{r}_2=\mathbf{t}_2^+\mathbf{r}, \quad \mathbf{t}_3^-\mathbf{r}_1=\mathbf{t}_1^+\mathbf{r}, \quad \mathbf{t}_3^-\mathbf{r}_2=\mathbf{t}_2^-\mathbf{r},$$

directly verifiable by the definition of the scalar product we have

$$2w_{\mu\nu}^{\pm}=\sum_{j=1}^{3}[\pm\cos\mathbf{t}_j^+\mathbf{r}+\cos\mathbf{t}_j^-\mathbf{r}],$$

a formula very similar to the representations (2. 3) and (2. 4) of $u_{mn}^{\pm}$ at the end of Section 2. A direct calculation leads to

$$w_{n,n+2m}^{\pm}=\pm w_{m,m+2n}^{\pm}=w_{m+n,m-n}^{\pm}=u_{mn}^{\pm}.$$

So by our assumption (5. 1) we have $C_{\mu\nu}=0$ if $\mu+\nu$ is even and the pair μ, ν is expressible in either of the three forms

$$\mu=n, \quad \nu=n+2m; \qquad \mu=m, \quad \nu=m+2n; \qquad \mu=m+n, \quad \nu=m-n$$

where m and n are non-negative integers satisfying $m\geqq n$. But each non-negative pair of integers μ, ν ($\mu+\nu\equiv 0 \mod 2$) is expressible in at least one of these manners. If $\mu\geqq\nu$, the third representation is valid, if $\mu\leqq\nu$ we always can find two non-negative integers a and b such that $\mu=a$, $\nu=a+2b$ and take $m=\max(a, b)$, $n=\min(a, b)$.

So we have $c_{\mu\nu}^{\pm}=0$ for each pair of non-negative integers μ, ν, whence by the completeness of the systems $U_{\mu\nu}^{\pm}$ we infer $f^{\pm}=0$ a.e. and finally $f=0$ a.e.

6. One can prove the statement of Theorem 1 about the systems $v_{\mu\nu}^{\pm}$ in exactly the same way the orthogonal and complete sets of functions

$$V_{\mu\nu}^{+}=\cos\frac{\mu\pi}{h}x\sin\frac{\nu\pi}{\sqrt{3}h}y \qquad (\mu=0, 1, \ldots, \ \nu=1, 2, \ldots)$$

and

$$V_{\mu\nu}^{-}=\sin\frac{\mu\pi}{h}x\cos\frac{\nu\pi}{\sqrt{3}h}y \qquad (\mu=1, 2, \ldots, \ \nu=0, 1, \ldots)$$

being used instead of $U_{\mu\nu}^{\pm}$.

7. Turning now to the proof of Theorem 2 the orthogonality of the elements of the systems $\{u_{mn}^{\pm}, v_{mn}^{\pm}\}$ on Δ is a simple consequence of their orthogonality properties on T further of their symmetry properties (3. 1).

The completeness on Δ may be shown in the following manner. Any function $F(x, y)$ L^2-integrable on Δ can be written in the form $F(x, y)=F^+(x, y)+F^-(x, y)$ where $F^{\pm}(x, y)=\pm F^{\pm}(x, -y)$ and $F^{\pm}(x, y)$ is L^2-integrable on Δ.

Suppose now that

$$\iint_{\Delta} F(x, y) u_{mn}^{\pm}\, dx\, dy = 0 \quad \text{and} \quad \iint_{\Delta} F(x, y) v_{mn}^{\pm}\, dx\, dy = 0 \tag{7. 1}$$

for each integer m and n satisfying $0 \leqq n \leqq m$. Choosing the upper signs this implies on account of (3. 1) that

$$\iint_{T} F^{+} u_{mn}^{\pm}\, dx\, dy = \iint_{T} F^{-} v_{mn}^{\pm}\, dx\, dy \quad (0 \leqq n \leqq m)$$

whence by Theorem 1 $F^{+} = F^{-} = 0$ a.e. on T, hence $F = 0$ a.e. on Δ. We are led to the same conclusion by using the lower signs in (7. 1).

8. Let A be the area and L the length of the perimeter of a domain D. Let further $N_1(\lambda)$ and $N_2(\lambda)$ denote the number of eigenvalues not exceeding λ of the differential equation (1. 1) — each counted with proper multiplicity — if u satisfies the first or second boundary condition on the boundary of D, respectively. The conjecture

$$N_j(\lambda) = \frac{\mathsf{A}}{4\pi}\lambda + (-1)^j \frac{\mathsf{L}}{4\pi}\sqrt{\lambda} + o(\sqrt{\lambda}) \qquad (j = 1, 2) \tag{8. 1}$$

is attributed to H. WEYL by N. V. KUZNECOV [2], though it seems to have been first stated explicitly — in the case of the first boundary condition problem — by F. H. BROWNELL [1]. KUZNECOV proved (8. 1) recently in the case of those domains for which (1. 1) can be solved by the method of separating the variables [2].

Now it is easy to show that (8. 1) is true for the domains T and Δ, too, where the method of separation of variables fails. All what one has to use is: (*a*) the area $\lambda h^2\sqrt{3}/(24\pi)$ of the elliptic sector $0 \leqq (4/3)\pi^2 h^{-2}(x^2 + xy + y^2) \leqq \lambda$, $0 \leqq y \leqq x$ is equal to the number of lattice points of integer coordinates lying in the interior of the sector plus half the number of lattice points on its boundary plus a term $o(\sqrt{\lambda})$, (*b*) there are $o(\sqrt{\lambda})$ lattice points on the curvilinear part of the boundary of the same sector.*

9. Finally some words about the isosceles right-angled triangular membrane $y \geqq 0$, $x \leqq \pi$, $y \leqq x$. In this case complete sets of eigenfunctions of the first and second

* This follows in an elementary way from a theorem of LANDAU [6] according to which the number of the lattice points in the interior of the ellipse $x^2 + xy + y^2 = \frac{3}{4}h^2\pi^{-2}\lambda$ differs from the area of the ellipse by $O(\lambda^{1/3})$. Indeed let us draw the straight lines $y = x$, $y = -x$, $y = 0$, $y = -x/2$ and let the common points of these lines and the ellipse be A, B, C and D, respectively, all these points lying in the half plane $x > 0$. If O is the origin, then OA and OB are the half-axes of the ellipse. The moduli of the determinants of the mappings $\bar{x} = x + y$, $\bar{y} = -y$ and $\bar{\bar{x}} = x + y$, $\bar{\bar{y}} = -x$ are both 1. These linear transformations map the interior of the elliptic sector OAC on the elliptic sectors OCD and ODB, respectively, in an area preserving manner. Hence the common area of these sectors is 1/12th of that of the ellipse.

The same mappings show that there are exactly as many lattice points in the interior of each of the sectors OAC, OCD and ODB. Finally a counting of the lattice points on the elliptic disk leads together with the theorem of LANDAU to the desired result.

boundary conditions are

$$W_{mn}^{-} = \sin mx \sin ny - \sin nx \sin my \qquad (m > n > 0)$$

and

$$W_{mn}^{+} = \cos mx \cos ny + \cos nx \cos my \qquad (m \geqq n \geqq 0),$$

respectively. Again, by using the symmetry properties of the functions $W_{mn}^{\pm}$ one can prove the completeness of these systems with the help of the same property of the eigenfunctions of the square. The eigenvalues corresponding both to W_{mn}^{-} and W_{mn}^{+} are

$$\lambda_{mn} = m^2 + n^2.$$

Once again one can show that the conjecture (8. 1) is valid.

10. *Remarks.* (i) Let $N(\lambda)$ be the number of eigenvalues less than λ for a domain of area A, where we prescribe that the eigenfunctions satisfy the first boundary condition problem on a part of the boundary of length L_1, while on the remaining part of the boundary of length L_2 they satisfy the second boundary condition problem. Examples of this kind of mixed boundary condition problems are the problems (1. 1), (1. 4) and (1. 1), (1. 5). Asymptotic formulae for $N(\lambda)$ can be established in both of these cases and also for the isosceles right-angled triangle in the following two cases: the eigenfunctions satisfy the first boundary condition on one of the cathetes, the second boundary condition on the other cathete and the first, or second boundary condition on the hypotenuse. In all of these four instances the number $N(\lambda)$ can be expressed by the formula

$$N(\lambda) = \frac{\mathsf{A}}{4\pi}\lambda + \frac{\mathsf{L}_2 - \mathsf{L}_1}{4\pi}\sqrt{\lambda} + o(\sqrt{\lambda}).$$

(ii) Let the nondecreasing sequences $\lambda_1^{(j)}, \lambda_2^{(j)}, \ldots$ denote the eigenvalues — written with proper multiplicity — of the j'th boundary condition problem for a domain of area A. A conjecture of G. PÓLYA [8; p. 52] states that

$$\lambda_n^{(2)} < 4\pi n/\mathsf{A} < \lambda_n^{(1)}. \tag{10. 1}$$

This conjecture can be verified for the three triangular domains dealt with in this paper.

(iii) We point out the curious fact that in the case of the 30°, 60°, 90° and 45°, 45°, 90° triangles the eigenvalues of the first and second boundary condition problems can be represented by similar formulae: λ_{mn} is a quadratic form of m and n where m and n are in both cases subject to the same conditions.

REFERENCES

[1] BROWNELL, F. H.: An extension of Weyl's asymptotic law of eigenvalues, *Pac. J. Math.* **5** (1955), 483—499.

[2] KUZNECOV, N. V.: Asymptotic distribution of eigenfrequencies of a plane membrane in the case of separable variables, *Differencial'nye Uravnenija* **2** (1966), 1385—1402.

[3] LAKSHMANA RAO, S. K.: On the vibrations of triangular membranes, *Journ. of the Indian Institute of Science*, **38** No. 1 (1956), 1—3.

[4] LAKSHMANA RAO, S. K.: An exposition of the classical problem of vibrations of membranes and plates, *Journ. of the Institution of Telecommunication Engineers*, **4** (1957), 96—101.
[5] LAMÉ, G.: *Leçons sur la théorie mathématique de l'élasticité des corps solides*, Paris, Bachelier, 1852.
[6] LANDAU, E.: Zur analytischen Zahlentheorie der definiten quadratischen Formen, *Sitzungsber. d. Kgl. Preussischen Akad.* **31** (1915), 458—476.
[7] PÓLYA, G.: A note on the principal frequency of a triangular membrane, *Quarterly of Appl. Math.* **8** (1951), 386.
[8] PÓLYA, G.: *Mathematics and Plausible Reasoning* (vol. II. *Patterns of Plausible Inference*), Princeton Univ. Press, Princeton, 1954.
[9] SETH, B. R.: Transverse vibrations of triangular membranes, *Proc. Indian Acad. Sci. ser. A.* **12** (1940), 487—490.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received October 23, 1968.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 63—74

ÜBER POISSONGESETZE AUF LOKALKOMPAKTEN GRUPPEN UND VERWANDTE FRAGEN

von

W. HAZOD und L. SCHMETTERER

I. Einleitung

Es ist wohlbekannt, daß jedes unendlich oft teilbare Wahrscheinlichkeitsgesetz, welches über den Borelschen Mengen einer lokalkompakten Abelschen Gruppe mit zweitem Abzählbarkeitsaxiom definiert ist, in eine schwach stetige einparametrige Halbgruppe von Wahrscheinlichkeitsverteilungen eingebettet werden kann [1]. Ein verwandtes Resultat für beliebige kompakte Gruppen wurde kürzlich von CARNAL [2] bewiesen. Hier soll nun die Frage geklärt werden, wie man Poissongesetze auf beliebigen lokalkompakten Gruppen kennzeichnen kann und damit die Frage nach der Einbettung von Wahrscheinlichkeitsgesetzen in gleichmäßig stetige Halbgruppen berührt werden. Darüber hinaus verallgemeinern wir diese Resultate auf beliebige Banachalgebren (mit Einheitselement). Damit führen wir Ergebnisse von URBANIK [3] und CARNAL [2] für kompakte Gruppen weiter und verallgemeinern teilweise Resultate von BÖGE [4].

Wir geben zunächst eine DEFINITION. Es sei G eine beliebige lokalkompakte topologische Gruppe. Mit $\mathcal{M}$ bezeichnen wir die Menge aller beschränkten regulären Maße über den Borelschen Mengen $\mathfrak{B}$ von G. Es sei C die Menge aller (komplexwertigen) stetigen Funktionen mit kompakten Träger über G, deren Supremumsnorm mit $\|\cdot\|$ bezeichnet werde. Für jedes $\nu\in\mathcal{M}$ definieren wir mit $\sup\limits_{\substack{f\in C\\ \|f\|\leqq 1}} |\nu(f)|$ die Norm $\|\nu\|$. $\mathcal{M}$ ist bezüglich der Addition und der Faltung (die wir als Multiplikation schreiben), sowie bezüglich der eben eingeführten Norm eine Banachalgebra. Sei $\mathbf{Z}\subset\mathcal{M}$ die Menge aller Wahrscheinlichkeitsmaße über $\mathfrak{B}$, das sind die nichtnegativen Maße, deren Norm gleich 1 ist. Mit H bezeichnen wir stets eine beliebige kompakte Untergruppe von G. Es seien $e_H\in\mathbf{Z}$ jene Maße, deren Träger H ist und welche dort mit dem (normierten) Haarschen Maß übereinstimmen. Genau die Maße e_H sind die Idempotenten von $\mathbf{Z}$ [5].

Für jedes $\mu\in\mathcal{M}$ definieren wir ein Maß $\tilde{\mu}\in\mathcal{M}$ gemäß $\tilde{\mu}(f)=\overline{\mu(\tilde{f})}$ für $f\in C$, wobei $\tilde{f}$ die Abbildung $x\to\overline{f(x^{-1})}$ über G ist. Mit $\delta_x\in\mathbf{Z}$, $x\in G$ bezeichnen wir das Maß, dessen Gesamtmasse in der einpunktigen Menge $\{x\}$ konzentriert ist.

$\mu\in\mathbf{Z}$ heißt unendlich oft teilbar, falls zu jedem natürlichen $n\geqq 1$ n-te Wurzeln, d.h. Elemente $\mu_{\frac{1}{n}}\in\mathbf{Z}$ existieren, mit

$$(\mu_{\frac{1}{n}})^n=\mu.$$

Die Wahrscheinlichkeitsverteilung $\mu=\mu_1\in\mathbf{Z}$ heißt sukzessiv unendlich teilbar, falls ein Homomorphismus $r\to\mu_r$ von der additiven Halbgruppe $\mathcal{R}^+$ der positiven rationalen Zahlen r in die Faltungshalbgruppe $\mathbf{Z}$ existiert.

Es sei $\mu \in \mathbf{Z}$ unendlich oft teilbar. Wir sagen, daß die Gesamtheit der Wurzeln $\mu_{\frac{1}{n}}$ eine Verträglichkeitsbedingung (V) erfüllt, falls für jedes $n \geqq 1$ und $k \geqq 1$ die Relation

(V) $$\mu_{\frac{1}{n}} = (\mu_{\frac{1}{k \cdot n}})^k$$

besteht. Es ist leicht zu sehen, daß diese Bedingung äquivalent zur Annahme ist, daß μ_1 sukzessiv unendlich oft teilbar ist. Der Begriff der Wurzel überträgt sich sofort auf eine beliebige Banachalgebra. In jeder Banachalgebra B mit Einheit e und Norm $\|\cdot\|$ ist die Reihe $e + \sum_1^\infty \frac{v^k}{k!}$ für jedes $v \in B$ konvergent, und der Grenzwert definiert die Exponentialfunktion, deren Wert für $v \in B$ mit $\exp(v)$ bezeichnet werden soll.

Für vertauschbare Elemente v, w gilt:

$$\exp(v) \cdot \exp(w) = \exp(w) \cdot \exp(v) = \exp(v + w)$$

Die Reihe $-\sum_1^\infty \frac{(e-v)^k}{k}$ konvergiert für alle $v \in B$ mit $\|v - e\| < 1$ und der Grenzwert wird mit $\log(v)$ bezeichnet. Bekanntlich läßt sich der Logarithmus für die Elemente von B mittels des Cauchy'schen Integralsatzes als (mehrdeutige) Abbildung einführen. Der Hauptzweig dieser Abbildung stimmt, kurz gesagt, mit der oben gegebenen Definition des Logarithmus überein.

Die Teilmenge $e_H \mathcal{M} e_H$ von $\mathcal{M}$ ist eine Banachalgebra mit Einheitselement e_H. Die zugehörige Exponentialfunktion bezeichnen wir mit $\exp_H$ und den entsprechenden Logarithmus mit $\log_H$. Das Wahrscheinlichkeitsgesetz $\mu \in \mathbf{Z}$ heißt Poissongesetz, wenn folgende Bedingungen erfüllt sind:

(a) Es existiert ein $v \in e_H \mathcal{M} e_H$ mit $v = v_1 - \|v_1\| e_H$, wobei v_1 ein nichtnegatives Maß aus $e_H \mathcal{M} e_H$ ist.

(b) Es gilt $\mu = \exp_H(v)$

Die Gesamtheit aller Maße v, die der Bedingung (a) genügen, soll mit $\mathcal{R}_H$ bezeichnet werden.

II. Charakterisierung der Poissongesetze über lokalkompakten topologischen Gruppen

Wir beweisen folgenden

SATZ 2. 1. *Es sei $\mu_1 \in \mathbf{Z}$ und unendlich oft teilbar. Für $n \geqq 1$ bezeichnen wir die n-te Wurzel mit $\mu_{\frac{1}{n}}$. Die Wurzeln $\{\mu_{\frac{1}{n}}\}$ mögen die in* I. *erklärte Verträglichkeitsbedingung* (V) *erfüllen. Es soll ein natürliches n_0 existieren, sodaß $\mu_{\frac{1}{n}} \in e_H \mathbf{Z} e_H$ für $n \geqq n_0$ gilt. Es seien $k(n)$ und $\omega(n)$ Folgen positiver Zahlen und es sei*

(2. 1) $$\lim_{n \to \infty} k(n) = 0$$

und

(2. 2) $$\lim_{n \to \infty} \omega(n) \cdot k(n) = \infty.$$

Es möge eine natürliche Zahl existieren, welche gleich n_0 angenommen werden kann, sodaß

(2. 3) $$\|\mu_{\frac{1}{n}} - e_H\| \leqq 2(1-k(n))$$

für alle $n \geqq n_0$ gilt.

Überdies sei

(2. 4) $$\sup_n \omega(n) \cdot \|\lambda_{\frac{1}{n}} - e_H\| < \infty$$

wobei $\lambda_{\frac{1}{n}} = \mu_{\frac{1}{n}} * \tilde{\mu}_{\frac{1}{n}}$ ist.

Dann ist μ_1 ein Poissongesetz, genügt daher den Bedingungen (a) *und* (b) *von* I. *und läßt sich also in der Form $\mu = \exp_H(v)$ darstellen und es gilt für jedes $n \geqq 1$ auch* $\mu_{\frac{1}{n}} = \exp_H\left(\frac{1}{n} \cdot v\right)$

BEWEIS:

Es sei v ein beliebiges Wahrscheinlichkeitsmaß aus $e_H \mathbf{Z} e_H$. Dann gilt

(2. 5) $$\|v - e_H\| = 2v(G-H),$$

weiters ist leicht zu sehen, daß die Ungleichung

(2. 6) $$v * \tilde{v}(G-H) \geqq v(G-H)v(H)$$

besteht.

Für $n \geqq n_0$ gilt daher $\lambda_{\frac{1}{n}}(G-H) \geqq \mu_{\frac{1}{n}}(G-H)\mu_{\frac{1}{n}}(H)$ und daher auch

(2. 7) $$\omega(n)\lambda_{\frac{1}{n}}(G-H) \geqq \omega(n)\mu_{\frac{1}{n}}(G-H)\mu_{\frac{1}{n}}(H).$$

Wendet man (2. 5) sinngemäß an, dann folgt aus (2. 3)

$$2\mu_{\frac{1}{n}}(G-H) \leqq 2(1-k(n)),$$

also

(2. 8) $$\mu_{\frac{1}{n}}(H) \geqq k(n)$$

für $n \geqq n_0$. (2. 7) und (2. 8) ergeben die Ungleichung

$$\omega(n)\lambda_{\frac{1}{n}}(G-H) \geqq \omega(n)k(n)\mu_{\frac{1}{n}}(G-H)$$

und dies impliziert wegen (2. 2) und (2. 4) die Beziehung

(2. 9) $$\mu_{\frac{1}{n}}(G-H) \to 0.$$

Es folgt also wieder nach (2. 5)

(2. 10) $$\|\mu_{\frac{1}{n}} - e_H\| \to 0.$$

Man kann annehmen, daß für $n \geqq n_0$

$$\|\mu_{\frac{1}{n}} - e_H\| < 1 \tag{2.11}$$

gilt. Dann ist $b = \log_H(\mu_{\frac{1}{n_0}})$ definiert und es folgt

$$\mu_{\frac{1}{n_0}} = \exp_H(b).$$

Aus (2. 10) und (2. 11) und (V) folgt bekanntlich, daß $\mu_{\frac{1}{n_0}}$ sogar Poissongesetz ist und sinngemäß die Bedingungen (a) und (b) von I. erfüllt sind. Weiters gilt für alle hinreichend großen natürlichen Zahlen m auch

$$\mu_{\frac{1}{n_0 m}} = \exp\left(\frac{1}{m} \cdot b\right) \quad \text{(Vergleiche Böge [4]).}$$

Überdies folgt $\mu_1 = (\mu_{\frac{1}{n_0}})^{n_0} = \exp(n_0 \cdot b)$.

Mit $v = n_0 b$ erhält man die gewünschte Darstellung von μ_1, da mit b auch v die Bedingung (a) von I. erfüllt.

Ist n eine beliebige natürliche Zahl $\geqq n_0$, dann existiert nach (2. 11) $b^* = \log_H(\mu_{\frac{1}{n}})$, es ist $\mu_{\frac{1}{n}} = \exp_H(b^*)$ und es gilt wieder für alle hinreichend großen m

$$\mu_{\frac{1}{n \cdot m}} = \exp_H\left(\frac{1}{m} b^*\right).$$

Somit ist einerseits

$$\log_H(\mu_{\frac{1}{n \cdot m \cdot n_0}}) = \frac{1}{m \cdot n_0} b^*$$

und andererseits gleich $\frac{1}{n \cdot m} b$ für hinreichend großes m.

Es folgt

$$\mu_{\frac{1}{n}} = \exp_H\left(\frac{n_0}{n} b\right) = \exp_H\left(\frac{1}{n} v\right).$$

Somit folgt aus der Verträglichkeitsbedingung für alle $n \geqq 1$ die Darstellung $\mu_{\frac{1}{n}} = \exp_H\left(\frac{n_0}{n} b\right) = \exp_H\left(\frac{1}{n} v\right)$.

Der Beweis lehrt auch, daß v von der gemäß (2. 11) gewählten natürlichen Zahl n_0 nicht abhängt. Anderseits ist es nicht möglich im allgemeinen eine Aussage über die Eindeutigkeit der Darstellung von μ_1 als Poissongesetz zu machen (Vgl. Bemerkung 2. 4). Auf die Bedeutung der Verträglichkeitsbedingung (V) hoffen wir in Kürze zurückkommen zu können.

BEMERKUNG 2. 2. Die Bedingung (2. 3) ist stets erfüllt, falls man nur

$$\sup_n \|\mu_{\frac{1}{n}} - e_H\| < 2 \tag{2.12}$$

voraussetzt. Aus der Bedingung (2. 4) folgt insbesondere die Existenz einer natürlichen Zahl n_1, sodaß

(2. 13) $$\|\lambda_{\frac{1}{n_1}} - e_H\| < 1.$$

Es ist naheliegend, anzunehmen, daß der Satz 2. 1 auch gilt, wenn man neben (V) die Voraussetzung (2. 12) und (2. 13) macht.

Dies ist auch der Fall, doch werden wir gleich eine etwas allgemeinere Aussage beweisen. Zunächst wollen wir jedoch darauf hinweisen, daß auch im Fall einer kompakten Abel'schen Gruppe die Bedingung (2. 12) nicht mehr abgeschwächt werden kann, ohne die Aussage des Satzes 2. 1 in Frage zu stellen, auch wenn man (V) aufrecht erhält und (2. 13) durch stärkere Voraussetzungen ersetzt. Dazu betrachten wir folgendes

BEISPIEL 2. 3.: Es sei G der Torus. Wir wählen eine Folge α_n positiver reeller Zahlen, sodaß $\sum_{n=1}^{\infty} \alpha_n$ konvergiert.

Für $n \geqq 1$ sei $x_n = e^{\frac{2\pi i}{n}}$. Wir erklären ein beschränktes Maß ν gemäß

$$\nu = \sum_{n=1}^{\infty} \alpha_n \delta_{x_n} - \left(\sum_{n=1}^{\infty} \alpha_n\right) \cdot \delta_{x_1}.$$

Offenbar gilt $\nu \in \mathscr{R}_{\{x_1\}}$, wobei $\{x_1\}$ die Untergruppe von G bezeichnet, die nur aus dem Einheitselement besteht. Somit ist $\mu_t = \exp_{\{x_1\}}(t \cdot \nu)$ für jedes reelle $t > 0$ Poissonmaß.

Für reelle $t > 0$ sei $y_t = e^{-2\pi i t}$ und wir definieren $\pi_t = \delta_{y_t} * \mu_t$. Es ist leicht zu sehen, daß π_t, $t > 0$, kein Poissonprozeß ist, die Verträglichkeitsbedingung ist jedoch erfüllt. Weiters gilt

$$\|\pi_{\frac{1}{n}} - \delta_{x_1}\| = 2\left(\pi_{\frac{1}{n}}(G - \{x_1\})\right) = 2\left(1 - \pi_{\frac{1}{n}}(\{x_1\})\right) = 2\left(1 - \mu_{\frac{1}{n}}(\{x_n\})\right) < 2.$$

Überdies ist $\pi_t * \tilde{\pi}_t = \mu_t * \tilde{\mu}_t = \exp\left(t(\nu + \tilde{\nu})\right)$ für $t > 0$ und daher Poissonmaß. Die Bedingung (2. 13) kann also sicher erfüllt werden.

BEMERKUNG 2. 4. Nach [4] ist das Maß $\nu \in \mathscr{R}_H$ welches gemäß (a), (b) von I. einen Poissonprozeß definiert, bereits durch die Wurzeln $\mu_{\frac{1}{n}} = \exp\left(\frac{1}{n}\nu\right)$ eindeutig festgelegt. Man kann nun untersuchen, ob ν schon durch das Maß μ_1 bestimmt ist. Man erhält jedoch leicht das folgende Resultat:

μ, a, b (mit $a \neq b$) seien reelle Maße in $e_H \mathscr{M} e_H$, es gelte

(2. 14) $$\mu = \exp_H(a) = \exp_H(b).$$

Dann gibt es ein Wahrscheinlichkeitsmaß $\mu \in e_H \mathbf{Z} e_H$ und Maße $a_1, b_1 \in \mathscr{R}_H$ mit $a_1 \neq b_1$, sodaß

$$\mu_1 = \exp_H(a_1) = \exp_H(b_1) \quad \text{gilt.}$$

Es kann somit μ_1 in zwei verschiedene Poissonprozesse eingebettet werden. Es gibt einfache Beispiele von Maßen die (2. 14) erfüllen, s. [9].

III. Verallgemeinerung auf Banachalgebren mit Einheitselement

Wir wollen jetzt einen Satz beweisen, welcher das in der Bemerkung 2. 2 angegebene Resultat in einen allgemeinen Rahmen einordnet.

SATZ 3. 1. *Es sei B eine (nicht notwendig kommutative) Banachalgebra mit Einheit e und mit einer Involution, die wir mit dem Symbol $\sim$ bezeichnen. Es sei a_1 ein Element von B mit folgenden Eigenschaften:*

(i) *Zu jedem natürlichen $n \geqq 1$ existiert ein $a_{\frac{1}{n}} \in B$ mit $(a_{\frac{1}{n}})^n = a_1$ und es ist $\|a_{\frac{1}{n}}\| \leqq 1$ für alle $n \geqq 1$.*

(ii) *Die Elemente $a_{\frac{1}{n}}$ erfüllen eine Verträglichkeitsbedingung d.h. es gilt für alle ganzen $n \geqq 1$ und $k \geqq 1$ die Relation*

$$a_{\frac{1}{n}} = (a_{\frac{1}{k \cdot n}})^k.$$

(iii) *Es ist* $\sup_n \|a_{\frac{1}{n}} - e\| < 2.$

(iv) *Es gibt ein $n_0 \geqq 1$ mit $\|a_{\frac{1}{n_0}} \cdot \tilde{a}_{\frac{1}{n_0}} - e\| < 1$ und $\|\tilde{a}_{\frac{1}{n_0}} \cdot a_{\frac{1}{n_0}} - e\| < 1$.*

Dann existiert ein $b \in B$, sodaß $a_{\frac{1}{n}} = \exp\left(\frac{1}{n} b\right)$ für alle $n \geqq 1$ gilt.

Für den Beweis benötigen wir eine Reihe von teils bekannten und teils wohl neuen Überlegungen, die wir in mehreren Hilfssätzen wiedergeben:

HILFSSATZ 3. 2. *Es sei f ein über den komplexen Zahlen definiertes Polynom $f(\lambda) = \sum_{k=0}^{n} c_k \lambda^k$, dann definiert für jedes $a \in B$ der Ausdruck $f(a) = \sum_{k=0}^{n} c_k a^k$ ein Element von B und es gilt für das Spektrum* Sp (a) *die Relation:*

$$f(\mathrm{Sp}\,(a)) = \mathrm{Sp}\,(f(a)).$$

Dieselbe Relation gilt für analytische Funktionen f, wenn Sp (a) *im (offenen) Definitionsbereich von f enthalten ist.*

Das ist der bekannte Spektralabbildungssatz (vgl. RICKART [6]).

HILFSSATZ 3. 3 [7]. *Es seien y_1 und y_2 vertauschbare Elemente von B, es sei* $\exp(y_1) = \exp(y_2)$, *dann ist $y_1 - y_2 = (2\pi i) \sum_{j=1}^{n} k_j e_j$, $n \geqq 1$ wobei die k_j ganze Zahlen sind, von denen höchstens eine 0 sein darf, und die e_j Idempotente $\neq 0$ sind und den Bedingungen $e_j e_k = \delta_{jk} e_j$ und $\sum_{j=1}^{n} e_j = e$ genügen.*

Falls y_1 der Bedingung Sp $(y_1) \cap \{(2k\pi i) + \mathrm{Sp}\,(y_1)\} = \emptyset$ *für alle ganzzahligen $k \neq 0$ genügt, folgt die Vertauschbarkeit von y_1 und y_2.*

Für jede komplexe Zahl λ bezeichnen wir nun mit arg λ das Argument von λ, also jene (eindeutig bestimmte) reelle Zahl, welche $-\frac{1}{2} \leqq \arg \lambda < \frac{1}{2}$ und $\lambda = |\lambda| e^{2\pi i \arg \lambda}$ erfüllt.

HILFSSATZ 3. 4. *Sei $c_1 \in B$, es gebe für jedes natürliche $n \geqq 1$ eine n-te Wurzel $c_{\frac{1}{n}} \in B$ die folgenden Bedingungen genügt:*

(α) $\|c_{\frac{1}{n}}\| \leqq 1$ *und jedes $c_{\frac{1}{n}}$ ist invertierbar*

(β) *Für alle natürlichen $k \geqq 1$ gilt $(c_{\frac{1}{k \cdot n}})^k = c_{\frac{1}{n}}$*

(γ) *Es gibt ein natürliches n_0, derart, daß für $n \geqq n_0$ gilt:*

$$\{\arg \lambda,\ \lambda \in \mathrm{Sp}\,(c_{\frac{1}{n}})\} \subseteq (-\tfrac{1}{3}, \tfrac{1}{3})$$

Dann gibt es ein $b \in B$, sodaß $c_{\frac{1}{n}} = \exp\left(\frac{1}{n} b\right)$ für $n \geqq 1$.

BEWEIS: Aus dem Hilfssatz 3. 2 folgt die Existenz eines positiven ε und einer natürlichen Zahl n_0, sodaß für $n \geqq n_0$

$$\mathrm{Sp}\,(c_{\frac{1}{n}}) \subset \left\{\lambda : |\lambda| \geqq \varepsilon^{\frac{1}{n}},\ |\arg \lambda| < \frac{1}{3}\right\} \text{ gilt.}$$

Daher liegt $\mathrm{Sp}\,(c_{\frac{1}{n}})$ im Definitionsbereich des Hauptastes der Logarithmenfunktion und es existiert $b(n) = \log c_{\frac{1}{n}}$ für $n \geqq n_0$. Ist $\lambda \in \mathrm{Sp}\,(b(n) - 2b(2n))$, so ist

$$|\mathrm{Im}\,\lambda| \leqq 2\pi \cdot 3 \cdot \max_{\substack{\lambda \in \mathrm{Sp}(c_{\frac{1}{n}}) \\ n \geqq n_0}} |\arg \lambda| < 2\pi \cdot 3 \cdot \frac{1}{3} = 2\pi. \tag{3. 1}$$

Andererseits ist offensichtlich $\exp\,(b(n) - 2b(2n)) = e$, daher ist nach Hilfssatz 3. 3 also $b(n) - 2b(2n)$ von der Form $2\pi i \sum_j k_j e_j$ und man sieht leicht, daß $\mathrm{Sp}\,(b(n) - 2b(2n)) = \{2\pi i k_j\}$ ist. Aus (3. 1) folgt nun, daß jedes $k_j = 0$ sein muß.

Damit zeigt man sofort:

$$b(n) = 2b(2n) = \cdots = 2^k b(2^k \cdot n) = \cdots$$

und

$$\max_{\lambda \in \mathrm{Sp}(c_{1/2^k n})} |\arg \lambda| < \frac{1}{3 \cdot 2^k} \quad \text{für} \quad n \geqq n_0. \tag{3. 2}$$

Ist nun r eine beliebige natürliche Zahl, dann gibt es ein natürliches k, sodaß $3 \cdot 2^k > r + 1$ ist. Daraus und aus (3. 2) folgert man in Analogie zu (3. 1), daß $|\mathrm{Im}\,\lambda| < 2\pi$ für alle $\lambda \in \mathrm{Sp}\,(b(2^k \cdot n) - r \cdot b(2^k \cdot n \cdot r))$ ist, und daraus folgt wiederum $b(n) = r \cdot b(r \cdot n)$ für $n \geqq n_0$ und für alle $r \geqq 1$.

Aus der Verträglichkeitsbedingung (β) ergibt sich, daß $n \log c_{\frac{1}{n}}$ für alle $n \geqq n_0$ gleich ein und demselben Element b^* ist und daher für alle $n \geqq 1$ auch $c_{\frac{1}{n}} = \exp\left(\frac{1}{n} b^*\right)$ gilt.

Es sei darauf hingewiesen, daß im Beweis wesentlich die Verträglichkeitsbedingung (β) verwendet wird. Läßt man (β) fallen, dann ist die Aussage des Hilfssatzes 3. 4 im allgemeinen falsch, wie das Beispiel 3. 9 zeigt. Jedoch kann man die Bedingung (β) entbehren, wenn man (γ) verschärft. Es gilt nämlich der

HILFSSATZ 3. 5. *Sei* $\{n_i\}$ *eine endliche oder unendliche Folge natürlicher Zahlen, es seien* $c_{\frac{1}{n_i}} \in B$ n_i*-te Wurzeln von* c_1. *Es werde vorausgesetzt:*

(α) $\|c_{\frac{1}{n_i}}\| \leqq 1$ *und jedes* $c_{\frac{1}{n_i}}$ *ist invertierbar.*

(γ') $$\{\arg \lambda,\ \lambda \in \mathrm{Sp}\,(c_{\frac{1}{n_i}})\} \subset \left(-\frac{1}{2n_i}, \frac{1}{2n_i}\right).$$

Dann gibt es ein $b \in B$ *mit* $c_{\frac{1}{n_i}} = \exp\left(\frac{1}{n_i} b\right)$ *für alle* n_i.

BEWEIS: Die Existenz der Logarithmen $\log c_{\frac{1}{n_i}}$ folgt wie im Hilfssatz 3. 4. Man überlegt sich weiter, daß aus (γ') und aus Hilfssatz 3. 3 die Vertauschbarkeit von $b(n_i) = \log c_{\frac{1}{n_i}}$ und $b(n_j) = \log c_{\frac{1}{n_j}}$ für jedes Paar (n_i, n_j) folgt. Daher gilt wiederum $\exp\left(n_i b(n_i) - n_j b(n_j)\right) = e$. Der Rest des Beweises wird wie im Hilfssatz 3. 4 geführt.

BEMERKUNG 3. 6. Die Bedingungen des Hilfssatzes 3. 4 sind insbesondere erfüllt, wenn $\|c_{\frac{1}{n}} - e\| < 1$ für alle $n \geqq 1$ vorausgesetzt wird. Dies entspricht einem Satz über Wahrscheinlichkeitsmaße in [4].

Für den Beweis des Satzes 3. 1 benötigen wir noch den

HILFSSATZ 3. 7 [8]: *Sei* r *eine natürliche Zahl, dann gibt es zu jedem reellen* α *einen Bruch* $\frac{p}{q}$ *mit* $1 \leqq q \leqq r$ *und* $(p, q) = 1$, *sodaß* $\left|\alpha - \frac{p}{q}\right| < \frac{1}{r \cdot q}$ ist.

BEWEIS des Satzes 3. 1

Aus der Bedingung (iv) folgt die Regularität von $a_{\frac{1}{n_0}} \cdot \tilde{a}_{\frac{1}{n_0}}$ und man überlegt sich leicht, daß dann auch $a_{\frac{1}{n_0}}$ und

(iv a) alle $a_{\frac{1}{n}}$ regulär sind.

Dann gibt es ein ε mit $0 < \varepsilon \leqq 1$, sodaß $\{|\lambda| < \varepsilon\} \cap \mathrm{Sp}\,(a_1) = \emptyset$ ist. Aus Hilfssatz 3. 2 folgt

$$\mathrm{Sp}\,(a_{\frac{1}{n}}) \cap \{|\lambda| < \varepsilon^{\frac{1}{n}}\} = \emptyset.$$

Für jedes $c \in B$ bezeichne $\varrho(c)$ den Spektralradius. Dann gilt bekanntlich $\|c\| \geqq \varrho(c)$. Somit folgt aus der Bedingung (iii) die Existenz eines reellen α mit $1 \leqq \alpha < 2$, sodaß

(iii a) $\varrho(a_{\frac{1}{n}} - e) < \alpha < 2$ für alle n gilt.

Da $\varepsilon^{\frac{1}{n}}$ gegen 1 konvergiert, gibt es eine natürliche Zahl n_1, derart, daß für $n \geqq n_1$ gilt: $\varepsilon^{\frac{1}{n}} > \alpha - 1$.

Nun behaupten wir, daß mit

$$\varphi_n = \frac{1}{2\pi} \cdot \arccos\left((\alpha-1)/\varepsilon^{\frac{1}{n}}\right) \tag{3. 1}$$

für alle $\lambda \in \mathrm{Sp}\left(a_{\frac{1}{n}}\right)$

$$|\arg \lambda| < \frac{1}{2} - \varphi_n$$

gilt. Es wurde ja bereits gezeigt, daß für alle $\lambda \in \mathrm{Sp}\left(a_{\frac{1}{n}}\right)$ die Relation $1 \geqq |\lambda| \geqq \varepsilon^{\frac{1}{n}}$ besteht. Würde aber $|\arg \lambda| \geqq \frac{1}{2} - \varphi_n$ für ein $\lambda \in \mathrm{Sp}\left(a_{\frac{1}{n}}\right)$ gelten, so würde für den Realteil folgen: $\mathrm{Re}\ \lambda = |\lambda| \cos(2\pi \arg \lambda)$. Aus $\alpha \geqq 1$ folgt $\varphi_n \leqq \frac{1}{4}$ und daher $\cos(2\pi \arg \lambda) \leqq 0$, somit gilt weiter

$$|\lambda| \cos(2\pi \arg \lambda) \leqq |\lambda| \cos\left(2\pi\left(\frac{1}{2} - \varphi_n\right)\right) = -|\lambda| \cos(2\pi\varphi_n) \leqq -\varepsilon^{\frac{1}{n}} \cos(2\pi\varphi_n),$$

also wäre nach (3. 1) auch

$$\mathrm{Re}\ \lambda \leqq 1 - \alpha \quad \text{und daher} \quad |\lambda - 1| \geqq \alpha$$

im Wiederspruch zu (iii a).

Für alle $n \geqq n_1$ ist, wie man leicht sieht, $\varphi_n \geqq \varphi_{n_1}$ und wir bezeichnen φ_{n_1} kurz mit φ.

Es gilt also für alle $\lambda \in \mathrm{Sp}\left(a_{\frac{1}{n}}\right)$ mit $n \geqq n_1$

$$|\arg \lambda| < \frac{1}{2} - \varphi.$$

Daher existieren die Logarithmen

$$b(n) = \log\left(a_{\frac{1}{n}}\right)$$

und es ist $a_1 = \exp(nb(n))$.

Nun führen wir die Bezeichnung

$$A_n = \{\arg \lambda, \lambda \in \mathrm{Sp}\left(a_{\frac{1}{n}}\right)\}$$

ein. Natürlich ist $A_n \subset \left[-\frac{1}{2}, \frac{1}{2}\right)$. Weiter gilt nach dem Spektralbildungssatz und nach (ii) $r \cdot A_{r \cdot n} = A_n$ für alle natürlichen n und r, wobei die Multiplikation mod 1 zu verstehen ist. Wir wollen zeigen, daß auch umgekehrt

$$A_{r \cdot n} = \frac{1}{r} A_n$$

ist.

Nach Hilfssatz 3. 7 gibt es zu festem n_2 und zu jedem $\beta\in[-\frac{1}{2},\frac{1}{2})$ einen Bruch $\frac{p}{q}$ mit $q\leqq n_2$, sodaß $\left|\beta-\frac{p}{q}\right|<\frac{1}{q\cdot n_2}$ ist. Wählt man $n_2\geqq\max(3,n_1)$, sodaß $\frac{1}{n_2}<\varphi$ und $n\geqq n_3=n_2^2$ ist, dann gibt es also zu jedem β ein $\frac{p}{q}$ mit $q\leqq n_2$, $(p,q)=1$ und

$$\left|\beta-\frac{p}{q}\right|<\frac{\varphi}{q}.$$

Es sei $\lambda\in\mathrm{Sp}\left(a_{\frac{1}{r\cdot n!}}\right)$, $r\geqq 1$ und $\beta=\arg\lambda$. Wäre $|\beta|\geqq\frac{1}{n_2}$ dann gäbe es ein $\frac{p}{q}$, $2\leqq q\leqq nn_2$, $p\neq 0$ mit $\left|\beta-\frac{p}{q}\right|<\frac{\varphi}{q}$. Wäre q gerade, so wäre $\frac{q}{2}\,\beta\in A_{\frac{2\cdot r\cdot n!}{q}}$ und $\left|\frac{q}{2}\beta-\frac{1}{2}\right|\leqq\frac{q}{2}\left|\beta-\frac{p}{q}\right|<\varphi$. Also muß q ungerade sein. Wählt man j so, daß $q^{j+1}\geqq n_2>q^j$ ist, dann gibt es in $A_{q^j\cdot r\cdot n!}$ ein β_1 mit $q^j\cdot\beta_1=\beta$, bzw. $\beta_1=\frac{\beta}{q^j}+\frac{k}{q^j}$ und es ist

$$\left|\beta_1-\frac{p+kq}{q^{j+1}}\right|=\left|\frac{\beta}{q^j}-\frac{p}{q^{j+1}}\right|\leqq\frac{1}{q^j}\left|\beta-\frac{p}{q}\right|<\frac{\varphi}{q^{j+1}}.$$

Dann gibt es ein natürliches $s<q^{j+1}$ mit

$$\frac{s(p+kq)}{q^{j+1}}\equiv\frac{1}{2}-\frac{1}{2q^{j+1}}$$

und daher wäre

$$\left|s\beta_1-\frac{1}{2}\right|\leqq\left|s\beta_1-\frac{s(p+kq)}{q^{j+1}}\right|+\frac{1}{2q^{j+1}}<\varphi.$$

Anderseits aber wäre $s\leqq n_2^2\leqq n$, somit $s\beta_1\in A_{\frac{q^j\cdot r\cdot n!}{s}}$ und daher $\left|s\cdot\beta_1-\frac{1}{2}\right|>\varphi$. Damit ist gezeigt

$$A_{r\cdot n!}\subset\left(-\frac{1-\varphi}{n_2},\frac{1-\varphi}{n_2}\right)\subset\left(-\frac{1}{3},\frac{1}{3}\right). \tag{3. 3}$$

Somit sind für alle natürlichen k der Gestalt $k=r\cdot n_3!$ die Voraussetzungen des Hilfssatzes 3. 4 erfüllt. Es existiert also ein $b\in B$ mit $a_{\frac{1}{k}}=\exp\left(\frac{1}{k}b\right)$ für alle k der Gestalt $k=r\cdot n_3!$, $r=1,2,\ldots$ Schließlich folgert man aus (ii) daß für alle natürlichen n gilt: $a_{\frac{1}{n}}=\exp\left(\frac{1}{n}b\right)$.

BEMERKUNG 3. 8: Satz 3. 1 gilt auch für beliebige Banachalgebren (mit Einheitselement), wenn die Bedingung (iv) durch (iv a) ersetzt wird. Nun wollen wir noch zeigen, daß der Hilfssatz 3. 5 nicht mehr wesentlich verbessert werden kann.

BEISPIEL 3. 9. B sei eine Banachalgebra mit Einheit $\mathbf{e}$ und mit unendlich vielen paarweise orthogonalen Idempotenten $\{z_j\}$. Es sei

$$u_{\frac{1}{n}} = e^{\frac{\pi i}{n}} \cdot z_n + e^{-\frac{\pi i}{n}} (\mathbf{e} - z_n)$$

für alle n. Dann sind die $u_{\frac{1}{n}}$ n-te Wurzeln von $-\mathbf{e}$ und wie man leicht sieht, ist

$$\mathrm{Sp}\,(u_{\frac{1}{n}}) = \{e^{\frac{\pi i}{n}}, e^{-\frac{\pi i}{n}}\}$$

also ist

$$\max_{\lambda \in \mathrm{Sp}\,(u_{\frac{1}{n}})} |\arg \lambda| = \frac{1}{2n}.$$

Es gibt jedoch sicher kein festes $b \in B$ mit $u_{\frac{1}{n}} = \exp\left(\frac{1}{n} b\right)$. Es müßte sonst $n \cdot \log(u_{\frac{1}{n}}) = m \cdot \log(u_{\frac{1}{m}})$ für alle n und m sein, oder auch

$$n \cdot \log u_{\frac{1}{n}} = n\left(\frac{\pi i}{n} z_n - \frac{\pi i}{n} (\mathbf{e} - z_n)\right) = 2\pi i z_n - \pi i \mathbf{e} = 2\pi i z_m - \pi i \mathbf{e}.$$

Daraus würde aber $z_n = z_m$ im Widerspruch zur Voraussetzung folgen.

BEISPIEL 3. 10. Die z_n seien wie im Beispiel 3. 9 definiert. Sei nun $v_0 = \mathbf{e}$, $v_1 = -\mathbf{e}$. Für jede Primzahl p und jedes natürliche n sei

$$v_{\frac{1}{p^n}} = e^{\frac{\pi i}{p^n}} z_p + e^{-\frac{\pi i}{p^n}} (\mathbf{e} - z_p).$$

Ist r rational, so hat r eine eindeutig bestimmte Darstellung $r = n + \sum \frac{k_i}{p_i^{m_i}}$ mit endlich vielen Primzahlen p_i und ganzen Zahlen $|k_i| < p_i^{m_i}$. Setzt man $v_r = v_{n + \sum \frac{k_i}{p_i^{m_i}}} = \prod_i \left(v_{\frac{1}{p_i^{m_i}}}\right)^{k_i} \cdot v_1^n$ so erhält man eine einparametrige Gruppe v_r. Insbesonders erfüllen die Wurzeln $v_{\frac{1}{n}}$ eine Verträglichkeitsbedingung, es gibt jedoch wiederum kein festes b mit $v_{\frac{1}{n}} = \exp\left(\frac{1}{n} b\right)$; es ist aber andererseits

$$\max_{\lambda \in \mathrm{Sp}\left(v_{\frac{1}{p^n}}\right)} |\arg \lambda| = \frac{1}{2p^n} \quad \text{für alle Primzahlpotenzen } p^n.$$

Das Beispiel 3. 10 lehrt auch, daß die Behauptung des Satzes 3. 1 im allgemeinen falsch wird, wenn man die Bedingung (iii) nur für eine (unendliche) Teilmenge der natürlichen Zahlen fordert.

BEMERKUNG 3. 11. Es lassen sich auch Algebren komplexer Maße angeben, die unendlich viele, paarweise orthogonale Idempotente besitzen, z. B. die Algebra der komplexen Maße einer kompakten Gruppe, die das direkte Produkt unendlich vieler kompakter Gruppen ist.

BEMERKUNG 3. 12. Der Satz 3. 1 kann noch verschärft werden, indem man ähnlich wie im Satz 2. 1 die Bedingung (iii) ersetzt durch (iii′) $\|a_{\frac{1}{n}} - e\| < \alpha_n < 2$, wobei $\{\alpha_n\}$ eine gegen 2 konvergente Folge positiver Zahlen ist. Da aber der Folge $\{\alpha_n\}$ weitere komplizierte Bedingungen auferlegt werden müssen und es bis jetzt nicht gelungen ist, eine ähnlich einfache Formulierung wie im Satz 2. 1 zu erreichen, verzichten wir auf die Wiedergabe von Einzelheiten.

LITERATUR

[1] PARTHASARATHY, K. R., RAO, R., VARADHAN, S. R. S.: Probability distributions on locally compact Abelian groups, *Illinois J. of Math.* **7** (1963) 337—369.
[2] CARNAL, H.: Unendlich oft teilbare Wahrscheinlichkeitsverteilungen auf kompakten Gruppen, *Math. Ann.* **153** (1964) 351—383.
[3] URBANIK, K.: Poisson distributions on a compact topological group, *Coll. Math.* **6** (1958) 13—24.
[4] BÖGE, W.: Zur Charakterisierung sukzessiv unendlich teilbarer Wahrscheinlichkeitsverteilungen auf lokalkompakten Gruppen. *Zeitschr. f. Wahrscheinlichkeitstheorie u. verw. Gebiete* **2** (1964) 380—394.
[5] PYM, J. S.: Idempotent measures on semigroups, *Pacific J. Math.* **12** (1962) 685—698.
[6] RICKART, CH. E.: *General theory of Banach Algebras,* D. v. Norstrand Comp. New York 1960.
[7] HILLE, E.: On Roots and Logarithms of Elements of a Complex Banach Algebra. *Math. Ann.* **136** (1958) 46—57.
[8] KOKSMA, J. F.: *Diophantische Approximationen.* Springer, Berlin 1935.
[9] BÖGE, W.: Über die Charakterisierung unendlich teilbarer Wahrscheinlichkeitsverteilungen, *J. reine u. angew. Math.* **201** (1959) 150—156.

Mathematisches Institut der Universität, Wien

(Eingegangen: 20. December 1968)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 75—87

ON THE ORDER OF CONVERGENCE OF FINITE-DIFFERENCE APPROXIMATIONS TO EIGENVALUES AND EIGENFUNCTIONS

by

L. VEIDINGER

1. Introduction

In some recent papers [1]—[4] I have studied the error in various finite-difference approximations to eigenvalues and eigenfunctions of self-adjoint elliptic differential operators. In the papers [1]—[3] first order methods, in the paper [4] a second order method has been considered.

In the present paper we shall be concerned with both first order and second order finite-difference analogs of eigenvalue problems. In Section 2 we shall obtain error bounds for the eigenfunctions of a first order finite-difference analog of the eigenvalue problem for the general second order self-adjoint elliptic differential operator (the corresponding error bounds for the eigenvalues have already been obtained in [1] and [2]). The first order method considered in Section 2 has been suggested by SAUL'EV in the paper [5]. In Section 3 we shall obtain error bounds for the eigenvalues and the eigenfunctions of a second order finite-difference analog of the fixed membrane problem. The second order method considered in Section 3 has been suggested by BRAMBLE in the paper [6]*.

2. Saul'ev's method for the general second order self-adjoint elliptic differential operator

Let R be a bounded open plane region whose boundary C consists of a finite number of piecewise-analytic simple closed curves. Denote by A_i $(i=1, 2, \ldots, n)$ the corners of C, i.e. those points on C where distinct analytic curves meet.

We consider the eigenvalue problem

$$Lu+\lambda u=0 \quad \text{in} \quad R, \tag{1}$$

$$u=0 \quad \text{on} \quad C,$$

where

$$Lu=\frac{\partial}{\partial x}\left[a(x, y)\frac{\partial u}{\partial x}\right]+\frac{\partial}{\partial x}\left[b(x, y)\frac{\partial u}{\partial y}\right]+$$

$$+\frac{\partial}{\partial y}\left[b(x, y)\frac{\partial u}{\partial x}\right]+\frac{\partial}{\partial y}\left[c(x, y)\frac{\partial u}{\partial y}\right]-f(x, y)u.$$

* In the paper [4] similar but weaker results were obtained for the well-known SHORTLEY—WELLER—MIKELADZE method. We have to remark, however, that the SHORTLEY—WELLER—MIKELADZE operator is not always self-adjoint in the sense defined in that paper and, consequently, the proofs where this property was used are incorrect.

Let the coefficients $a(x, y)$, $b(x, y)$, $c(x, y)$, $f(x, y)$ be analytic in an open region G containing the closure of R in its interior. Suppose that at all points of R

$$a\xi^2 + 2b\xi\eta + c\eta^2 \geqq \alpha(\xi^2 + \eta^2) \quad (\alpha = \text{const} > 0)$$

for all real ξ, η. Moreover, we assume that $f(x, y) \geqq 0$.

Let $\lambda^1 \leqq \lambda^2 \leqq \ldots$ be the eigenvalues of the problem (1) and let $u^1, u^2 \ldots$ be the corresponding eigenfunctions. We may assume that the eigenfunctions $u^1, u^2, \ldots$ are orthonormal, in the sense that

$$\iint_R u^k u^l \, dx \, dy = \delta_{kl} = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases}$$

Suppose the infinite plane of the region R is subdivided by two families of parallel lines into a square net. Let the lines of the net be $x = mh$ and $y = nh$ $(m, n = 0, \pm 1, \pm 2, \ldots)$. The points (mh, nh) will be called the nodes of the net. The smallest squares bounded by four lines of the net are called meshes of the net. Denote by S_h the set of all nodes of the plane.

Let R^* be then union of all meshes contained in R and let C^* be the boundary of R^*. Let R_h^* consist of all the interior nodes of R^* and let C_h^* be the net boundary of R_h^*. Let $\bar{R}_h^* = R_h^* \cup C_h^*$. Denote by R_h the set of all nodes in R. The points of intersection of the net lines with the boundary C form the set C_h.

We define

$$V_x(P) = h^{-1}[V(E) - V(P)], \quad V_{\bar{x}}(P) = h^{-1}[V(P) - V(W)],$$
$$V_y(P) = h^{-1}[V(N) - V(P)], \quad V_{\bar{y}}(P) = h^{-1}[V(P) - V(S)],$$

where $V = V(P)$ is any real-valued function defined on $\bar{R}_h^*$,

$$E = (x_P + h, y_P), \quad N = (x_P, y_P + h), \quad W = (x_P - h, y_P), \quad S = (x_P, y_P - h)$$

are the four neighbours of the node $P = (x_P, y_P)$.

If V and W are any two functions defined on R_h^*, then we define the scalar product and the norm of these functions by

$$(V, W) = \sum_{P \in R_h^*} V(P) W(P) h^2, \quad \|V\| = (V, V)^{\frac{1}{2}}.$$

The problem (1) is approximated by the finite-difference problem

$$L_h U + \lambda_h U = 0 \quad \text{in} \quad R_h^*, \tag{2}$$
$$U = 0 \quad \text{on} \quad C_h^*,$$

where

$$L_h U = 0{,}5[(aU_x)_{\bar{x}} + (aU_{\bar{x}})_x + (bU_y)_{\bar{x}} + (bU_{\bar{y}})_x +$$
$$+ (bU_x)_{\bar{y}} + (bU_{\bar{x}})_y + (cU_{\bar{y}})_y + (cU_y)_{\bar{y}}] - fU.$$

It is easy to see that the matrix corresponding to the operator L_h is symmetric and, consequently, the operator L_h is self-adjoint in the sense that

$$(V, L_h W) = (W, L_h V)$$

for any two functions V, W vanishing on C_h^*.

Let $\lambda_h^1 \leqq \lambda_h^2 \leqq \ldots$ be the eigenvalues of the problem (2) and let $U^1, U^2, \ldots$ be the corresponding eigenfunctions. We may assume that the eigenfunctions $U^1, U^2, \ldots$ are orthonormal, in the sense that

$$(U^k, U^l) = \delta_{kl}.$$

Consider the quadratic form

$$H_h(Z) = 0{,}5h^2 \sum_{P \in S_h} \{a(P)\left[(Z_x(P))^2 + (Z_{\bar{x}}(P))^2\right] +$$

$$+ 2b(P)[Z_x(P)Z_y(P) + Z_{\bar{x}}(P)Z_{\bar{y}}(P)] + c(P)\left[(Z_y(P))^2 + (Z_{\bar{y}}(P))^2\right] +$$

$$+ 2f(P)[Z(P)]^2\},$$

where $Z = Z(P)$ is any function defined at the nodes which vanishes outside R_h^*. We define

$$\|\delta Z\| = \left\{h^2 \sum_{P \in S_h} \left[(Z_x(P))^2 + (Z_y(P))^2\right]\right\}^{\frac{1}{2}}.$$

It is easy to see that

$$\|\delta Z\|^2 = O(H_h(Z)). \tag{3}$$

LEMMA 1. *For* $k = 1, 2, \ldots$

$$c_1 k < \lambda_h^k < c_2 k, \tag{4}$$

where c_1 *and* c_2 *are positive constants depending only on the region* R *and the coefficients of the operator* L.

For a PROOF see [7], p. 88.

LEMMA 2. *Let* $Z = Z(P)$ *be any function defined at the nodes which vanishes outside* R_h^*. *Then for* h *sufficiently small*

$$\max_{P \in R_h^*} |Z(P)| < c_3 |\log h|^{\frac{1}{2}} \|\delta Z\|, \tag{5}$$

where c_3 *is a positive constant depending only on the region* R.

For a PROOF see [8], p. 239.

THEOREM 1. *If* λ^k *is a simple eigenvalue of the problem* (1), *then for* h *sufficiently small*

$$\max_{P \in R_h^*} |u^k(P) - U^k(P)| < c_4 h^{\frac{1}{2}} |\log h|^{\frac{1}{2}}, \tag{6}$$

where c_4 *is a positive constant depending only on* k, *the region* R *and the coefficients of the operator* L. *If* $\lambda^k = \lambda^{k+1} = \cdots = \lambda^{k+m-1}$ *is an* m*-fold multiple eigenvalue of the problem* (1), *then for* h *sufficiently small*

$$\max_{P \in R_h^*} \left| u^k(P) - \sum_{i=0}^{m-1} \beta_h^{k+i} U^{k+i}(P) \right| < c_5 h^{\frac{1}{2}} |\log h|^{\frac{1}{2}}, \tag{7}$$

where $\beta_h^k, \ldots, \beta_h^{k+m-1}$ *are real coefficients and* c_5 *is a positive constant depending only on* k, *the region* R *and the coefficients of the operator* L.

PROOF. Let us first suppose that λ^k is a simple eigenvalue of the problem (1). Let $z^k = u^k - U^k$. Then

$$L_h z^k + \lambda_h^k z^k = \Phi_h^k \quad \text{in} \quad R_h^*,$$
$$z^k = u^k \quad \text{on} \quad C_h^*,$$

where $\Phi_h^k = L_h u^k - Lu^k + (\lambda_h^k - \lambda^k)u^k$. It is easy to see that $z^k = v^k + w^k$, where v^k is the solution of the problem

$$L_h v^k = \Phi_h^k \quad \text{in} \quad R_h^*,$$
$$v^k = u^k \quad \text{on} \quad C_h^* \tag{8}$$

and w^k is a solution of the problem

$$L_h w^k + \lambda_h^k w^k = -\lambda_h^k v^k \quad \text{in} \quad R_h^*,$$
$$w^k = 0 \quad \text{on} \quad C_h^*.$$

It has been proved in [2] that

$$\lambda^k - \lambda_h^k = O(h).$$

Using this result the function v^k can be estimated in the same way as the truncation error in the finite-difference approximation to the solution of the Dirichlet problem (see [9]). Thus we obtain

$$\|v^k\| = O(h^{\frac{1}{2}}) \tag{9}$$

and

$$\max_{P \in R_h^*} |v^k(P)| = O(h^{\frac{1}{2}} |\log h|^{\frac{1}{2}}). \tag{10}$$

Let $c_k = (w^k, U^k)$ and $W^k = w^k - c_k U^k$. The function W^k is a solution of the problem

$$L_h W^k + \lambda_h^k W^k = -\lambda_h^k v^k \quad \text{in} \quad R_h^*,$$
$$W^k = 0 \quad \text{on} \quad C_h^*$$

and satisfies the condition $(W^k, U^k) = 0$. Hence it follows that

$$W^k = \sum_{\substack{j=1 \\ (j \neq k)}}^{p} \frac{\lambda_h^k}{\lambda_h^j - \lambda_h^k} (v^k, U^j) U^j, \tag{11}$$

where p is the number of nodes in R_h^*. From (11), using the orthonormality of the eigenfunctions $U^1, U^2, \ldots$ and applying Schwarz's inequality we get

$$\|W^k\|^2 \leqq \|v^k\|^2 \sum_{\substack{j=1 \\ (j \neq k)}}^{p} \left(\frac{\lambda_h^k}{\lambda_h^j - \lambda_h^k} \right)^2.$$

Hence, using (4) and (9) we have

$$\|W^k\| = O(\|v^k\|) = O(h^{\frac{1}{2}}). \tag{12}$$

By the definition of W^k we have

$$W^k = u^k - U^k - v^k - c_k U^k = u^k - v^k - (1+c_k)U^k$$

and, consequently,

$$u^k = W^k + v^k + (1+c_k)U^k.$$

Hence, using Schwarz's inequality, (9) and (12) we obtain

$$(u^k, u^k) = (1+c_k)^2 + O(h^{\frac{1}{2}}). \tag{13}$$

But it is easy to show that

$$(u^k, u^k) = 1 + O(h). \tag{14}$$

Inserting (14) into (13) we get

$$2c_k + c_k^2 = O(h^{\frac{1}{2}}). \tag{15}$$

From (15) it follows that either $c_k = O(h^{\frac{1}{2}})$ or $c_k = -2 + O(h^{\frac{1}{2}})$. But if the sign of the function U^k is chosen so that $(u^k, U^k) \geqq 0$, then the second possibility is excluded and, consequently, $c_k = O(h^{\frac{1}{2}})$. Hence by (12) and the definition of W^k it follows that

$$\|w^k\| = O(h^{\frac{1}{2}}). \tag{16}$$

By the definition of w^k we have

$$-(w^k, L_h w^k) - \lambda_h^k (w^k, w^k) = \lambda_h^k (w^k, v^k).$$

Hence, applying the finite-difference analog of Green's first identity, we obtain

$$H_h(w^k) = \lambda_h^k (w^k, w^k) + \lambda_h^k (w^k, v^k). \tag{17}$$

From (17), using Schwarz's inequality, (9) and (16) we get

$$H_h(w^k) = O(h). \tag{18}$$

Hence by (3) and (5) it follows that

$$\max_{P \in R_h^*} |w^k(P)| = O(h^{\frac{1}{2}} |\log h|^{\frac{1}{2}}). \tag{19}$$

Combining (10) and (19) we obtain (6).

If $\lambda^k = \lambda^{k+1} = \cdots = \lambda^{k+m-1}$ is an m-fold multiple eigenvalue of the problem (1), then let

$$\bar{z}^k = u^k - \sum_{i=0}^{m-1} \beta_h^{k+i} U^{k+i},$$

where $\beta_h^{k+i} = (u^k - v^k, U^{k+i})$. The function $\bar{z}^k$ is a solution of the problem

$$L_h \bar{z}^k + \lambda_h^k \bar{z}^k = \bar{\Phi}_h^k \quad \text{in} \quad R_h^*,$$

$$\bar{z}^k = u^k \quad \text{on} \quad C_h^*,$$

where

$$\bar{\Phi}_h^k = \Phi_h^k - \sum_{i=0}^{m-1} (\lambda_h^k - \lambda_h^{k+i}) \beta_h^{k+i} U^{k+i}.$$

It is easy to see that $\bar{z}^k = v^k + \bar{w}^k$, where v^k is the solution of the problem (8) and $\bar{w}^k$ is a solution of the problem

$$L_h \bar{w}^k + \lambda_h^k \bar{w}^k = -\lambda_h^k v^k - \sum_{i=0}^{m-1} (\lambda_h^k - \lambda_h^{k+i}) \beta_h^{k+i} U^{k+i} \quad \text{in} \quad R_h^*, \tag{20}$$

$$\bar{w}^k = 0 \quad \text{on} \quad C_h^*.$$

The function $\bar{w}^k$ satisfies for $i = 0, 1, \ldots, m-1$ the condition $(\bar{w}^k, U^{k+i}) = 0$. Thus from (20) it follows that

$$\bar{w}^k = \sum_{\substack{j=1 \\ (j \neq k, k+1, \ldots, k+m-1)}}^{p} \frac{\lambda_h^k}{\lambda_h^j - \lambda_h^k} (v^k, U^j) U^j.$$

Hence by arguments almost identical with those used in the case of a simple eigenvalue we obtain

$$\max_{P \in R_h^*} |\bar{w}^k(P)| = O\left(h^{\frac{1}{2}} |\log h|^{\frac{1}{2}}\right)$$

and

$$\max_{P \in R_h^*} |\bar{z}^k(P)| = O\left(h^{\frac{1}{2}} |\log h|^{\frac{1}{2}}\right).$$

This completes the proof of Theorem 1.

Let A_i be a corner of C, with interior angle $\pi\alpha_i$ $(0 < \alpha_i < 2)$ and let

$$x^* = k_{A_i} x + l_{A_i} y, \quad y^* = m_{A_i} x + n_{A_i} y \tag{21}$$

be a linear transformation which transforms the operator L into the normal form at the point A_i. The transformation (21) transforms the angle $\pi\alpha_i$ into an angle $\pi\alpha_i^*$ $(0 < \alpha_i^* < 2)$. It is easy to see that $\alpha_i^* < 1$ if and only if $\alpha_i < 1$.

If $b(x, y) \equiv 0$, then the function v^k defined by the problem (8) satisfies instead of (10) the sharper relation (see [9])

$$\max_{P \in R_h^*} |v^k(P)| = O(h^{\beta^*}), \tag{22}$$

where

$$\beta^* = \begin{cases} 1, & \text{if} \quad \max\limits_{i=1,\ldots,n} \alpha_i^* < 1 \text{ or if there are no corners,} \\ \dfrac{1}{\max\limits_{i=1,\ldots,n} \alpha_i^*} - \varepsilon, & \text{if} \quad \max\limits_{i=1,\ldots,n} \alpha_i^* \geqq 1. \end{cases}$$

Using (22) instead of (10) we have instead of (6) and (7) the sharper estimates

$$\max_{P \in R_h^*} |z^k(P)| = O\left(h^{\beta^*} |\log h|^{\frac{1}{2}}\right) \tag{23}$$

and

$$\max_{P \in R_h^*} |\bar{z}^k(P)| = O\left(h^{\beta^*} |\log h|^{\frac{1}{2}}\right). \tag{24}$$

Thus we have proved the following theorem.

THEOREM 2. *Assume that* $b(x, y) \equiv 0$. *Then the functions* $z^k(P)$ *and* $\bar{z}^k(P)$ *satisfy* (23) *and* (24), *respectively.*

3. Bramble's method for the Laplace operator

In this section we consider the fixed membrane problem

$$\begin{aligned} \Delta u + \lambda u &= 0 \quad \text{in} \quad R, \\ u &= 0 \quad \text{on} \quad C, \end{aligned} \tag{25}$$

which is a special case of the problem (1).

An interior node $P=(x_P, y_P)$ is called regular if the four links joining P to its four neighbours $E = (x_P+h, y_P)$, $N = (x_P, y_P+h)$, $W=(x_P-h, y_P)$, $S = (x_P, y_P-h)$ all lie within the closed region $\bar{R}$. Let T_h be the set of all regular nodes and let $B_h = R_h - T_h$. The BRAMBLE operator $\Delta^{(h)}$ can be defined as follows. If $P \in T_h$, then let

$$\Delta^{(h)}V(P) = h^{-2}[V(E)+V(N)+V(W)+V(S)-4V(P)],$$

where $V=V(P)$ is any real-valued function defined on $\bar{R}_h = R_h \cup C_h$. If $P \in B_h$ and, for example, $E \notin \bar{R}_h$, $N \notin \bar{R}_h$, then let

$$\Delta^{(h)}V(P) = h^{-2}\left[\frac{1}{\alpha}V(E')+\frac{1}{\beta}V(N')+V(S)+V(W)-\left(\frac{\alpha+1}{\alpha}+\frac{\beta+1}{\beta}\right)V(P)\right],$$

where $E'=(x_P+\alpha h, y_P)$, $N' = (x_P, y_P+\beta h)$ are the points of C_h which lie closest to P on the corresponding net lines.

If V and W are any two functions defined on R_h, then we define the scalar product and the norm of these functions by

$$(V, W)_1 = \sum_{P \in R_h} V(P)W(P)h^2, \quad \|V\|_1 = (V, V)_1^{\frac{1}{2}}.$$

It is easy to see that the matrix corresponding to the operator $\Delta^{(h)}$ is symmetric and, consequently, the operator $\Delta^{(h)}$ is self-adjoint, in the sense that

$$(V, \Delta^{(h)}W)_1 = (W, \Delta^{(h)}V)_1 \tag{26}$$

for any two functions V, W vanishing on C_h.

The problem (25) is approximated by the finite-difference problem

$$\begin{aligned} \Delta^{(h)}V + \mu_h V &= 0 \quad \text{in} \quad R_h, \\ V &= 0 \quad \text{on} \quad C_h. \end{aligned} \tag{27}$$

Let $\mu_h^1 \leqq \mu_h^2 \leqq \ldots$ be the eigenvalues of the problem (27) and let $V^1, V^2, \ldots$ be the corresponding eigenfunctions. We may assume that the eigenfunctions $V^1, V^2 \ldots$ are orthonormal in the sense that

$$(V^k, V^l)_1 = \delta_{kl}.$$

LEMMA 3. *The eigenfunction V^k is bounded in R_h.*

PROOF. In the paper [6] a finite-difference analog of Green's function $G_h(P, Q)$ is defined by

$$\Delta_P^{(h)} G_h(P, Q) = -h^{-2}\delta(P, Q), \qquad P \in R_h,$$

$$G_h(P, Q) = \delta(P, Q), \qquad P \in C_h,$$

where

$$\delta(P, Q) = \begin{cases} 0, & \text{if } P \neq Q, \\ 1, & \text{if } P = Q. \end{cases}$$

The subscript P in the symbol $\Delta_P^{(h)}$ means that the operator is to be applied with respect to the variable P.

It is proved in [6] that

$$h^2 \sum_{Q \in R_h} G_h^2(P, Q) = O(1), \qquad P \in R_h. \tag{28}$$

Another finite-difference analog of Green's function $\bar{G}_h(P, Q)$ can be defined by

$$\Delta_P^{(h)} \bar{G}_h(P, Q) = -h^{-2}\delta(P, Q), \qquad P \in R_h.$$

$$\bar{G}_h(P, Q) = 0, \qquad P \in C_h.$$

It is easy to show that

$$\bar{G}_h(P, Q) = \sum_{j=1}^{q} \frac{V^j(P)V^j(Q)}{\mu_h^j}, \tag{29}$$

where q is the number of nodes in R_h. From (29) using the orthonormality of the eigenfunctions $V^1, V^2, \ldots$ we get

$$h^2 \sum_{Q \in R_h} \bar{G}_h^2(P, Q) = \sum_{j=1}^{q} \frac{[V^j(P)]^2}{(\mu_h^j)^2}. \tag{30}$$

By the definition of $G_h(P, Q)$ and $\bar{G}_P(P, Q)$ we have

$$G_h(P, Q) - \bar{G}_h(P, Q) = O(1). \tag{31}$$

From (28), (30) and (31) it follows that

$$\sum_{j=1}^{q} \frac{[V^j(P)]^2}{(\mu_h^j)^2} = O(1). \tag{32}$$

But it is easy to show (see [10], p. 214) that $\mu_h^k \to \mu^k$ and, consequently, from (32) immediately follows that the eigenfunction V^k is bounded in R_h.

THEOREM 3. *Let $A_1, A_2, \ldots, A_n$ be the corners of C with interior angles $\pi\alpha_1, \ldots, \pi\alpha_n$, respectively. Assume that the distance of the corner A_i from the nearest net line is at least $c_6 h$, where c_6 is a positive constant independent of h. Then for $k = 1, 2, \ldots$*

$$\lambda^k - \mu_h^k = O(h^\delta), \tag{33}$$

where

$$\delta = \begin{cases} 2, & \text{if } \max\limits_{i=1,\ldots,n} \alpha_i < 1 \text{ or if there are no corners,} \\ \dfrac{2}{\max\limits_{i=1,\ldots,n} \alpha_i} - \varepsilon, & \text{if } \max\limits_{i=1,\ldots,n} \alpha_i \geqq 1, \end{cases}$$

ε is any positive real number. Moreover, if λ^k is a simple eigenvalue of the problem (25), *then for h sufficiently small*

$$|u^k(P)-V^k(P)| < c_7\left(\sum_{i=1}^{n} h^{\delta_i}[r(P, A_i)]^{-\nu_i}+h^{\delta}|\log h|^{\frac{1}{2}}\right), \tag{34}$$

where $r(P, A_i)$ is the distance of the node P from the corner A_i, $\delta_i = \min\left(2, \frac{2}{\alpha_i}-2\varepsilon\right)$, $\nu_i = \min\left(1, \frac{1}{\alpha_i}-\varepsilon\right)$ and c_7 is a positive constant depending only on k and the region R. If $\lambda^k=\lambda^{k+1}=\cdots=\lambda^{k+m-1}$ is an m-fold multiple eigenvalue of the problem (25), *then for h sufficiently small*

$$\left|u^k(P)-\sum_{i=0}^{m-1}\gamma_h^{k+i}V^{k+i}(P)\right| < c_8\left(\sum_{i=1}^{n} h^{\delta_i}[r(P, A_i)]^{-\nu_i}+h^{\delta}|\log h|^{\frac{1}{2}}\right), \tag{35}$$

where $\gamma_h^k, \ldots, \gamma_h^{k+m-1}$ are real coefficients and c_8 is a positive constant depending only on k and the region R.

PROOF. We consider only the case when λ^k is a simple eigenvalue of the problem (25). The modifications needed in the case when λ^k is a multiple eigenvalue are obvious.

Let $Z^k = u^k - V^k$. The function Z^k satisfies the equation

$$\Delta^{(h)}Z^k+\mu_h^k Z^k = \Psi_h^k \quad \text{in} \quad R_h \tag{36}$$

and the boundary condition

$$Z^k=0 \quad \text{on} \quad C_h,$$

where $\Psi_h^k = \Delta^{(h)}u^k - \Delta u^k + (\mu_h^k-\lambda^k)u^k$. Multiplying (36) through by V^k and using (26) we get

$$(V^k, \Psi_h^k)_1 = (V^k, \Delta^{(h)}u^k - \Delta u^k)_1 + (\mu_h^k-\lambda^k)(u^k, V^k)_1 = 0,$$

whence

$$\lambda^k-\mu_h^k = \frac{(V^k, \Delta^{(h)}u^k-\Delta u^k)_1}{(V^k, u^k)_1} = \frac{\sum\limits_{P\in R_h} V^k(P)[\Delta^{(h)}u^k(P)-\Delta u^k(P)]h^2}{\sum\limits_{P\in R_h} V^k(P)u^k(P)h^2}. \tag{37}$$

Let

$$M_i(P) = \max\,[M_i^{(1)}(P), M_i^{(2)}(P)],$$

where

$$M_i^{(1)}(P) = \sup_{x_P-h_W<x<x_P+h_E}\left|\frac{\partial^i u^k(x, y)}{\partial x^i}\right|, \quad M_i^{(2)}(P) = \sup_{y_P-h_S<y<y_P+h_N}\left|\frac{\partial^i u^k(x, y)}{\partial y^i}\right|.$$

It is well-known that the function u^k is analytic in R and it is analytic on C, excluding the corners (see [11], p. 179 and [12]). Thus from our assumption on the placement of the net it follows that $M_i(P)$ is finite for all i and for all $P\in R_h$. By Taylor's theorem we have

$$|\Delta^{(h)}u^k(P)-\Delta u^k(P)| \leqq \begin{cases} \frac{1}{6}M_4(P)h^2, & P\in T_h, \\ 2M_2(P), & P\in B_h. \end{cases} \tag{38}$$

Let A_i be a corner of C with interior angle $\pi\alpha_i$ $(0<\alpha_i<2)$ and denote by $r(P, A_i)$ the distance of the node P from the corner A_i. Denote by D_{A_i,r_1} the set of all nodes of T_h and B_h, respectively, whose distance from A_i is less than a fixed positive real number r_1. Using (38), Lemma 3 and the well-known estimates of the derivatives of the function $u^k(x, y)$ in the neighbourhood of the corner A_i (see [13]) we obtain

$$\sum_{P\in D_{A_i,r_1}} V^k(P)[\Delta^{(h)}u^k(P)-\Delta u^k(P)]h^2 = O\Big(\sum_{P\in D_{A_i,r_1}} [r(P,A_i)]^{\frac{1}{\alpha_i}-4}h^4\Big) = \tag{39}$$

$$= O\Bigg(\sum_{1\leq m^2+n^2\leq\left(\frac{r_1}{h}\right)^2} (m^2+n^2)^{\frac{1}{2\alpha_i}-2}h^{\frac{1}{\alpha_i}}\Bigg) = \begin{cases} O(h^2), & \text{if } \alpha_i<\frac{1}{2},\\ O(h^{2-\varepsilon}), & \text{if } \alpha_i=\frac{1}{2},\\ O(h^{\frac{1}{\alpha_i}}), & \text{if } \alpha_i>\frac{1}{2}\end{cases}$$

and

$$\sum_{P\in E_{A_i,r_1}} V^k(P)[\Delta^{(h)}u^k(P)-\Delta u^k(P)]h^2 = O\Big(\sum_{P\in E_{A_i,r_1}} [r(P,A_i)]^{\frac{1}{\alpha_i}-2}h^2\Big) = \tag{40}$$

$$= O\Bigg(\sum_{1\leq n\leq\frac{r_1}{h}} n^{\frac{1}{\alpha_i}-2}h^{\frac{1}{\alpha_i}}\Bigg) = \begin{cases} O(h), & \text{if } \alpha_i<1,\\ O(h^{1-\varepsilon}), & \text{if } \alpha_i=1,\\ O(h^{\frac{1}{\alpha_i}}), & \text{if } \alpha_i>1.\end{cases}$$

Summing (39) and (40) over all corners of C and using the interior regularity of the function $u^k(x, y)$ we get

$$\sum_{P\in R_h} V^k(P)[\Delta^{(h)}u^k(P)-\Delta u^k(P)]h^2 = O(h^\gamma), \tag{41}$$

where

$$\gamma = \begin{cases} 1, & \text{if } \max\limits_{i=1,\dots,n}\alpha_i<1 \text{ or if there are no corners,}\\ 1-\varepsilon, & \text{if } \max\limits_{i=1,\dots,n}\alpha_i=1,\\ \dfrac{1}{\max\limits_{i=1,\dots,n}\alpha_i}, & \text{if } \max\limits_{i=1,\dots,n}\alpha_i>1.\end{cases}$$

But it is easy to show that $\|u^k-V^k\|_1\to 0$ as $h\to 0$ (see [10], p. 214). Hence it follows that

$$\sum_{P\in R_h} V^k(P)u^k(P)h^2 \to 1 \quad \text{as} \quad h\to 0. \tag{42}$$

Substituting (41) into (37) and using (42) we get

$$\lambda^k-\mu_h^k = O(h^\gamma). \tag{43}$$

It is easy to see that $Z^k = \overline{V}^k + \overline{W}^k$, where $\overline{V}^k$ is the solution of the problem

$$\Delta^{(h)}\overline{V}^k = \Psi_h^k \quad \text{in} \quad R_h,$$

$$\overline{V}^k = 0 \quad \text{on} \quad C_h$$

and $\overline{W}^k$ is a solution of the problem

$$\Delta^{(h)}\overline{W}^k + \mu_h^k \overline{W}^k = -\mu_h^k \overline{V}^k \quad \text{in} \quad R_h,$$

$$\overline{W}^k = 0 \quad \text{on} \quad C_h.$$

Using (43) the function $\overline{V}^k$ can be estimated in the same way as the truncation error in the finite-difference approximation to the solution of the Dirichlet problem (see [14]). Thus we obtain

$$|\overline{V}^k(P)| = O\left(\sum_{i=1}^{n} h^{\delta_i}[r(P, A_i)]^{-\nu_i} + h^{\gamma}\right) = O(h^{\gamma}), \tag{44}$$

where $\delta_i = \min\left(2, \frac{2}{\alpha_i} - 2\varepsilon\right)$ and $\nu_i = \min\left(1, \frac{1}{\alpha_i} - \varepsilon\right)$. Hence by arguments almost identical with those used in the proof of Theorem 1 we get

$$|\overline{W}^k(P)| = O(h^{\gamma}|\log h|^{\frac{1}{2}}). \tag{45}$$

From (44) and (45) it follows that

$$|Z^k(P)| = O(h^{\gamma}|\log h|^{\frac{1}{2}}). \tag{46}$$

Hence, using the estimates of the function $u^k(x, y)$ and its derivatives in the neighbourhood of the corner A_i we obtain

$$|V^k(P)| = O([r(P, A_i)]^{\frac{1}{\alpha_i}} + h^{\gamma}|\log h|^{\frac{1}{2}}), \qquad P \in D_{A_i, r_1} \tag{47}$$

and

$$|V^k(P)| = O(h[r(P, A_i)]^{\frac{1}{\alpha_i}-1} + h^{\gamma}|\log h|^{\frac{1}{2}}), \qquad P \in E_{A_i, r_1}. \tag{48}$$

Using (47) we get instead of (39) the sharper estimate

$$\sum_{P \in D_{A_i, r_1}} V^k(P)[\Delta^{(h)}u^k(P) - \Delta u^k(P]h^2 = \tag{49}$$

$$= O\left(\sum_{P \in D_{A_i, r_1}} [r(P, A_i)]^{\frac{2}{\alpha_i}-4} h^4 + \sum_{P \in D_{A_i, r_1}} [r(P, A_i)]^{\frac{1}{\alpha_i}-4} h^{4+\gamma}|\log h|^{\frac{1}{2}}\right) =$$

$$= O\left(\sum_{1 \leqq m^2+n^2 \leqq \left(\frac{r_1}{h}\right)^2} (m^2+n^2)^{\frac{1}{\alpha_i}-2} h^{\frac{2}{\alpha_i}} + \right.$$

$$\left. + \sum_{1 \leqq m^2+n^2 \leqq \left(\frac{r_1}{h}\right)^2} (m^2+n^2)^{\frac{1}{2\alpha_i}-2} h^{\frac{1}{\alpha_i}+\gamma}|\log h|^{\frac{1}{2}}\right) = O(h^2 + h^{\frac{1}{\alpha_i}+\gamma-\varepsilon}).$$

Similarly, using (48) we get instead of (40) the sharper estimate

$$\sum_{P\in E_{A_i,r_1}} V^k(P)[\Delta^{(h)}u^k(P)-\Delta u^k(P)]h^2 = \tag{50}$$

$$= O\left(\sum_{P\in E_{A_i,r_1}} [r(P,A_i)]^{\frac{2}{\alpha_i}-3}h^3+\sum_{P\in E_{A_i,r_1}} [r(P,A_i)]^{\frac{1}{\alpha_i}-2}h^{2+\gamma}|\log h|^{\frac{1}{2}}\right) =$$

$$= O\left(\sum_{1\leqq n\leqq \frac{r_1}{h}} n^{\frac{2}{\alpha_i}-3}h^{\frac{2}{\alpha_i}}+\sum_{1\leqq n\leqq \frac{r_1}{h}} n^{\frac{1}{\alpha_i}-2}h^{\frac{1}{\alpha_i}+\gamma}|\log h|^{\frac{1}{2}}\right] = O(h^2+h^{\frac{1}{\alpha_i}+\gamma-\varepsilon}+h^{1+\gamma-\varepsilon}).$$

Using (49) and (50) instead of (39) and (40) we obtain instead of (39) the sharper estimate (33).

Using (33) we get instead of (44) the sharper estimate

$$|\overline{V}^k(P)| = O\left(\sum_{i=1}^{n} h^{\delta_i}[r(P,A_i)]^{-\nu_i}+h^{\delta}\right). \tag{51}$$

From (51) it follows that

$$\sum_{P\in D_{A_i,r_1}} [\overline{V}^k(P)]^2h^2 = O\left(\sum_{P\in D_{A_i,r_1}} [r(P,A_i)]^{-2\nu_i}h^{2\delta_i+2}+h^{2\delta}\right) = \tag{52}$$

$$= O\left(\sum_{1\leqq m^2+n^2\leqq\left(\frac{r_1}{h}\right)^2} (m^2+n^2)^{-\nu_i}h^{2\delta_i+2-2\nu_i}+h^{2\delta}\right) = O(h^{2\delta})$$

and

$$|\overline{V}^k(P)| = O(h^{\delta}), \quad \text{if} \quad r(P,A_i) > r_1, \qquad i = 1, 2, \ldots, n. \tag{53}$$

Summing (52) over all corners of C and using (53) we get

$$\|\overline{V}^k\|_1 = O(h^{\delta}). \tag{54}$$

From (54) by arguments almost identical with those used in the proof of Theorem 1 we obtain

$$|\overline{W}^k(P)| = O(h^{\delta}|\log h|^{\frac{1}{2}}). \tag{55}$$

Combining (51) and (55) we obtain (34). This completes the proof of Theorem 3.

REFERENCES

[1] Вейдингер, Л.: О вычислении собственных значений мембраны методом конечных разностей, *Ж. Вычисл. Мат. и Мат. Физ.* **4** (1964) 1037—1044.

[2] Вейдингер, Л.: Об оценке погрешности при нахождении собственных значений методом конечных разностей, *Ж. Вычисл. Мат. и Мат. Физ.* **5** (1965) 806—815.

[3] VEIDINGER, L.: О разностном методе Pólya, *Studia Sci. Math. Hungar.* **2** (1967) 193—199.

[4] Вейдингер, Л.: О вычислении собственных значений и собственных функций оператора Лапласа методом конечных разностей, *Ж. Вычисл. Мат. и Мат. Физ.* **6** (1966) 687—698.

[5] Саульев, В. К.: *К вопросу решения задачи о собственных значениях методом конечных разностей, в сб. „Вычисл. матем. и вычисл. техн." 2.* Изд-во АН СССР, Москва, 1955, 116—144.

[6] BRAMBLE, J. H.: Error estimates for difference methods in forced vibration problems. *J. SIAM Numer. Anal.* **3** (1966) 1—12.

[7] Саульев, В. К.: *Об оценке погрешности при нахождении собственных функций методом конечных разностей, в сб. „Вычисл. матем." 1.* Изд-во АН СССР, Москва, 1957, 87—115.

[8] BRAMBLE, J. H.: A second order finite difference analog of the first biharmonic boundary value problem, *Num. Math.* **9** (1966) 236—249.

[9] VEIDINGER, L.: On the order of convergence of finite-difference approximations to the solution of the Dirichlet problem in a domain with corners, *Studia Sci. Math. Hungar.* **3** (1968) 337—343.

[10] Ладыженская, О. А.: *Смешанная задача для гиперболического уравнения.* Гостехиздат, Москва, 1953.

[11] BERNSTEIN, D. L.: *Existence theorems in partial differential equations*, Ann. Math. Studies, **23**, Princeton Univ. Press, 1950.

[12] MORREY, C. B. and NIRENBERG, L.: On the analyticity of linear elliptic systems of partial differential equations, *Communs Pure and Appl. Math.* **10** (1957) 271—290.

[13] LEHMAN, R. S.: Developments at an analytic corner of solutions of partial differential equations, *J. Math. and Mech.* **8** (1959) 727—760.

[14] LAASONEN, P.: On the discretization error of the Dirichlet problem in a plane region with corners, *Ann. Acad. Sci. Fenn. A. I.* **408** (1967) 1—16.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received January 2, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 89—96

THE QUEUE *GI/M/1* WITH TRAFFIC INTENSITY ONE

by

N. U. PRABHU

1. Introduction

This paper is a continuation of an earlier paper by the author [5], which dealt with a single server queueing system $GI/M/1$ with a recurrent input and exponential service times. The waiting time $Y(t)$ for this queue was defined as follows:

$Y(t)$ = time which has elapsed since the arrival of the customer being served (if any) at time t.
= 0 if the counter is unoccupied.

The distribution of $Y(t)$ was obtained in [5] and this was used to derive the distributions of the queue-length $Q(t)$ and the conventionally defined waiting time $W(t)$. The stochastic processes $Y(t)$, $Q(t)$ and $W(t)$ have non-null limiting distributions if, and only if, the traffic intensity ϱ_2 is less than unity.

In the present paper we are concerned with the case where $\varrho_2 = 1$. We shall prove that for the above processes suitably normalized limiting distributions exist in this case.

The notations used here are the same as those of [5]. Thus the interarrival times $\{u_i,\ i \geqq 1\}$ are assumed to be independent random variables with a common distribution function (d.f.) $B(x)$, and the service times have the exponential density $\lambda e^{-\lambda x}$ $(0 < x < \infty)$. Let $\psi(\theta)$ be the Laplace—Stieltjes (L. S.) transform of $B(x)$. We assume that

$$0 < \varrho_2 = [-\lambda\psi'(0)]^{-1} < \infty, \qquad 0 < \sigma^2 = \lambda\psi''(0) < \infty. \tag{1}$$

Let $t_0, t_1, t_2, \ldots$ be the epochs of successive departures from the system. Then

$$Y(t_n) = Y(t_n - 0)$$

$$Y(t) = \max\,[0,\ Y(t_n - 0) - u_{n+1} + (t - t_n)] \tag{2}$$

$$t_n < t < t_{n+1} \qquad (n = 0, 1, \ldots).$$

The process $Y(t)$ is not Markovian, and the methods employed [5] to investigate it used the duality relationship between the given system and the one obtained from it by interchanging the inter-arrival times and the service times. Thus the second system is the queue $M/G/1$ where customers arrive in a Poisson process at a rate λ and the service times have the d.f. $B(x)$; its traffic intensity is $\varrho = \varrho_2^{-1}$. BRODY [1] and IGLEHART [4] have investigated this $M/G/1$ system in the case $\varrho = 1$. Let $W(t)$ be the virtual waiting time in this system, and

$$F(x;\ 0, t) = \Pr\{W(t) = 0 \mid W(0) = x\}. \tag{3}$$

It is known [5] that

$$\int_0^\infty e^{-\theta t} F(x; 0, t)\, dt = \frac{e^{-x\eta(\theta)}}{\eta(\theta)} \tag{4}$$

where $\eta = \eta(\theta)$ is the unique root of the equation

$$\eta = \theta + \lambda - \lambda\psi(\eta) \tag{5}$$

with $\eta(\infty) = \infty$. When $\varrho = 1$, $\eta(0+) = 0$ and (5) gives

$$\eta = \theta + \lambda - \lambda\left[\psi(0) + \eta\psi'(0) + \frac{1}{2}\eta^2\psi''(0) + o(\eta^2)\right].$$

Using (1) we obtain after simplification

$$\frac{1}{\eta(\theta)} = \frac{\sigma}{\sqrt{2}}\theta^{-\frac{1}{2}} + o\left(\theta^{-\frac{1}{2}}\right) \quad (\theta \to 0+). \tag{6}$$

Using this result and a Tauberian theorem BRODY [1] proved Lemma 1. Lemma 2 is also an easy consequence of (6) and holds for more general processes; see PRABHU [6]. These two lemmas are basic to our investigation of the system $GI/M/1$.

LEMMA 1. *In the system $M/G/1$ with traffic intensity one,*

$$F(0; 0, t) \sim \frac{\sigma}{\sqrt{2\pi}} t^{-\frac{1}{2}} + o\left(t^{-\frac{1}{2}}\right) \quad (t \to \infty). \tag{7}$$

LEMMA 2. *In the system $M/G/1$, let $T(x) = \inf\{t: W(t) = 0 | W(0) = x\}$ with $x > 0$. If the traffic intensity is one, then*

$$\lim_{x\to\infty} \Pr\left\{\frac{\sigma^2 T(x)}{x^2} \leqq y\right\} = G_{\frac{1}{2}}(y) \tag{8}$$

where $G_{\frac{1}{2}}(y)$ is the stable d.f. with index $\frac{1}{2}$ given by

$$G_{\frac{1}{2}}(y) \begin{cases} = \int_0^y \frac{1}{\sqrt{2\pi x^3}} e^{-\frac{1}{2x}}\, dx & (y > 0) \\ = 0 & (y \leqq 0). \end{cases} \tag{9}$$

PROOF. If is known [5] that $\mathsf{E}[e^{-\theta T(x)}] = e^{-x\eta(\theta)}$. Therefore

$$\mathsf{E}\left[e^{-\theta \frac{\sigma^2 T(x)}{x^2}}\right] = e^{-x\eta\left(\frac{\theta\sigma^2}{x^2}\right)} \to e^{-\sqrt{2\theta}} \tag{10}$$

and since $e^{-\sqrt{2\theta}}$ is the L.S. transform of the d.f. (9), the lemma follows from the continuity theorem.

2. The distribution of $Y(t)$

Let us assume that the system $GI/M/1$ starts at time $t=0$ with one customer, who has just commenced his service. Then $Y(0)=0$. Let $Z_1, Z_2, \ldots$ be the successive busy cycles in the system; the random variables $\{Z_n\}$ are independent and the L.S. transform of their common d.f. was found in [5] to be

$$\varphi(\theta) = \mathsf{E}(e^{-\theta Z_i}) = \frac{\psi(\theta)-\psi(\eta)}{1-\psi(\eta)}. \tag{11}$$

If $\varrho_2=1$, Z_i is finite with probability one, but $\mathsf{E}(Z_i)=\infty$. We have then the following

THEOREM 1. *In the system $GI/M/1$ with traffic intensity one, the sequence of busy cycles $\{Z_n\}$ belongs to the domain of attraction of a stable distribution with index $\frac{1}{2}$. In fact*

$$\lim_{n\to\infty} \Pr\left\{\frac{4(Z_1+Z_2+\cdots+Z_n)}{\sigma^2 n^2} \leqq x\right\} = G_{\frac{1}{2}}(x) \tag{12}$$

where $G_{\frac{1}{2}}(x)$ is defined in (9).

PROOF: When $\varrho_2=1$ we have as $\theta\to 0+$,

$$\lambda-\lambda\psi(\theta) = \lambda-\lambda\left[\psi(0)+\theta\psi'(0)+\frac{1}{2}\theta^2\psi''(0)+o(\theta^2)\right] = \theta-\frac{1}{2}\sigma^2\theta^2+o(\theta^2)$$

and similarly

$$\lambda-\lambda\psi(\eta) = \eta-\frac{1}{2}\sigma^2\eta^2+o(\eta^2).$$

Therefore, using (6) we obtain

$$\frac{1-\varphi(\theta)}{\theta} = \frac{1}{\theta}\cdot\frac{1-\psi(\theta)}{1-\psi(\eta)} = \frac{1-\frac{1}{2}\sigma^2\theta+o(\theta)}{\eta}\left[1-\frac{1}{2}\sigma^2\eta+o(\eta)\right]^{-1} = \tag{13}$$

$$= \frac{\sigma}{\sqrt{2}}\theta^{-\frac{1}{2}}+o(\theta^{-\frac{1}{2}}) \quad (\theta\to 0+).$$

It follows from a Tauberian theorem that

$$\Pr\{Z_1>z\} \sim \frac{\sigma}{\sqrt{2\pi}} z^{-\frac{1}{2}} \quad (z\to\infty) \tag{14}$$

and the desired result follows from a known theorem (FELLER [3], pp. 424—425). More directly we find, using (13), that

$$\mathsf{E}\left[e^{-\theta\frac{4(Z_1+Z_2+\cdots+Z_n)}{\sigma^2 n^2}}\right] = \left[\varphi\left(\frac{4\theta}{\sigma^2 n^2}\right)\right]^n = \left[1-\frac{\sqrt{2\theta}}{n}+o\left(\frac{1}{n}\right)\right]^n \to e^{-\sqrt{2\theta}}$$

and the result follows as in Lemma 2.

COROLLARY 1. *Let* $U(t) = 1 +$ *expected number of busy cycles during a time-interval* $(0, t]$. *If* $\varrho_2 = 1$, *then*

$$U(t) \sim \frac{2^{\frac{3}{2}} t^{\frac{1}{2}}}{\sigma\sqrt{\pi}} \qquad (t \to \infty). \tag{16}$$

PROOF: We have

$$U(t) = 1 + \sum_{n=1}^{\infty} \Pr\{Z_1 + Z_2 + \cdots + Z_n \leqq t\}$$

so that

$$U^*(\theta) = \int_0^{\infty} e^{-\theta t} U(dt) = \frac{1}{1-\varphi(\theta)} = \frac{1-\psi(\eta)}{1-\psi(\theta)}. \tag{17}$$

Proceeding as in (13) we find that

$$U^*(\theta) \sim \frac{\sqrt{2}}{\sigma} \theta^{-\frac{1}{2}} \qquad (\theta \to 0+). \tag{18}$$

Since $U(t)$ is a monotone function of t it follows from a Tauberian theorem that

$$U(t) \sim \frac{\sqrt{2}}{\sigma\Gamma\left(\frac{3}{2}\right)} t^{\frac{1}{2}} \qquad (t \to \infty)$$

and this is the desired result, since $\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\sqrt{\pi}$.

THEOREM 2. *If the system* $GI/M/1$ *has* $\varrho_2 = 1$, *then as* $t \to \infty$,

$$t^{\frac{1}{2}} \Pr\left\{\frac{Y(t)}{\sigma\sqrt{t}} \leqq x,\ Y(\tau) > 0 \ \ (0 < \tau \leqq t)\right\} \to \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{19}$$

PROOF: In [5] it was proved that

$$\Pr\{Y(t) \leqq x, \quad Y(\tau) > 0 \qquad (0 < \tau \leqq t)\} = F(x;\ 0,\ t), \tag{20}$$

so the result (19) can be written as

$$t^{\frac{1}{2}} F(\sigma x\sqrt{t}\,;\ 0,\ t) \to \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{21}$$

For $x = 0$ this reduces to Lemma 1; the Tauberian theorem is applicable because $F(0;\ 0,\ t)$ is a monotone (non-increasing) function of t, as is evident from

$$F(0;\ 0,\ t) = \Pr\{Y(\tau) > 0 \qquad (0 < \tau \leqq t)\}. \tag{22}$$

This fact seems to have been overlooked by BRODY [1] in his proof of (7). To prove (21) for $x > 0$ we recall the definition of the first passage time $T(x)$ given in Lemma 2. Let $G(x, t)$ be the d.f. of $T(x)$. Then considering the ordinary passage from $x\,(>0)$

to 0 and the first passage from x to 0 in the process $W(t)$ we obtain the relation

$$F(x;0,t)=\int_0^t G(x,d\tau)F(0;0,t-\tau)=\int_0^1 G(x,t\,du)F(0;0,t-tu). \tag{23}$$

Using Lemmas 1 and 2 we can then find that

$$t^{\frac{1}{2}}F(\sigma x\sqrt{t};0,t)=\int_0^1 G\left(\sigma x\sqrt{t};\frac{\sigma^2x^2t}{\sigma^2}\cdot\frac{du}{x^2}\right)t^{\frac{1}{2}}F(0;0,t-tu)\to \tag{24}$$

$$\to\int_0^1\frac{x^3}{\sqrt{2\pi u^3}}e^{-\frac{1}{2}\frac{x^2}{u}}\left(\frac{du}{x^2}\right)\frac{\sigma}{\sqrt{2\pi}}(1-u)^{-\frac{1}{2}}.$$

The transformation $\frac{x^2}{u}=x^2+y^2$ reduces the last integral to

$$\frac{2\sigma}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\int_0^\infty\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy=\frac{\sigma}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}.$$

This proves (21) for $x>0$.

THEOREM 3. *For the waiting time process $Y(t)$ in the queue $GI/M/1$ with traffic intensity one, we have*

$$\lim_{t\to\infty}\Pr\left\{\frac{Y(t)}{\sqrt{t}}\geqq x\right\}\begin{cases}=2\int_x^\infty e^{-\frac{1}{2}u^2}du & (x\geqq 0)\\ =1 & (x<0).\end{cases} \tag{25}$$

PROOF: We have

$$\Pr\{Y(t)\geqq x\}=\int_{0-}^t U(d\tau)\,F(x;0,t-\tau) \tag{26}$$

as was shown in [5]. We can write this as

$$\Pr\left\{\frac{Y(t)}{\sigma\sqrt{t}}\geqq x\right\}=\int_{0-}^1 U(t\,du)\,F(\sigma x\sqrt{t};0,t-tu)=$$

$$=\int_{0-}^1\frac{U(t\,du)}{U(t)}t^{\frac{1}{2}}F(\sigma x\sqrt{t};0,t-tu)t^{-\frac{1}{2}}U(t).$$

Using Corollary 1, Theorem 2 and the fact that as $t\to\infty$, the measure $U(t\,du)/U(t)$ tends to the measure with density $\frac{1}{2}u^{-\frac{1}{2}}$ (see FELLER [3], p. 447) we find that

$$\Pr\left\{\frac{Y(t)}{\sigma\sqrt{t}}\geqq x\right\}\to\frac{2}{\sqrt{2\pi}}\int_0^1\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\frac{x^2}{1-u}}\frac{du}{\sqrt{u(1-u)}}= \tag{27}$$

$$=\frac{2}{\sqrt{2\pi}}\int_x^\infty y\,dy\int_0^1\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\frac{y^2}{1-u}}\frac{du}{u^{\frac{1}{2}}(1-u)^{\frac{3}{2}}}.$$

The transformation $\frac{y^2}{1-u} = u^2+y^2$ reduces the last integral to

$$\frac{2}{y} e^{-\frac{1}{2}y^2} \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}\, dv = \frac{1}{y} e^{-\frac{1}{2}y^2} \tag{28}$$

and so the double integral in (27) reduces to the desired expression.

3. The queue-length and the waiting time

Let $Q(t)$ be the number of customers present in the system (including the one being served, if any) at time t. Clearly, if the counter is occupied, then $Q(t)-1$ is the number of arrivals during the time the customer being served has spent in the system up to that time. Therefore

$$Q(t) \begin{cases} = 1+\max\{n: u_1+u_2+\cdots+u_n \leqq Y(t)\} & \text{if} \quad Y(t)>0 \\ = 0 & \text{if} \quad Y(t)=0. \end{cases} \tag{29}$$

The virtual waiting time $W(t)$ is given by

$$W(t) \begin{cases} = v_1'-v_2+\cdots+v_{Q(t)} & \text{if} \quad Q(t)>0 \\ = 0 & \text{if} \quad Q(t)=0, \end{cases} \tag{30}$$

where v_1' is the residual service time of the customer at the counter, and $v_2, v_3, \ldots, v_{Q(t)}$ are the service times of the customers behind him.

In Theorems 5 and 6 we investigate the limiting behavior of $Q(t)$ and $W(t)$ in the case $\varrho_2 = 1$. We recall from [5] that when $\varrho_2 \geqq 1$ both $Y(t)$ and $Q(t)$ diverge to $+\infty$ in distribution. The following theorem is a stronger version of this result.

THEOREM 4. *In the queue GI/M/1 with* $\varrho_2 \geqq 1$, *as* $t \to \infty$,

$$Y(t) \to \infty, \quad Q(t) \to \infty \tag{31}$$

with probability one.

PROOF: From the definition of the process $Y(t)$ it is clear that with probability one

$$Y(t) \geqq t - X(t) \tag{32}$$

where $X(t)$ is the compound Poisson process with the L.S. transform

$$\mathsf{E}[e^{-\theta X(t)}] = e^{-\lambda t+\lambda t\psi(\theta)} \qquad (\theta>0).$$

The process $t-X(t)$ has stationary independent increments, and $\mathsf{E}[t-X(t)] = (1-\varrho_2^{-1})t$. Therefore if $\varrho_2 \geqq 1$ we have

$$\limsup_{t\to\infty} [t-X(t)] = +\infty \tag{33}$$

with probability one (see SPITZER [7] for a discussion of the discrete case of partial

sums of independent and identically distributed random variables). From (32) we therefore obtain

$$\limsup_{t\to\infty} Y(t) \geqq \liminf_{t\to\infty} Y(t) \geqq \limsup_{t\to\infty} [t - X(t)] = +\infty, \tag{34}$$

so that $\lim_{t\to\infty} Y(t) = +\infty$ with probability one. To complete the proof, let $N(y) = \max\{n\colon u_1 + u_2 + \cdots + u_n \leqq y\}$ for $y > 0$, so that $N(y)$ is the number of arrivals in a time-interval $(0, y]$. We can then write (29) as

$$Q(t) \begin{cases} = 1 + N[Y(t)] & \text{if} \quad Y(t) > 0 \\ = 0 & \text{if} \quad Y(t) = 0. \end{cases} \tag{35}$$

Now it is known as $y \to \infty$, $N(y) \to \infty$ with probability one. Letting $y \to \infty$ through values of $Y(t)$ we therefore see that $Q(t) \to \infty$ with probability one.

THEOREM 5. *In the queue GI/M/1 with* $\varrho_2 = 1$,

$$\lim_{t\to\infty} \Pr\left\{\frac{Q(t)}{\lambda\sigma\sqrt{t}} \geqq x\right\} \begin{cases} = 2\int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\, dy & (x \geqq 0) \\ = 1 & (x < 0). \end{cases} \tag{36}$$

PROOF: For $Y(t) > 0$ we can write (35) as

$$\frac{Q(t)}{\lambda\sigma\sqrt{t}} = \frac{1}{\lambda\sigma\sqrt{t}} + \frac{N[Y(t)]}{\lambda Y(t)} \cdot \frac{Y(t)}{\sigma\sqrt{t}}. \tag{37}$$

The first term on the right side of (37) tends to zero as $t \to \infty$. For the second term we recall a result of DOOB [2], which states that with probability one

$$\frac{N(y)}{y} \to \frac{1}{\mathsf{E}(u_n)} \tag{38}$$

as $y \to \infty$. Also, according to Theorem 4, $Y(t) \to \infty$ with probability one, so when $y \to \infty$ through values of $Y(t)$, (38) gives

$$\frac{N[Y(t)]}{Y(t)} \to \frac{1}{\mathsf{E}(u_n)} = \lambda \quad \text{as} \quad t \to \infty, \tag{39}$$

since $\varrho_2 = [\lambda \mathsf{E}(u_n)]^{-1} = 1$. Further, as $t \to \infty$, $Y(t)/\sigma\sqrt{t}$ has the limiting distribution given by (25). Therefore the left side in (37) has the limiting distribution given by (36).

THEOREM 6. *In the queue GI/M/1 with* $\varrho_2 = 1$,

$$\lim_{t\to\infty} \Pr\left\{\frac{W(t)}{\sigma\sqrt{t}} \geqq x\right\} \begin{cases} = 2\int_x^\infty e^{-\frac{1}{2}y^2}\, dy & (x \geqq 0) \\ = 1 & (x < 0). \end{cases} \tag{40}$$

PROOF: We can write (30) as

$$\frac{W(t)}{\sigma\sqrt{t}} = \lambda \cdot \frac{v_1' + v_2 + \cdots + v_{Q(t)}}{Q(t)} \cdot \frac{Q(t)}{\lambda\sigma\sqrt{t}}. \tag{41}$$

Here the random variables $v_1', v_2, \ldots$ are independent and have the exponential distribution with mean λ^{-1}, so by the strong law of large numbers

$$\frac{v_1' + v_2 + \cdots + v_n}{n} \to \frac{1}{\lambda} \quad \text{as} \quad n \to \infty \tag{42}$$

with probability one. By Theorem 4, $Q(t) \to \infty$ with probability one since $\varrho_2 = 1$. Therefore when $n \to \infty$ through values of $Q(t)$, (42) gives

$$\frac{v_1' + v_2 + \cdots + v_{Q(t)}}{Q(t)} \to \frac{1}{\lambda} \quad \text{as} \quad t \to \infty, \tag{43}$$

with probability one. Further, as $t \to \infty$, $Q(t)/\lambda\sigma\sqrt{t}$ has the limiting distribution given by (36). This completes the proof.

REFERENCES

[1] BRODY, S. M.: On a limit theorem of the theory of queues. *Ukrain. Mat. Z.* **15** (1963), 76—79.
[2] DOOB, J. L.: Renewal theory from the point of view of the theory of probability. *Trans. Amer. Math. Soc.* **63** (1948), 422—438.
[3] FELLER, W.: *An Introduction to Probability Theory and its Applications,* Vol. II, J. Wiley, New York. (1966).
[4] IGLEHART, D. L.: Limit theorems for queues with traffic intensity one. *Ann. Math. Statist.* **36** (1965), 1437—1449.
[5] PRABHU, N. U.: A waiting time process in the queue *GI/M/1*. *Acta Math. Acad. Sci. Hungar.* **15** (1964), 363—371.
[6] PRABHU, N. U.: Some new results in storage theory. *J. Appl. Prob.* **5** (1968), 452—460
[7] SPITZER, F.: A combinatorial lemma and its application to probability theory. *Trans. Amer. Math. Soc.* **82** (1956), 323—339.

Cornell University, New York

(Received August 7, 1968)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 97—98

ON A PROBLEM OF STEINHAUS

by

D. WATERMAN

Let $\{f_n\}$, $n=1, 2, \ldots$, be a finite or countably infinite sequence of measurable functions on (0, 1). Consider the system

$$\{f_1^{m_1}(x)\cdot f_2^{m_2}(x)\cdot\cdots\cdot f_n^{m_n}(x)\}, \tag{1}$$

where $m_i=0,1, 2, \ldots$, $n=1, 2, \ldots$. In 1937 [3] STEINHAUS raised the following question: If $\{f_n\}$ is a sequence of stochastically independent functions, what are the necessary and sufficient conditions that the system (1) be closed in L^2?

We say that $\{f_n\}$ is *maximal* if there is a set Z of measure zero such that if x_1 and $x_2 \notin Z$ and $f_n(x_1)=f_n(x_2)$ for every n, then $x_1=x_2$.

The following result answers the question and shows that the notion of independence is irrelevant here.

THEOREM. *If $\{f_n\}$ is a sequence of bounded measurable functions, then the system* (1) *is closed in L^2 if and only if $\{f_n\}$ is maximal.*

RÉNYI introduced the notion of maximality in 1952 [2] and demonstrated its sufficiency, but not its necessity, remarking, in fact, that the problem of STEINHAUS was still unsolved [2, p. 287]. Professor RÉNYI has kindly brought to our attention a prior demonstration of this result from a different approach by R. F. GUNDY [1].

We will now assume the closure of the system (1) in L^2 and produce a set Z with the required properties.

Consider the function $f(x)=x$. There exists a sequence $\{P_n\}$ of linear combinations of elements of the system (1) such that

$$\|P_n-f\|_2 \to 0 \quad \text{as} \quad n\to\infty.$$

Thus there is a sequence $\{n_k\}$ such that

$$P_{n_k}(x)\to x \quad \text{a.e. in} \quad (0, 1) \quad \text{as} \quad k\to\infty.$$

Let

$$Z=\{x: P_{n_k}(x)\not\to x\}.$$

Evidently $m(Z)=0$. Now if x_1 and $x_2 \notin Z$ and $f_n(x_1)=f_n(x_2)$ for every n, we have

$$x_1 = \lim P_{n_k}(x_1) = \lim P_{n_k}(x_2) = x_2 .$$

REFERENCES

[1] GUNDY, R. F.: Complete systems in L_2 and a theorem of Rényi, *Mich. Math. J.* **12** (1965), 161—167.

[2] RÉNYI, A.: On a conjecture of H. Steinhaus, *Annales Soc. Math. Polonaise (Rocznik Polskiego Tow. Matematycznego)* **25** (1952), 279—287.

[3] STEINHAUS, H.: La théorie et les applications des fonctions indépendantes au sens stochastique, *Colloque consacré à la Théroie des Probabilités, Part V, Actualités Sci. et Ind.* **738** (1938), 57—73.

Syracuse University, Syracus, New York

(Received November 3, 1968)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 99—125

ON SECOND ORDER ESTIMATES FOR THE CODING THEOREM AND ITS STRONG CONVERSE

by

U. AUGUSTIN

Introduction

The paper deals with memoryless channels for discrete time without stationarity requirements or conditions on the alphabets. We are concerned with estimates for the maximal length $N_t(\varepsilon)$ of ε-codes for the time t when ε $(0<\varepsilon<1)$ is fixed and t increases.

The author describes in [1] necessary and sufficient conditions for the coding theorem and its strong converse for those channels i.e. conditions for the validity of $|\ln N_t(\varepsilon) - C_t| = o(t)$ $(0<\varepsilon<1)$ where C_t is a non-stationary substitute for $t \cdot C$ (C capacity).

Here we will discuss estimates of the form

$$|\ln N_t(\varepsilon) - C_t + \Phi^{-1}(\varepsilon)\, D_t(\varepsilon)| = o(\sqrt{t}) \qquad (\varepsilon \text{ fixed},\ 0<\varepsilon<1)$$

where Φ^{-1} is the inverse function of the normal distribution function and $D_t(\varepsilon)$ a variance parameter depending on t and ε. Such estimates have been derived in [4] for stationary finite alphabet channels without memory by methods, which allow no direct extensions to our situation. However, our main results in § 5 are of a weaker type than those in [4].

§ 1. The Model for the Channel

Let the triple (X, F, M) be given, where $X \neq \emptyset$ denotes a set, F a σ-field on X, $M = \{p\}$ a non-empty set of probabilities on (X, F).

DEFINITION 1. 1. (ε-code and $N(M, \varepsilon)$):
A finite ε-code $(0<\varepsilon<1)$ for (X, F, M) (or M) is a finite sequence $\{(p^i, E_i)\}_{1\leqq i\leqq N}$ of pairs (p^i, E_i) $(p^i \in M,\ E_i \in F)$ where the E_i are pairwise disjoint and $p^i(E_i) > \varepsilon$ holds $(1 \leqq i \leqq N)$. N is called the length of the code. We set

$N(M, \varepsilon) := \sup \{N \text{ natural}: N \text{ is length of an } \varepsilon\text{-code for } M\}$.

Let $[1, t] := \{v \text{ natural}: 1 \leqq v \leqq t\}$ (the time) and let for every $v \in [1, t]$ (X_v, F_v, M_v) be given similarly as above, but not necessarily copies of each other. We denote

$$X_{[1,t]} := \prod_{v=1}^{t}{}^{\times} X_v \quad \text{(product space)},$$

$$F_{[1,t]} := \left(\prod_{v=1}^{t}{}^{\times} F_v\right) \quad \text{(product } \sigma\text{-field)},$$

$$M_{[1,t]} := M_1 \times \cdots \times M_t := \{p_1 \times \cdots \times p_t : p_v \in M_v\} \quad \text{on} \quad (X_{[1,t]}, F_{[1,t]}).$$

DEFINITION 1. 2 (channel without memory):
$M_{[1,t]}$ is called a channel without memory for the discrete time $[1, t]$.

REMARK: The reader may consider (X, F, M) in the following way: Interpret X as the set of output symbols of a channel and M as the set of all transition probabilities from a set of input symbols to X (for a fixed time). Seen from the decoder the structure of the set of transition probabilities only or (less) the structure of the set of output sources is interesting for the relation between decoding error and length of a code. One finds therefore, that definition 1. 2 is a version of the classical definition of a channel without memory (see [3]).

We are going to derive estimates for $N(M_{[1,t]}, \varepsilon)$ by deriving estimates on $N(M', \varepsilon)$ for finite subsets $M' \subseteqq M_{1,t]}$.

§ 2. Stochastic inequalities for $N(M, \cdot)$

LEMMA 2. 1 (Maximal code theorem, see [3]):

Let $M' = \{p^{(j)}\}_{1 \leqq j \leqq n}$ *be a finite set of probabilities on* (X, F), $a^{(1)}, \ldots, a^{(n)} \geqq 0$ *real with* $\sum_{j=1}^{n} a^{(j)} = 1$, $q := \sum_{j=1}^{n} a^{(j)} p^{(j)}$. *Then for any* $S > 0$ *holds*

$$N(M', \varepsilon) > S \cdot \left[\sum_{j=1}^{n} a^{(j)} p^{(j)} \left\{ \frac{dp^{(j)}}{dq} > S \right\} - \varepsilon \right].$$

We recall the short PROOF in [3]:
Assume $p^{(j)} \left\{ \frac{dp^{(j)}}{dq} > S \right\} > \varepsilon$ for at least one j (otherwise the inequality in the lemma is trivially true). Then there is an ε-code $\{(p^i, E_i)\}_{1 \leqq i \leqq N}$ of M' for some N which is of maximal length under the condition that $E_i \subseteqq \left\{ \frac{dp^i}{dq} > S \right\}$ $(1 \leqq i \leqq N)$. Hence $p^{(j)} \left(\left\{ \frac{dp^{(j)}}{dq} > S \right\} \cap \operatorname{compl} \left(\bigcup_{i=1}^{N} E_i \right) \right) \leqq \varepsilon$ $(1 \leqq j \leqq n)$ (or N is not maximal). Therefore $p^{(j)} \left\{ \frac{dp^{(j)}}{dq} > S \right\} - \varepsilon \leqq p^{(j)} \left(\bigcup_{i=1}^{N} E_i \right)$ and

$$\sum_{j=1}^{n} a^{(j)} p^{(j)} \left\{ \frac{dp^{(j)}}{dq} > S \right\} - \varepsilon \leqq \sum_{j=1}^{n} a^{(j)} p^{(j)} \left(\bigcup_{i=1}^{N} E_i \right) = q \left(\bigcup_{i=1}^{N} E_i \right) \leqq$$

$$\leqq S^{-1} \sum_{i=1}^{N} p^i(E_i) \text{ because } \frac{dp^i}{dq} > S \text{ on } E_i \; p^i\text{-a.e.}$$

LEMMA 2. 2. *Let* $\{(p^i, E_i)\}$ *be an* ε*-code of length* N *for* M. *Then*

$$\frac{1}{N} > \frac{1}{N} \sum_{i=1}^{N} S_i^{-1} \left[\varepsilon - p^i \left\{ \frac{dp^i}{d\lambda} > S_i \right\} \right]$$

for arbitrary $S_1, \ldots, S_N > 0$ *and any probability* λ *with* $p^i \ll \lambda$ $(1 \leqq i \leqq N)$.

PROOF:

$$S_i \lambda(E_i) \geqq p^i\left(\left\{\frac{dp^i}{d\lambda} \leqq S_i\right\} \cap E_i\right) \geqq p^i(E_i) - p^i\left\{\frac{dp^i}{d\lambda} > S_i\right\} > \varepsilon - p^i\left\{\frac{dp^i}{d\lambda} > S_i\right\}.$$

Hence

$$\frac{1}{N} \geqq \frac{1}{N} \sum_{i=1}^{N} \lambda(E_i) > \frac{1}{N} \sum_{i=1}^{N} S_i^{-1}\left[\varepsilon - p^i\left\{\frac{dp^i}{d\lambda} > S_i\right\}\right].$$

It is useful for abbreviations later on to have the following definition introducing a substitute for simultaneous probabilities:

DEFINITION 2. 3 (Simultaneous probability $\tilde{p}$ for M):
Let (X, F, M) be given and let $M' = \{p^{(j)}\}$ $1 \leqq j \leqq n$ be any finite subset of M (n arbitrary), furthermore, let $a^{(j)} \geqq 0$ $(1 \leqq j \leqq n)$ be a real numbers with $\sum_{j=1}^{n} a^{(j)} = 1$.

The probability $\tilde{p}$ defined on $(1, \ldots, n) \times X$ by $\tilde{p}(j \times E) := a^{(j)} p^{(j)}(E)$ $(E \in F)$ is called simultaneous probability of M (determined by $\{a^{(j)}\}$ and $\{p^{(j)}\}$). We denote the front-marginal probability of $\tilde{p}$ on $(1, \ldots, n)$ by a $(a\{j\} = a^{(j)})$ and the back-marginal of $\tilde{p}$ on F by $q\left(q = \sum_{j=1}^{n} a^{(j)} p^{(j)}\right)$. $\tilde{f}$ denotes the Radon-Nikodym density $\tilde{f} := \frac{d\tilde{p}}{da \times dq}$. The set of all simultaneous probabilities derived in this manner from M is denoted by $\tilde{M}$.

It should be remarked that $0 \leqq \tilde{f} \leqq \max\left\{\frac{1}{a^{(j)}} : a^{(j)} > 0\right\}$ holds $\tilde{p}$-a.e. and $a \times q$-a.e.

In particular it follows from lemma 2. 2 and lemma 2. 3:

(1) a) $$N(M, \varepsilon) \geqq S \cdot \sup_{\tilde{p} \in \tilde{M}} (\tilde{p}\{\tilde{f} > S\} - \varepsilon) \quad (S \geqq 0),$$

b) $$\frac{1}{N(M, \varepsilon)} \geqq \frac{1}{S} \inf_{\tilde{p} \in \tilde{M}} (\varepsilon - \tilde{p}\{\tilde{f} > S\}) \quad (S > 0).$$

$\left(\text{For b) set in lemma 2.2 } S_1, \ldots, S_N = S,\ \lambda = \frac{1}{N} \sum_{i=1}^{N} p^i\right)$.

This implies:

(2) $N(M, \cdot)$ is a finite function for $0 < \varepsilon < 1$ iff

$$\sup_{\tilde{p} \in \tilde{M}} \tilde{p}\{\tilde{f} > S\} \to 0 \qquad (S \to \infty).$$

(One may interpret (1) as a result on distribution functions, f.i. one has: $\sup_{\tilde{p} \in \tilde{M}} \tilde{p}\{\tilde{f} > S\} \to 0$ $(S \to \infty)$ iff the set $\{H_{\tilde{p}} : \tilde{p} \in \tilde{M}\}$ of distribution functions $H_{\tilde{p}}(y) := \tilde{p}\{\tilde{f} < y\}$ is conditionally compact with respect to the Levy-metric.)

Furthermore, one similarly obtains such sharper results as:

(3) a) $\int_0^1 d\varepsilon \left(\ln N(M, \varepsilon)\right)^d < \infty$ implies $\sup_{\tilde{p} \in \tilde{M}} \int d\tilde{p} \, (\ln^+ \tilde{f})^d < \infty \quad (d \geqq 1)$,

b) $\sup_{\tilde{p} \in \tilde{M}} \int d\tilde{p} \, (\ln^+ \tilde{f})^d < \infty$ implies $\left(\ln N(M, \varepsilon)\right)^d = O\left(\frac{1}{\varepsilon}\right) (\varepsilon \to 0) \quad (d \geqq 1)$.

(Remark: $o\left(\frac{1}{\varepsilon}\right)$ is i.g. wrong in b) and $\left(\ln N(M, \varepsilon)\right)^d = O\left(\frac{1}{\varepsilon}\right)$ does not imply $\sup_{\tilde{p} \in \tilde{M}} \int d\tilde{p} (\ln^+ \tilde{f})^d < \infty$. For a more complete discussion see [1]). (1), (2), (3) may explain how the finiteness of the estimating parameters which we will use later on is connected with the growth of $N(M_v, \cdot)$ $(\varepsilon \to 0)$ $(1 \leqq v \leqq t)$.

We will apply lemma 2. 1 and lemma 2. 2 in the following way for $M_{[1,t]} = M_1 \times \cdots \times M_t$: Let be given for every M_v $(1 \leqq v \leqq t)$ a probability $\tilde{p}_v \in \tilde{M}_v$ with $\tilde{f}_v$ according to definition 2. 3 and let A_v be a finite set for $\tilde{p}_v$ where the front-marginal of $\tilde{p}_v$ is defined. The probability on $\prod_{v=1}^{t \times} (A_v \times X_v)$ and $\left(\prod_{v=1}^{t \times} A_v\right) \times \left(\prod_{v=1}^{t \times} X_v\right)$, respectively, which is given by $\tilde{p}_{[1,t]} := \tilde{p}_1 \times \cdots \times \tilde{p}_t$ satisfies $\tilde{p}_{[1,t]} \in \tilde{M}_{[1,t]}$ and $\tilde{f}_{[1,t]}$ is the product of the $\tilde{p}_{[1,t]}$-independent random variables $\tilde{f}_v$. Denoting $\{\tilde{p}_1 \times \cdots \times \tilde{p}_t : : \tilde{p}_v \in \tilde{M}_v\}$ by $\tilde{M}_1 \times \cdots \times \tilde{M}_t$ one obtains as a corollary of Lemma 2. 1:

LEMMA 2. 4. $\tilde{M}_1 \times \cdots \times \tilde{M}_t \subseteq \tilde{M}_{[1,t]}$; *and for every real* S *and* $\tilde{p}_{[1,t]} = \tilde{p}_1 \times \cdots \times \tilde{p}_t \in \tilde{M}_1 \times \cdots \times \tilde{M}_t$ *holds*

a) $N(M_{[1,t]}, \varepsilon) > S\left(\tilde{p}_{[1,t]} \left\{\prod_{v=1}^{t \times} \tilde{f}_v > S\right\} - \varepsilon\right) \quad (0 < \varepsilon < 1)$,

b) $N(M_{[1,t]}, \varepsilon) > \exp[S]\left(\tilde{p}_{[1,t]} \left\{\sum_{v=1}^{t} \ln \tilde{f}_v > S\right\} - \varepsilon\right) \quad (0 < \varepsilon < 1)$.

We obtain from lemma 2. 2:

LEMMA 2. 5: *Let* $\{(p^i, E_i)\}_{1 \leqq i \leqq N}$ *be an* ε*-code of length* N *for*

$$M_{[1,t]} \, (0 < \varepsilon < 1), \; q := q_1 \times \cdots \times q_t := \left(\frac{1}{N} \sum_{i=1}^{N} p_1^i\right) \times \cdots \times \left(\frac{1}{N} \sum_{i=1}^{N} p_t^i\right)$$

(where p_v^i *is the component of* p^i *in* M_v*). Then*

a) $\frac{1}{N} \geqq \frac{1}{N} \sum_{i=1}^{N} S_i^{-1} \left[\varepsilon - p^i \left\{\frac{dp^i}{dq} > S_i\right\}\right] \quad (S_1, \ldots, S_N > 0)$,

b) $\frac{1}{N} \geqq \frac{1}{N} \sum_{i=1}^{N} \exp[-S_i] \left(\varepsilon - p^i \left\{\sum_{v=1}^{t} \ln \frac{dp_v^i}{dq_v} > S_i\right\}\right) \quad (S_1, \cdots, S_N \text{ real})$.

It will be the main problem from now on to derive estimates for the right hand sides of lemma 2. 4. b) and lemma 2. 5. b) and to represent these estimates in the same form.

§ 3. Coding Theorem and Its Strong Converse

DEFINITION 3. 1. We denote for M:

$$C(M) := \sup_{\tilde{p}\in\tilde{M}} \int d\tilde{p} \ln \tilde{f},$$

$$D(M) := \sup_{\tilde{p}\in\tilde{M}} \left[\int d\tilde{p} \left(\ln \tilde{f} - \int d\tilde{p} \ln \tilde{f}\right)^2\right]^{1/2}.$$

REMARK: For $p \ll \lambda$ holds $\int dp \left|\ln^- \left(\frac{dp}{d\lambda}\right)\right| = \int d\lambda \frac{dp}{d\lambda} \left|\ln^- \left(\frac{dp}{d\lambda}\right)\right| \leqq \frac{1}{e}$. (Let $z(x) = x \ln x$. One has $z(x) \geqq -\frac{1}{e}$ and $z(x) \to 0$ $(x \to 0)$. Similarly

$$\int dp \left(\ln^- \left(\frac{dp}{d\lambda}\right)\right)^2 = \int d\lambda \frac{dp}{d\lambda} \left(\ln^- \left(\frac{dp}{d\lambda}\right)\right)^2 \leqq 4e^{-2}$$

(discuss $x \ln^2 x$ for $0 < x < 1$). With this and from the fact that $\tilde{f}$ $(\geqq 0)$ is a $\tilde{p}$-a. e. uniformly bounded function follows:

The integrals $\int d\tilde{p} \ln^k \tilde{f}$ $(k = 1, 2)$ are finite.

THEOREM 3. 2 (Coding Theorem, see [1]): *Let* $C(M_v) < \infty$ $(1 \leqq v \leqq t)$. *Then for* $0 < \varepsilon < 1$, $0 < b < 1 - \varepsilon$ holds

$$\ln N(M_{[1,t]}, \varepsilon) \geqq \sum_{v=1}^{t} C(M_v) - \left[\frac{1}{1-\varepsilon-b} \sum_{v=1}^{t} D^2(M_v)\right]^{1/2} + \ln b.$$

PROOF: Set in lemma 2.4 b)

$$h := \ln \tilde{f}_{[1,t]} - \int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} = \sum_{v=1}^{t} \left(\ln \tilde{f}_v - \int d\tilde{p}_v \ln \tilde{f}_v\right).$$

Then by Chebyshev's inequality:

$$\tilde{p}_{[1,t]} \left\{|h| > \left[\frac{1}{1-\varepsilon-b} \int d\tilde{p}_{[1,t]} h^2\right]^{1/2}\right\} < 1 - \varepsilon - b,$$

hence

$$\tilde{p}_{[1,t]} \left\{\ln \tilde{f}_{[1,t]} > \int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} - \left(\frac{1}{1-\varepsilon-b} \int d\tilde{p}_{[1,t]} h^2\right)^{1/2}\right\} > \varepsilon + b$$

and therefore

$$\ln N(M_{[1,t]}, \varepsilon) \geqq \sup_{\tilde{p}_{[1,t]} \in \tilde{M}_1 \times \cdots \times \tilde{M}_t} \left(\int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} - \left[\frac{1}{1-\varepsilon-b} \sum_{v=1}^{t} D^2(M_v)\right]^{1/2}\right) +$$

$$+ \ln(\varepsilon + b - \varepsilon) \geqq \sum_{v=1}^{t} C(M_v) - \left[\frac{1}{1-\varepsilon-b} \sum_{v=1}^{t} D^2(M_v)\right]^{1/2} + \ln b.$$

THEOREM 3. 3 (Strong Converse of the Coding Theorem, see [1]):
Let $\{(p^i, E_i)_{1 \leqq i \leqq N}$ *be any* ε*-code of length* N *for* $M_{[1,t]}$ $(0 < \varepsilon < 1)$. *Then for* $0 < b < 1$ holds:

$$\ln N \leqq \sum_{v=1}^{t} \left(\frac{1}{N} \sum_{i=1}^{N} \int dp_v^i \ln \frac{dp_v^i}{dq_v} \right) + \left[\frac{1}{\varepsilon(1-b)} \sum_{v=1}^{t} \left(\frac{1}{N} \sum \int dp_v^i \ln^2 \left(\frac{dp_v^i}{dq_v} \right) - \right. \right.$$

$$\left. \left. - \left(\frac{1}{N} \sum_{i=1}^{N} \int dp_v^i \ln \frac{dp_v^i}{dq_v} \right)^2 \right) \right]^{1/2} - \ln(\varepsilon \cdot b),$$

where

$$q_v = \frac{1}{N} \sum_{i=1}^{N} p_v^i.$$

The right hand side of this inequality is

$$\leqq \sum_{v=1}^{t} C(M_v) + \left[\frac{1}{\varepsilon(1-b)} \sum_{v=1}^{t} D^2(M_v) \right]^{1/2} - \ln(\varepsilon \cdot b).$$

PROOF: With the notation of lemma 2. 5 and putting

$$S_i = \int dp^i \ln \frac{dp^i}{dq} + \left[\frac{1}{\varepsilon(1-b)} \int dp^i \left(\ln \frac{dp^i}{dq} - \int dp^i \ln \frac{dp^i}{dq} \right)^2 \right]^{1/2}$$

one obtains from lemma 2. 5:

$$\frac{1}{N} \geqq \frac{1}{N} \sum_{i=1}^{N} \exp \left[- \int dp^i \ln \frac{dp^i}{dq} - \right.$$

$$\left. - \left[\frac{1}{\varepsilon(1-b)} \int dp^i \left(\ln \frac{dp^i}{dq} - \int dp^i \ln \frac{dp^i}{dq} \right)^2 \right]^{1/2} \right] (\varepsilon - \varepsilon(1-b)) \geqq$$

$$\geqq \exp \left[- \frac{1}{N} \sum_{i=1}^{N} \int dp^i \ln \frac{dp^i}{dq} - \right.$$

$$\left. - \left[\frac{1}{\varepsilon(1-b)} \frac{1}{N} \sum_{i=1}^{N} \int dp^i \left(\ln \frac{dp^i}{dq} - \int dp^i \ln \frac{dp^i}{dq} \right)^2 \right]^{1/2} + \ln(\varepsilon b) \right].$$

The last inequality follows from Jensen's inequality for the convex functions $\exp[y]$ and $-\sqrt{y}$, respectively. But

$$\frac{1}{N} \sum_{i=1}^{N} \int dp^i \ln \frac{dp^i}{dq} = \sum_{v=1}^{t} \left(\frac{1}{N} \sum_{i=1}^{N} \int dp_v^i \ln \frac{dp_v^i}{dq_v} \right)$$

and

$$\frac{1}{N} \sum_{i=1}^{N} \int dp^i \left(\ln \frac{dp^i}{dq} - \int dp^i \ln \frac{dp^i}{dq} \right)^2 = \sum_{v=1}^{t} \frac{1}{N} \sum_{i=1}^{N} \left(\int dp_v^i \ln^2 \frac{dp_v^i}{dq_v} - \left(\int dp_v^i \ln \frac{dp_v^i}{dq_v} \right)^2 \right) \leqq$$

$$\leqq \sum_{v=1}^{t} \left[\frac{1}{N} \sum \int dp_v^i \ln^2 \frac{dp_v^i}{dq_v} - \left(\frac{1}{N} \sum_{i=1}^{N} \int dp_v^i \ln \frac{dp_v^i}{dq_v} \right)^2 \right].$$

The right hand sides can be expressed by means of simultaneous probabilities (because of the special choice of q) and one obtains the theorem.

COROLLARY 3. 4 (of theorem 3. 2 and theorem 3. 3). Let $C(M_v) < \infty$ $(1 \leqq v \leqq t)$ and let $\{(p^i, E_i)\}_{1 \leqq i \leqq N}$ be any ε-code for $M_{[1,t]}$ of maximal length,

$$q := q_1 \times \cdots \times q_t := \left(\frac{1}{N}\sum_{i=1}^{N} p_1^i\right) \times \cdots \times \left(\frac{1}{N}\sum_{i=1}^{N} p_t^i\right).$$

Then

$$0 \leqq \sum_{v=1}^{t} C(M_v) - \frac{1}{N}\sum_{i=1}^{N} \int dp^i \ln \frac{dp^i}{dq} \leqq K_1(\varepsilon)\left[\sum_{v=1}^{t} D^2(M_v)\right]^{1/2} + K_2(\varepsilon)$$

where $K_1(\varepsilon)$, $K_2(\varepsilon)$ are finite and depend only on ε. This corollary will be used in § 5 in order to derive tighter upper bounds on $N(M_{[1,t]}, \varepsilon)$ for $\frac{1}{2} < \varepsilon < 1$.

First we are going to improve the lower bound of theorem 3. 2 and, in case that $0 < \varepsilon \leqq \frac{1}{2}$, the upper bound of theorem 3. 3. This can be done by means of BERRY's inequality (see f.i. [5]) which says: For independent random variables $r_1, \ldots, r_t$ with expectation values $\int d\mathsf{P} r_v = 0$ and finite third absolute moments holds

$$\left|\mathsf{P}\left\{\sum_{v=1}^{t} r_v < yD_t\right\} - \Phi_{(y)}\right| < C\frac{R_t^3}{D_t^3}, \tag{4}$$

where

$$D_t := \left[\sum_{v=1}^{t}\int d\mathsf{P} r_v^2\right]^{1/2}, \quad R_t := \left[\sum_{v=1}^{t}\int d\mathsf{P}|r_v|^3\right]^{1/3},$$

$\Phi_{(y)}$ is the standard normal distribution, C is an absolute constant. (It is known that $C = 1{\cdot}322$ is a possible choice for C).

Let the inverse function of Φ be Ψ and Ψ' its derivative. For $\mathrm{A} := \sup\left(\Psi'(\varepsilon), \Psi'\left(\frac{1-\varepsilon}{2}\right)\right)$ and any $\delta\left(0 \leqq \delta \leqq \frac{1-\varepsilon}{2}\right)$ holds $\Psi(\varepsilon+\delta) \leqq \Psi(\varepsilon) + A\delta$.

Thus, (4) implies:

$$\mathsf{P}\left\{\sum_{v=1}^{t} r_v > -\Psi_{(\varepsilon)} D_t - A\delta D_t\right\} > \varepsilon + \delta - 1{,}5\frac{R_t^3}{D_t^3} \tag{5}$$

for $0 \leqq \delta \leqq \frac{1-\varepsilon}{2}$.

1. Assume $2\frac{R_t^3}{D_t^3} \leqq \frac{1-\varepsilon}{2}$ and put $\delta = 2\frac{R_t^3}{D_t^3}$.

With (5) follows:

$$\mathsf{P}\left\{\sum_{v=1}^{t} r_v > -\Psi_{(\varepsilon)} D_t - 2A\left(\frac{1-\varepsilon}{4}\right)^{2/3} R_t\right\} \geqq \tag{6}$$

$$\geqq \mathsf{P}\left\{\sum_{v=1}^{t} r_v > -\Psi_{(\varepsilon)} D_t - 2A\frac{R_t^3}{D_t^3} D_t\right\} > e + \frac{1}{2}\frac{R_t^3}{D_t^3} \geqq \varepsilon + \frac{1}{2\sqrt{t}}.$$

$$\left(\text{One has } \frac{R_t^3}{D_t^3} \geqq \frac{1}{\sqrt{t}} \text{ because } \left[\frac{1}{t}\sum_{v=1}^{t}\int d\mathsf{P} r_v^2\right]^{1/2} \leqq \left[\frac{1}{t}\sum_{v=1}^{t}\int d\mathsf{P}|r_v|^3\right]^{1/3}\right).$$

2. Assume $2\frac{R_t^3}{D_t^3} > \frac{1-\varepsilon}{2}$. Then

$$(7)\qquad \mathsf{P}\left\{\sum_{v=1}^{t} r_v > -\Psi_{(\varepsilon)} D_t - \left(|\Psi_{(\varepsilon)}| + \left(\frac{2}{1-\varepsilon}\right)^{1/2}\right)\frac{4}{1-\varepsilon}\,\frac{R_t^3}{D_t^3}\,D_t\right\} \geqq$$

$$\geqq \mathsf{P}\left\{\sum_{v=1}^{t} r_v > -\Psi_{(\varepsilon)} D_t - \left(|\Psi_{(\varepsilon)}| + \left(\frac{2}{1-\varepsilon}\right)^{1/2}\right)\left(\frac{4}{1-\varepsilon}\right)^{1/3} R_t\right\} \geqq$$

$$\geqq \mathsf{P}\left\{\sum_{v=1}^{t} r_v > -\left(\frac{2}{1-\varepsilon}\right)^{1/2} D_t\right\} > \varepsilon + \frac{1-\varepsilon}{2}$$

(Chebyshev's inequality). Both cases together give

$$(8)\qquad \text{a)}\quad \mathsf{P}\left\{\sum_{v=1}^{t} r_v > -\Psi_{(\varepsilon)} D_t - K_1(\varepsilon) R_t\right\} > \varepsilon + \min\left(\frac{1-\varepsilon}{2}, \frac{1}{2}\, t^{-1/2}\right),$$

$$\text{b)}\quad \mathsf{P}\left\{\sum_{v=1}^{t} r_v > -\Psi_{(\varepsilon)} D_t - K_2(\varepsilon)\,\frac{R_t^3}{D_t^3}\,D_t\right\} > \varepsilon + \min\left(\frac{1-\varepsilon}{2}, \frac{1}{2}\, t^{-1/2}\right),$$

(where $K_1(\varepsilon), K_2(\varepsilon) \geqq 0$ are (finite) functions of ε). Replacing r_v by $-r_v$ and ε by $1-\varepsilon$ in (8) one similarly obtains:

$$(9)\qquad \text{a)}\quad \mathsf{P}\left\{\sum_{v=1}^{t} r_v > -\Psi_{(\varepsilon)} D_t + K_1(1-\varepsilon) R_t\right\} < \varepsilon - \min\left(\frac{\varepsilon}{2}, \frac{1}{2}\, t^{-1/2}\right),$$

$$\text{b)}\quad \mathsf{P}\left\{\sum_{v=1}^{t} r_v > -\Psi_{(\varepsilon)} D_t + K_2(1-\varepsilon)\,\frac{R_t^3}{D_t^3}\right\} < \varepsilon - \min\left(\frac{\varepsilon}{2}, \frac{1}{2}\, t^{-1/2}\right).$$

We use (8) and (9) instead of Chebyshev's inequality for stronger estimates on $N(M_{[1,t]}, \varepsilon)$.

It directly follows:

THEOREM 3.5. *Let be given a sequence* $\{M_v\}$ $(v = 1, 2, \ldots)$ *with*

$$\sup_v \sup_{\tilde{p}_v \in \tilde{M}_v} \int d\tilde{p} \int d\tilde{p}_v (\ln \tilde{f})^3 < \infty.$$

Then:

a) $\ln N(M_{[1,t]}, \varepsilon) \geqq$

$$\geqq \sup_{\tilde{p}_{[1,t]}} \left(\int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} - \Psi_{(\varepsilon)}\left[\int d\tilde{p}_{[1,t]}\left(\ln \tilde{f}_{[1,t]} - \int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]}\right)^2\right]^{1/2}\right) - O(t^{1/3})$$

for every fixed ε $(0 < \varepsilon < 1)$; $(O(t^{1/3})$ *depends on* ε,

$$\tilde{p}_{[1,t]} = \tilde{p}_1 \times \cdots \times \tilde{p}_t \in \tilde{M}_1 \times \cdots \times \tilde{M}_t).$$

b) $\ln N(M_{[1,t]}, \varepsilon) \leqq$

$$\leqq \sup_{\tilde{P}_{[1,t]}} \left(\int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} - \Psi_{(\varepsilon)} \left[\int d\tilde{p}_{[1,t]} \left(\ln \tilde{f}_{[1,t]} - \int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} \right)^2 \right]^{1/2} \right) + O(t^{1/3})$$

for every fixed ε $(0 < \varepsilon \leqq 1/2)$; $(O(t^{1/3})$ *depends on* ε,

$$\tilde{p}_{[1,t]} = \tilde{p}_1 \times \cdots \times \tilde{p}_t \in \tilde{M}_1 \times \cdots \times \tilde{M}_t).$$

PROOF: a) Set $\mathsf{P} = \tilde{p}_{[1,t]}$ and $r_v = \ln \tilde{f}_v - \int d\tilde{p}_v \ln \tilde{f}_v$ in (8) a) and use lemma 2. 4 b).

b) Set $\mathsf{P} = p^i$ and $r_v = \ln \dfrac{dp_v^i}{dq_v} - \int dp_v^i \ln \dfrac{dp_v^i}{dq_v}$

in (9) a) $\left(\text{where } q_v = \dfrac{1}{N} \sum_{i=1}^{N} p_v^i \right)$ and apply lemma 2. 5 b). One uses Jensen's inequality exactly the same way as for theorem 3. 3.

REMARK: $\sup_v \sup_{\tilde{p}_v \in \tilde{M}_v} \int d\tilde{p}_v (\ln \tilde{f}_v)^3 < \infty$ iff $\sup_v \sup_{\tilde{p}_v \in \tilde{M}_v} \int d\tilde{p}_v |\ln \tilde{f}_v|^3 < \infty$ follows with a similar conclusion as at the beginning of § 3. Using a method of small perturbations of measures, one obtains together with (8) b) and (9) b) the

THEOREM 3. 6: *Let be given a sequence* $\{M_v\}$ $(v = 1, 2, \ldots)$, *where for every* M_v *exists a probability* λ_v *with* $\dfrac{dp_v}{d\lambda_v} \leqq K$ *for all* $p_v \in M_v$ $(v = 1, 2, \ldots)$. *Then*:

a) $\ln N(M_{[1,t]}, \varepsilon) \geqq$

$$\geqq \sup_{\tilde{P}_{[1,t]}} \left(\int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} - \Psi_{(\varepsilon)} \left[\int d\tilde{p}_{[1,t]} \left(\ln \tilde{f}_{[1,t]} - \int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} \right)^2 \right]^{1/2} \right)$$

$- O(\ln t)$ *for every fixed* ε $(0 < \varepsilon < 1)$; $(O(\ln t)$ *depends on* ε, $\tilde{p}_{[1,t]} = \tilde{p}_1 \times \cdots$ $\cdots \times \tilde{p}_t \in \tilde{M}_1 \times \cdots \times \tilde{M}_t)$.

b) $\ln N(M_{[1,t]}, \varepsilon) \leqq$

$$\leqq \sup_{\tilde{P}_{[1,t]}} \left(\int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} - \Psi_{(\varepsilon)} \left[\int d\tilde{p}_{[1,t]} \left(\ln \tilde{f}_{[1,t]} - \int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} \right)^2 \right]^{1/2} \right)$$

$+ O(\ln t)$ *for every* ε $(0 < \varepsilon < 1/2)$; $(O(\ln t)$ *depends on* ε, $\tilde{p}_{[1,t]} = \tilde{p}_1 \times \cdots$ $\cdots \times \tilde{p}_t \in \tilde{M}_1 \times \cdots \times \tilde{M}_t)$.

PROOF:

$$\text{Let } \overline{M}_v := \{\bar{p}_v : = (1 - t^{-2}) p_v + t^{-2} \lambda_v : p_v \in M_v\},$$

$\overline{M}_{[1,t]} := \overline{M}_1 \times \cdots \times \overline{M}_t$, $\tilde{\bar{p}}_v$ the notation for the simultaneous probability with respect to $\overline{M}_v$, $\tilde{g}_v$ the simultaneous density of

$$\tilde{\bar{p}}_v, \tilde{\bar{p}}_{[1,t]} := \tilde{\bar{p}}_1 \times \cdots \times \tilde{\bar{p}}_t, \tilde{g}_{[1,t]} := \tilde{g}_1 \cdots \tilde{g}_t.$$

$$\bar{p}^i = (p_1^i (1 - t^{-2}) + \lambda_1 t^{-2}) \times \cdots \times (p_t^i (1 - t^{-2}) + t^{-2} \lambda_t) \text{ in } \overline{M}_{[1,t]}$$

corresponds to $p^i = p_1^i \times \cdots \times p_t^i$ in $M_{[1,t]}$. If $\{(\bar{p}^i E_i)\}_{1 \leqq i \leqq N}$ is an ε-code for $\overline{M}_{[1,t]}$ then $(1-t^{-2})^t p^i(E_i) + \frac{1}{t} > \varepsilon$ and therefore $\{(p^i, E_i)\}_{1 \leqq i \leqq N}$ is an $\varepsilon\left(1-\frac{1}{\varepsilon t}\right)$-code for $M_{[1,t]}$ $(t > t_0)$. If $\{(p^i, E_i)\}_{1 \leqq i \leqq N}$ then $\varepsilon < p^i(E_i) \leqq (1-t^{-2})^{-t} \bar{p}^i(E_i)$ and therefore $\{(\bar{p}^i, E_i)\}_{1 \leqq i \leqq N}$ is an $\varepsilon\left(1-\frac{1}{\varepsilon t}\right)$-code for $\overline{M}_{[1,t]}$ $(t > t_0)$. We have:

$$(10) \qquad N\left(\overline{M}_{[1,t]}, \varepsilon\left(1+\frac{1}{\varepsilon t}\right)\right) \leqq N(M_{[1,t]}, \varepsilon) \leqq N\left(\overline{M}_{[1,t]}, \varepsilon\left(1-\frac{2}{t}\right)\right) \quad (t > t_0).$$

Observe now that

$$\int d\tilde{\bar{p}}_v \left|\ln \tilde{g}_v - \int d\tilde{\bar{p}} \ln \tilde{g}_v\right|^3 \leqq \int d\tilde{\bar{p}}_v \left(\ln \tilde{g}_v - \int d\tilde{\bar{p}} \ln \tilde{g}_v\right)^2 \cdot 2 \cdot (\text{a.e. sup} \left|\ln \tilde{g}_v\right|)$$

and that

$$t^2 K \geqq \frac{t^2 \, d\tilde{\bar{p}}_v}{da_v \times d\lambda_v} \geqq \tilde{g}_v \geqq \frac{t^{-2} \, d\lambda_v}{K d\lambda_v}.$$

Hence $|\ln \tilde{g}_v| \leqq \ln (t^2 K)$.

Now use (8) b) and (9) b) for $\mathsf{P} = \tilde{\bar{p}}_{[1,t]}$ and

$$\sum r_v = \ln \tilde{g}_{[1,t]} - \int d\tilde{\bar{p}}_{[1,t]} \ln \tilde{g}_{[1,t]}.$$

Here the last inequalities give an estimate $\frac{R_t^3}{D_t^3} \leqq \frac{\ln (t^2 K)}{D_t}$.

Therefore we obtain:

$$(11) \qquad \text{a)} \quad \ln N\left(\overline{M}_{[1,t]}, \varepsilon\left(1+\frac{1}{\varepsilon t}\right)\right) \geqq$$

$$\geqq \sup_{\tilde{\bar{p}}_{[1,t]}} \left(\int d\tilde{\bar{p}}_{[1,t]} \ln \tilde{g}_{[1,t]} - \Psi\left(\varepsilon\left(1+\frac{1}{\varepsilon t}\right)\right)\left[\int d\tilde{\bar{p}}_{[1,t]} \left(\ln \tilde{g}_{[1,t]} - \int d\tilde{\bar{p}}_{[1,t]} \ln \tilde{g}_{[1,t]}\right)^2\right]^{1/2}\right) -$$

$$- O(\ln t) \quad (t > t_0(\varepsilon), 0 < \varepsilon < 1),$$

$$\text{b)} \quad \ln N\left(\overline{M}_{[1,t]}, \varepsilon\left(1-\frac{2}{t}\right)\right) \leqq$$

$$\leqq \sup_{\tilde{\bar{p}}_{[1,t]}} \left(\int d\tilde{\bar{p}}_{[1,t]} \ln \tilde{g}_{[1,t]} - \Psi\left(\varepsilon\left(1-\frac{2}{t}\right)\right)\left[\int d\tilde{\bar{p}}_{[1,t]} \left(\ln \tilde{g}_{[1,t]} - \int d\tilde{\bar{p}}_{[1,t]} \ln \tilde{g}_{[1,t]}\right)^2\right]^{1/2}\right) +$$

$$+ O(\ln t) \quad \left(t > t_0(\varepsilon), 0 < \varepsilon < \frac{1}{2}\right).$$

We have to replace $\tilde{\bar{p}}, \tilde{g}$ in these estimates by $\tilde{p}, \tilde{f}$. The following lemmas will take care of this and complete the proof.

LEMMA 3.7: *In the situation of theorem* 3.6 *holds*

a) $\int d\tilde{p}_v \ln \tilde{f}_v \leqq \ln K \quad (\tilde{p}_v \in \tilde{M}_v)$,

b) $\int d\tilde{p}_v \ln^2 \tilde{f}_v \leqq 4(\ln^2 K + 4e^{-2})$.

PROOF:

$$\text{a)} \quad \int d\tilde{p}_v \ln \tilde{f}_v = \sum_{j=1}^{n} a^{(j)} \int dp_v^{(j)} \ln \frac{dp_v^{(j)}}{d\left(\sum_{k=1}^{n} a^{(k)} p_v^{(k)}\right)} =$$

$$= \sum_{j=1}^{n} a^{(j)} \int dp_v^{(j)} \ln \frac{dp_v^{(j)}}{d\lambda_v} - \int d\left(\sum a^{(k)} p_v^{(k)}\right) \ln \frac{d\left(\sum a(k) p_v(k)\right)}{d\lambda_v} \leqq$$

$$\leqq \sum a^{(j)} \int dp_v^{(j)} \ln \frac{dp_v^{(j)}}{d\lambda_v} \leqq \ln K.$$

$$\text{b)} \quad \int d\tilde{p}_v \ln^2 \tilde{f}_v \leqq 2 \int d\tilde{p}_v \ln^2 \frac{d\tilde{p}_v}{da_v \times d\lambda_v} + 2 \int dq_v \ln^2 \frac{dq_v}{d\lambda_v} \leqq 4(\ln^2 K + 4e^{-2}),$$

(see the beginning of § 3).

LEMMA 3. 8: *With the notation of theorem* 3. 6 *holds for* $1 \leqq v \leqq t$, $t \geqq 2$:

$$\left| \int d\tilde{p}_v \ln \tilde{f}_v - \int d\bar{p} \ln \tilde{g}_v \right| \leqq c \cdot \frac{1}{t}$$

with c depending only on K.

PROOF:

$$\int d\tilde{\bar{p}}_v \ln \tilde{g}_v = \sum_{j=1}^{n} a^{(j)} \int d\left(p_v^{(j)}(1-t^{-2}) + \lambda_v t^{-2}\right) \ln \left(\frac{d\left(p_v^{(j)}(1-t^2) + \lambda_v t^{-2}\right)}{d\left(q_v(1-t^{-2}) + \lambda_v t^{-2}\right)} \right) \leqq$$

$$\leqq (1-t^{-2}) \sum_{j=1}^{n} a^{(j)} \int dp_v^{(j)} \ln \frac{dp_v^{(j)}(1-t^{-2})}{dq_v(1-t^{-2})} + t^{-2} \int d\lambda_v \ln \frac{d\lambda_v t^{-2}}{d\lambda_v t^{-2}} =$$

$$= (1-t^{-2}) \int d\tilde{p}_v \ln \tilde{f}_v$$

because $\sum_{i=1}^{m} a^i \ln \frac{a^i}{b^i} \geqq \left(\sum_{i=1}^{m} a^i\right) \ln \frac{\left(\sum_{i=1}^{m} a^i\right)}{\left(\sum_{i=1}^{m} b^i\right)}$ for $a^i, b^i \geqq 0$ arbitrary.

This well known inequality can be checked by induction. On the other hand

$$\int d\tilde{\bar{p}} \ln \tilde{g}_v = (1-t^{-2}) \int d\tilde{p}_v \ln \tilde{g}_v + t^{-2} \int da_v \times d\lambda_v \ln \tilde{g}_v \geqq$$

$$\geqq (1-t^{-2}) \int d\tilde{p}_v \ln \frac{d\tilde{p}_v (1-t^{-2})}{da_v \times d\left(q_v(1-t^{-2}) + \lambda_v t^{-2}\right)} +$$

$$+ t^{-2} \int d\lambda_v \ln \frac{t^{-2} d\lambda_v}{d\left(q_v(1-t^{-2}) + \lambda_v t^{-2}\right)} =$$

$$= z(1-t^{-2}) + (1-t^{-2}) \int d\tilde{p} \ln \frac{d\tilde{p}_v}{da_v \times d(q_v(1-t^{-2}) + \lambda_v t^{-2})} + z(t^{-2}) +$$

$$+ t^{-2} \int d\lambda_v \ln \frac{d\lambda_v}{d(q_v(1-t^{-2}) + \lambda_v t^{-2})} \geqq z(1-t^{-2}) + z(t^{-2}) + (1-t^{-2}) \int d\tilde{p}_v \ln \tilde{f}_v$$

(where $z(x) = x \ln x$), because the integral with the factor t^{-2} is non-negative and

$$\int d\tilde{p}_v \ln \frac{d\tilde{p}_v}{da_v \times d(q_v(1-t^{-2}) + \lambda_v t^{-2})} \geqq \int d\tilde{p}_v \ln \tilde{f}_v,$$

which follows similarly as lemma 3. 7 a).

$$0 \leqq -z(1-t^{-2}) - z(t^{-2}) \leqq c_1 t^{-2} + \frac{1}{t}\left(\frac{2}{e}\right)$$

for $t \geqq 2$ and some constant c_1. $\frac{1}{t^2} \int d\tilde{p}_v \ln \tilde{f}_v \leqq \frac{1}{t^2} \ln K$. This together completes the proof.

LEMMA 3. 9: *With the notation of theorem* 3. 6 *holds for* $1 \leqq v \leqq t$, $t \geqq 2$:

$$\left| \int d\tilde{\tilde{p}}_v \ln^2 \tilde{g}_v - \int d\tilde{p}_v \ln^2 \tilde{f}_v \right| \leqq c_2 \frac{1}{t}$$

where c_2 *depends only on* K.

PROOF: Let a_v be the front marginal probability of $\tilde{p}_v$ and $\tilde{\tilde{p}}_v$, respectively. Furthermore let q_v and q_v' be the back marginal probabilities of $\tilde{p}_v$ and $\tilde{\tilde{p}}$, respectively. $q_v' = q_v(1-t^{-2}) + \lambda_v t^{-2}$. Finally let $h(y) := y \ln^2 y + y - 1$. $h(y)$ is a monatonically increasing function.

1.
$$\int d\tilde{\tilde{p}}_v \ln^2 \tilde{g}_v = \int d(a_v \times q_v') \tilde{g}_v \ln^2 \tilde{g}_v =$$

$$= \int d(a_v \times q_v') h(\tilde{g}_v) \geqq \int d(a_v \times q_v') h\left(\frac{d\tilde{p}_v(1-t^{-2})}{da_v \times dq_v'}\right) =$$

$$= (1-t^{-2}) \int d\tilde{p}_v \ln^2 \left(\frac{d\tilde{p}_v}{da_v \times dq_v'}\right) + 2z(1-t^{-2}) \int d\tilde{p}_v \ln \frac{d\tilde{p}_v}{da_v \times dq_v'} - t^{-2} \geqq$$

$$\geqq (1-t^{-2}) \int d\tilde{p}_v \ln^2 \frac{d\tilde{p}_v}{da_v \times dq_v'} + 2z(1-t^{-2})(\ln K - \ln(1-t^{-2})) - t^{-2},$$

because

$$\int d\tilde{p}_v \ln \frac{d\tilde{p}_v}{da_v \times dq_v'} \leqq \int d\tilde{p}_v \ln \frac{d\tilde{p}_v}{da_v \times dq_v(1-t^{-2})} =$$

$$= \int d\tilde{p}_v \ln \tilde{f}_v - \ln(1-t^{-2}) \leqq \ln K - \ln(1-t^{-2}).$$

We estimate $\int d\tilde{p}_v \ln^2 \frac{d\tilde{p}_v}{da_v \times dq'_v}$ later.

2. $$\int d\tilde{\tilde{p}}_v \ln^2 \tilde{g}_v = (1-t^{-2}) \int d\tilde{p}_v \ln^2 \tilde{g}_v + t^{-2} \int da_v \times d\lambda_v \ln^2 \tilde{g}_v \leqq$$

$$\leqq (1-t^{-2}) \int d\tilde{p}_v \ln^2 \tilde{g}_v + t^{-2} \ln^2 (t^2 K) =$$

$$= (1-t^{-2}) \int d\tilde{p}_v \ln^2 \tilde{g}_v + \frac{K^{1/2}}{t} \left(\frac{4}{tK^{1/2}} \ln^2 \left(\frac{1}{tK^{1/2}} \right) \right).$$

$$\frac{4}{tK^{1/2}} \ln^2 \left(\frac{1}{tK^{1/2}} \right) \leqq 16e^{-2}$$

and

$$\int d\tilde{p}_v \ln^2 \tilde{g}_v \leqq \int d\tilde{p}_v \ln^2 \frac{d\tilde{p}_v}{da_v \times dq'_v} + \int d\tilde{p}_v \ln^2 \frac{d\tilde{p}_v}{d(\tilde{p}_v(1-t^{-2}) + a_v \times \lambda_v t^{-2})} +$$

$$+ 2 \left(\int d\tilde{p}_v \ln^2 \frac{d\tilde{p}_v}{da_v \times dq'_v} \right)^{1/2} \left(\int d\tilde{p}_v \ln^2 \frac{d\tilde{p}_v}{d(\tilde{p}_v(1-t^{-2}) + a_v \times \lambda_v t^{-2})} \right)^{1/2}.$$

$$\int d\tilde{p}_v \ln^2 \frac{d\tilde{p}_v}{d(\tilde{p}_v(1-t^{-2}) + a_v \times \lambda_v t^{-2})} =$$

$$= \int d(\tilde{p}_v(1-t^{-2}) + a_v \times \lambda_v t^{-2}) h \left(\frac{d\tilde{p}_v}{d(\tilde{p}_v(1-t^{-2}) + a_v \times \lambda_v t^{-2})} \right) \leqq$$

$$\leqq h \left(\frac{1}{1-t^{-2}} \right) = O(t^{-2}) \qquad (t^{-2} \to 0).$$

Again, it remains to estimate $\int d\tilde{p}_v \ln^2 \frac{d\tilde{p}_v}{da_v \times dq'_v}$.

3. $$\int \tilde{p}_v \ln^2 \frac{d\tilde{p}_v}{da_v \times dq'_v} = \int d\tilde{p}_v \ln^2 \tilde{f}_v + \int dq_v \ln^2 \frac{dq_v}{dq'_v} + 2 \int d\tilde{p}_v (\ln \tilde{f}_v) \cdot \left(\ln \frac{da_v \times dq_v}{da_v \times dq'_v} \right).$$

$$\left| \int d\tilde{p}_v \ln \tilde{f}_v \ln \frac{da_v \times dq_v}{da_v \times dq'_v} \right| \leqq \left(\int d\tilde{p}_v \ln \tilde{f}_v \right)^{1/2} \cdot \left(\int dq_v \ln^2 \frac{dq_v}{dq'_v} \right)^{1/2}.$$

$$\int d\tilde{p}_v \ln^2 \tilde{f}_v \leqq 2(\ln^2 K + 4e^{-2})$$

and

$$\int dq_v \ln^2 \frac{dq_v}{dq'_v} = \int dq'_v h \left(\frac{dq_v}{dq'_v} \right) \leqq \int dq'_v h \left(\frac{1}{1-t^{-2}} \right) = O(t^{-2}).$$

Hence $\left|\int d\tilde{p}_v \left(\ln^2 \frac{d\tilde{p}_v}{da_v \times dq'_v} - \ln^2 \tilde{f}_v\right)\right| = O\left(\frac{1}{t}\right)$,

$\left(O\left(\frac{1}{t}\right)\right.$ depending on $\left.K\right)$.

Now one only has to put these three parts together.

It follows from lemma 3. 7, 3. 8, 3. 9 that one changes the right hand sides of (11) about a term $O(1)$ if one replaces $\tilde{\tilde{p}}_{[1,t]}$ by $\tilde{p}_{[1,t]}$ and $\tilde{g}_{[1,t]}$ by $\tilde{f}_{[1,t]}$ in (11).

This completes the proof of theorem 3. 6.

§ 4. Properties of the Estimating Parameters

Up to now it has been possible to obtain sharp estimates for the length of codes by means of sharp stochastic inequalities and convexity methods. However, for the upper estimate of $N(M_{[1,t]}, \varepsilon)$ for $\varepsilon \geqq \frac{1}{2}$ we need additional information, because convexity methods cannot be used in the same way as earlier. $(\Psi(\varepsilon) \leqq 0$ for $\varepsilon \leqq \frac{1}{2}$, $\Psi(\varepsilon) > 0$ for $\varepsilon > \frac{1}{2})$. Furthermore, the bounds of § 3 do not show good enough, how $N(M_{[1,t]}, \varepsilon)$ behaves with increasing time t.

LEMMA 4. 1: $C(M_1 \times M_2) = C(M_1) + C(M_2)$.

PROOF:

1. $$\sum_{i=1}^{n} a^{(i)} \int dp_1^{(i)} \ln \frac{dp_1^{(i)}}{dp\left(\sum_{j=1}^{n} a^{(j)} p^{(j)}\right)} + \sum_{k=1}^{m} b^{(k)} \int dp_2^{(k)} \ln \frac{dp_2^{(k)}}{d\left(\sum_{l=1}^{m} b^{(l)} p_2^{(l)}\right)} =$$

$$= \sum_{i,k} a^{(i)} b^{(k)} \int d(p_1^{(i)} \times p_2^{(k)}) \ln \frac{d(p_1^{(i)} \times p_2^{(k)})}{d\left(\sum_{j,l} a^{(j)} b^{(l)} p_1^{(j)} \times p_2^{(l)}\right)} \leqq C(M_1 \times M_2).$$

Hence $C(M_1) + C(M_2) \leqq C(M_1 \times M_2)$.

2. Let $p^{(i)} = p_1^{(i)} \times p_2^{(i)} \in M_1 \times M_2$ be given $(1 \leqq i \leqq n)$.

$$\sum_{i=1}^{n} a^{(i)} \int dp^{(i)} \ln \frac{dp^{(i)}}{d\left(\sum_{j=1}^{n} a^{(j)} p^{(j)}\right)} \leqq \sum_{i=1}^{n} a^{(i)} \int dp^{(i)} \ln \frac{dp^{(i)}}{d\left(\sum a^{(j)} p_1^{(j)}\right) \times d\left(\sum a^{(k)} p_2^{(k)}\right)}$$

similarly follows as in the proof of Lemma 3. 7 a). Splitting the right hand term into a sum

$$\sum_{i=1}^{n} a^{(i)} \int dp_1^{(i)} \ln \frac{dp_1^{(i)}}{d\left(\sum a^{(k)} p_1^{(k)}\right)} + \sum_{i=1}^{n} a^{(i)} \int dp_2^{(i)} \ln \frac{dp_2^{(i)}}{d\left(\sum a^{(k)} p_2^{(k)}\right)}$$

and supremizing the left hand side of the inequality gives $C(M_1 \times M_2) \leqq$ $\leqq C(M_1) + C(M_2)$.

LEMMA 4. 2: *Let* $C(M)<\infty$. *Then there is exactly one probability* μ *for* M *s. t. for every* $\tilde{p}\in\tilde{M}$ *(with back marginal probability* q*)* $\int d\tilde{p}\ln\tilde{f}>C(M)-\delta$ *implies* $\|q-\mu\|\leqq 8\left(\delta+\min\left(\frac{1}{e}, c\delta^{\frac{1}{4}}\right)\right)$ *(where* $c>0$ *is a universal constant,* $\|\cdot\|$ *the total variation norm).*

PROOF:

Let $\{p^{(1)}, \ldots, p^{(n)}\}\subseteqq M$, $\quad a^{(j)}b^{(j)}\geqq 0 \quad (1\leqq j\leqq n)$,

$$\sum_{j=1}^{n} a^{(j)} = \sum_{j=1}^{n} b^{(j)} = 1, \qquad q' := \sum a^{(j)}p^{(j)}, \qquad q'' := \Sigma b^{(j)}p^{(j)},$$

where $\sum a^{(j)}\int dp^{(j)}\ln\frac{dp^{(j)}}{dq'}>C(M)-\delta$

and

$$\sum b^{(j)}\int dp^{(j)}\ln\frac{dp^{(j)}}{pq''}>C(M)-\delta$$

be satisfied.

Then $C(M)-\delta<\frac{1}{2}\sum a^{(j)}\int dp^{(j)}\ln\frac{dp^{(j)}}{dq'}+\frac{1}{2}\sum b^{(j)}\int dp^{(j)}\ln\frac{dp^{(j)}}{dq''}\leqq$

$$\leqq\frac{1}{2}\sum a^{(j)}\int dp^{(j)}\ln\frac{dp^{(j)}}{dq'}+\frac{1}{2}\sum b^{(j)}\int dp^{(j)}\ln\frac{dp^{(j)}}{dq''}+$$

$$+\left(\frac{1}{2}\int dp'\ln\frac{dq'}{\frac{1}{2}d(q'+q'')}+\frac{1}{2}\int dq''\ln\frac{dq''}{\frac{1}{2}d(q'+q'')}\right)=$$

$$=\sum\frac{a^{(j)}+b^{(j)}}{2}\int dp^{(j)}\ln\frac{dp^{(j)}}{\frac{1}{2}d(q'+q'')}\leqq C(M).$$

Therefore

$$0\leqq\frac{1}{2}\int dq'\ln\frac{dq'}{\frac{1}{2}d(q'+q'')}\leqq\delta.$$

One finds the following inequality (which can be derived easily from Chebishev's inequality) in [2]: If p, q are probabilities, $p\ll q$, then

$$\int dp\left|\ln\frac{dp}{dq}\right|\leqq\int dp\ln\frac{dp}{dq}+\min\left(\frac{1}{e}, c\left(\int dp\ln\frac{dp}{dq}\right)^{1/2}\right)$$

(with a universal constant $c>0$).

Hence $\frac{1}{4}\|q'-q''\| = \frac{1}{2}\left\| q'-\left(\frac{1}{2}q'+\frac{1}{2}q''\right)\right\| =$

$$= \int d\left(\frac{1}{2}q'+\frac{1}{2}q''\right)\left(\frac{dq'}{d\left(\frac{1}{2}q'+\frac{1}{2}q''\right)}-1\right)^{+} \leqq$$

$$\leqq \int d\left(\frac{1}{2}q'+\frac{1}{2}q''\right) z^{+}\left(\frac{dq'}{d\left(\frac{1}{2}q'+\frac{1}{2}q''\right)}\right) = \int dq' \ln^{+} \frac{dq'}{d\left(\frac{1}{2}q'+\frac{1}{2}q''\right)}$$

gives

$$\|q'-q''\| < 8\left(\delta + \min\left(\frac{1}{e}, c\delta^{1/2}\right)\right).$$

Then the lemma follows from the norm completeness of the set of all probabilities on the σ-field F.

The next observation is due to SHANNON (see f.i. [3]). It may be formulated as

LEMMA 4.3: *Let* $M = \{p^{(1)}, \ldots, p^{(n)}\}$. *Then there is exactly one probability* $r \in \mathrm{co}(M)$ $(\mathrm{co}(M)$ *the convex hull of* $M)$ *s.t.* $\int dp \ln \frac{dp}{dr} \leqq C(M)$ *for every* $p \in M$. *Moreover* $C(M) = \sup\left\{\int dp \ln \frac{dp}{dr} : p \in M\right\}$.

PROOF:

$C(M)$ is finite $(C(M) \leqq \ln n)$. For some $\tilde{p} \in \tilde{M}$ holds

$$C(M) = \int d\tilde{p} \ln \tilde{f} = \int d\tilde{p} \ln \frac{d\tilde{p}}{da \times dq} = \sum_{j=1}^{n} a^{(j)} \int dp^{(j)} \ln \frac{dp^{(j)}}{d\left(\sum_{k=1}^{n} a^{(k)} p^{(k)}\right)}$$

where q is the required probability r: $C(M) \leqq \sup\left\{\int dp \ln \frac{dp}{dr} : p \in M\right\}$ follows from the last equality. For the same $\tilde{p}$, $0 \leqq b \leqq 1$ and $p \in M$ fixed let $g(b) :=$ $= b \int d\tilde{p} \ln \frac{d\tilde{p}}{da \times d(br+(1-b)p)} + (1-b) \int dp \ln \frac{dp}{d(br+(1-b)p)} \leqq C(M)$. Then $g(b) = bC(M) + b \int dr \ln \frac{dr}{d(br+(1-b)p)} + (1-b) \int dp \ln \frac{dp}{d(br+(1-b)p)}$ and the left derivative at 1 is $g'(1) = C(M) - \int dp \ln \frac{dp}{dr}$. Because $g(b) \leqq C(M)$, $g(1) = C(M)$, one has $g'(1) \geqq 0$. One obtains the uniqueness of r as follows: Let λ be another probability with $\int dp \ln \frac{dp}{d\lambda} \leqq C(M)$ for all $p \in M$. Then for the $\tilde{p}$ above holds

$$\int d\tilde{p} \ln \frac{d\tilde{p}}{da \times d\lambda} = C(M) = \int d\tilde{p} \ln \frac{d\tilde{p}}{da \times dr} = \int d\tilde{p} \ln \frac{d\tilde{p}}{da \times d\lambda} + \int dr \ln \frac{dr}{d\lambda}.$$

But $\int dr \ln \frac{dr}{d\lambda} > 0$ for every $\lambda \neq r$ because of the strict concavity of $\ln y$.

LEMMA 4. 4: *For arbitrary M with $C(M) < \infty$ there is exactly one probability μ s.t.*

$$\int dp \ln \frac{dp}{d\mu} \leqq C(M) \text{ for all } p \in M. \text{ For this } \mu \text{ holds } C(M) =$$

$$= \sup \left\{ \int dp \ln \frac{dp}{d\mu} : p \in M \right\}. \ \mu \text{ is identical with the probability } \mu \text{ of lemma 4.2.}$$

We sketch a PROOF:

Let $\{M^k\}_{k=1,2,\ldots}$ be a sequence of finite subsets of M with $M^k \subseteq M^{k+1}$ and $C(M^k) \xrightarrow{k \to \infty} C(M)$. Furthermore let r_k be the probability for M^k which satisfies $\int dp \ln \frac{dp}{dr_k} \leqq C(M^k)$ for all $p \in M^k$. $\|r_k - \mu\| \to 0 \ (k \to \infty)$ (μ according to lemma 4. 2). One has for $p \in M'$ fixed and $s > 0$ fixed

$$0 \leqq -\frac{1}{s} \ln \int dp \left(\frac{dp}{dr_k} \right)^{-s} \leqq \int dp \ln \frac{dp}{dr_k} \leqq C(M^k) \leqq C(M) \text{ for all } k.$$

(Differentiation with respect to s shows that

$$-\ln \int dp \left(\frac{dp}{dr_k} \right)^{-s} \text{ is a concave function of } s, \text{ and}$$

$$\lim_{s \to 0} -\frac{1}{s} \ln \int dp \left(\frac{dp}{dr_k} \right)^{-s} = \int dp \ln \frac{dp}{dr_k}.)$$

$$\int d\lambda_k \left(\frac{dp}{d\lambda_k} \right)^{1-s} \left(\frac{dr_k}{d\lambda_k} \right)^{s} = \int dp \left(\frac{dp}{dr_k} \right)^{-s} \xrightarrow{k \to \infty} \int dp \left(\frac{dp}{d\mu} \right)^{-s} : =$$

$$= \int d\lambda \left(\frac{dp}{d\lambda} \right)^{1-s} \left(\frac{d\mu}{d\lambda} \right)^{s}, \text{ where } \lambda := \frac{1}{2}\mu + \frac{1}{2}p, \ \lambda_k := \frac{1}{2}r_k + \frac{1}{2}p.$$

$$-\frac{1}{s} \ln \int dp \left(\frac{dp}{d\mu} \right)^{-s} \leqq C(M) \text{ for all } s > 0.$$

p is absolutely continuous with respect to μ. $p \not\ll \mu$ would imply that there is a support E of $p \wedge \mu$ with $p(E) < 1$.

$$\text{But } C(M) \geqq -\frac{1}{s} \ln \int dp \left(\frac{dp}{d\mu} \right)^{-s} = -\frac{1}{s} \ln \int_E d\mu \left(\frac{dp}{d\mu} \right)^{1-s} \geqq \frac{1-s}{s} \ln p(E) \to \infty$$

$(s \to 0)$.

One shows

$$C(M) \geqq -\frac{1}{s} \ln \int dp \left(\frac{dp}{d\mu} \right)^{-s} \xrightarrow{s \to 0} \int dp \ln \frac{dp}{d\mu}.$$

The uniqueness of μ similarly follows as in lemma 4. 3 or lemma 4. 2.

LEMMA 4. 5: *Let* $C(M_1 \times M_2) < \infty$. *Then the probability* μ *for* $M_1 \times M_2$ *with* $\int dp \ln \frac{dp}{d\mu} \leqq C(M_1 \times M_2)$ *for all* $p \in M_1 \times M_2$ *is a product probability.* (This is a consequence of lemma 4. 1, lemma 4. 4.)

We need in addition properties of the integrals $\int d\tilde{p} \ln^2 \tilde{f}$ when $\int d\tilde{p} \ln \tilde{f} \to C(M)$. Assume that $\sup_{\tilde{p} \in \tilde{M}} \int d\tilde{p} |\ln \tilde{f}|^{2+\delta}$ is finite for some $\delta > 0$. Then the set $\{F = F_{\tilde{p}} : \tilde{p} \in \tilde{M}\}$ of distribution functions $F_{\tilde{p}}(y) := \tilde{p}\{\ln \tilde{f} < y\}$ is totally bounded with respect to the Levy-metric. Furthermore, $L(M)$, the completion of this set, is a compact set of distribution functions on which $l_1(F) := \int y \, dF(y)$ and $l_2(F) := \int y^2 \, dF(y)$ are continuous functionals of F.

$L(M)_C := \{F \in L(M) : l_1(F) = C(M)\}$ is a compact convex subset of $L(M)$. Set $\sigma_F := \left[\int \left(y - \int y \, dF(y)\right)^2 dF(y)\right]^{\frac{1}{2}}$ for $F \in L(M)$ and

$$D_{(\varepsilon)}(M) := \begin{cases} \max \{\sigma_F : F \in L(M)_C & \text{if } 0 < \varepsilon < \frac{1}{2} \\ \min_{F \in L(M)_C} & \text{if } \frac{1}{2} \leqq \varepsilon < 1. \end{cases} \tag{12}$$

One has $L(M)_C = \bigcap_{b>0} \{F \in L(M) : l_1(F) > C(M) - b\}$ and it follows therefore from the compactness of $L(M)$ and $L(M)_C$ that for any Levy-neighbourhood of $L(M)_C$ a set $\{F : l_1(F) > C(M) - b\}$ is contained in this neighbourhood if $b > 0$ is sufficiently small. We have:

LEMMA 4. 6: *Suppose* $\sup_{\tilde{p} \in \tilde{M}} \int d\tilde{p} \, |\ln \tilde{f}|^{2+\delta} < \infty$. *Then for every* $d > 0$ *there is sufficiently small* $b > 0$ *s.t.*

$$\sigma_F \in [D_{(\frac{1}{4})}(M) - d, \, D_{(\frac{1}{4})}(M) + d] \quad \text{if} \quad l_1(F) > C(M) - b \qquad (F \in L(M)).$$

LEMMA 4.7:

Suppose $\int d\tilde{p} \, |\ln \tilde{f}|^{2+\delta} < K$ *for* $\tilde{p} \in \tilde{M}_{[1,2]}$. *Then* $D_{(\varepsilon)}(M_{[1,2]}) =$

$$= [D^2_{(\varepsilon)}(M_1) + D^2_{(\varepsilon)}(M_2)]^{1/2}.$$

We sketch a PROOF: Let $F \in L(M_{[1,2]})_C$. Then F is Levy-limit of distribution functions

$$\sum_{j=1}^{n} a^{(j)} F^{(j)}_{(y)} = \sum_{j=1}^{n} a^{(j)} p^{(j)} \left\{\ln \frac{dp^{(j)}}{d\mu_1 \times d\mu_2} < y\right\} =$$

$$= \sum_{j=1}^{n} a^{(j)} p^{(j)} \left\{\ln \frac{dp_1^{(j)}}{d\mu_1} + \ln \frac{dp_2^{(j)}}{d\mu_2} < y\right\} = \sum_{j=1}^{n} a^{(j)} (F_1^{(j)} * F_2^{(j)})_{(y)}$$

(μ_1, μ_2 according to lemma 4.5 and because of lemma 4. 2).

F behaves as an average of convolutions and

$$\sum a^{(j)} \left(\int y\, dF_{(y)}^{(j)} - \sum a^{(j)} \int y\, dF_{(y)}^{(j)}\right)^2 \to 0 \qquad \left(\sum a^{(j)} F_{(y)}^{(j)} \to F\right)$$

because $\int y\, dF_{(y)}^{(j)} \leqq C(M)$ and $\sum a^{(j)} \int y\, dF_{(y)}^{(j)} \to C(M)$.

The condition $\int d\tilde{p}\, |\ln \tilde{f}|^{2+\delta}$ takes care of the tails $\{y\colon |y| > R\}$ in the estimates.

REMARK: STRASSEN [4] uses a similar definition as (12) for his estimates for the finite alphabet stationary channel without memory for a classification of the input sources.

§ 5. Upper estimates of $N(M_{[1,t]}, \varepsilon)$ for $\frac{1}{2} < \varepsilon < 1$ and bounds for $N(M_{[1,t]}, \varepsilon)$ belonging to input sources which generate the capacity

THEOREM 5. 1: *Let a sequence* $\{M_v\}_{v=1,2,\ldots}$ *be given where the* M_v *are copies of each other, and suppose* $\sup_{\tilde{p}_1 \in \tilde{M}_1} \int d\tilde{p}_1\, |\ln \tilde{f}_1|^{2+\delta} < \infty$ *for some* $\delta > 0$.

Then

$$|\ln N(M_{[1,t]}, \varepsilon) - tC(M_1) + \Psi_{(\varepsilon)} D_{(\varepsilon)}(M_1) t^{1/2}| = o(t^{1/2}).$$

We will use again small perturbations of probabilities to derive the estimates. Furthermore, we conduct the proof in such a way that the later results can be seen:

PROOF: Going back to theorem 3. 5 a), we obtain, setting

$$Q = \sup_{\tilde{p}_{[1,t]}} \left(\int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} - \Psi_{(\varepsilon)} \left[\int d\tilde{p}_{[1,t]} \left(\ln \tilde{f}_{[1,t]} - \int d\tilde{p}_{[1,t]}\ \ln\ \tilde{f}_{[1,t]} \right)^2 \right]^{\frac{1}{2}} \right):$$

$$\ln N(M_{[1,t]}, \varepsilon) + o(t^{1/2}) \geqq Q \geqq C(M_{[1,t]} - \Psi_{(\varepsilon)} D_{(\varepsilon)}(M_{[1,t]}) =$$

$$= tC(M_1) - \Psi_{(\varepsilon)} D_{(\varepsilon)}(M_1) t^{1/2}.$$

The standard interpolation inequalities of inequality (4) (see f.i. [5]) take care of the remainder term in the estimate with $o(t^{\frac{1}{2}}) = O\left(t^{\frac{1}{2+\delta}}\right)$, as a consequence of the condition $\sup \int d\tilde{p}_1\, |\ln \tilde{f}|^{2+\delta} < \infty$.

Moreover, similarly to theorem 3. 5 b), it follows that for $0 < \varepsilon \leqq \frac{1}{2}$

$$\ln N(M_{[1,t]}, \varepsilon) - o(t^{1/2}) \leqq Q \leqq tC(M_1) - \Psi_{(\varepsilon)} D_{(\varepsilon)}(M_1) t^{1/2} + o(t^{1/2})$$

holds, where the last inequality sign is a consequence of

$$C(M_{[1,t]}) \geqq \int d\tilde{p}_{[1,t]} \ln \tilde{f}_{[1,t]} \geqq C(M_{[1,t]}) - o(t^{1/2})$$

(corollary 3. 4) and of lemma 4. 6.

Now let $\frac{1}{2}<\varepsilon<1$ and let $\{(p^i, E_i)\}$ $1 \leqq i \leqq N$ be an ε-code of maximal length for $M_{[1,t]}$,

$$q := q_1 \times \ldots \times q_t := \left(\frac{1}{N}\sum_{i=1}^{N} p_1^i\right) \times \cdots \times \left(\frac{1}{N}\sum_{i=1}^{N} p_t^i\right),$$

$$\tilde{E} := \frac{1}{N}\sum_{i=1}^{N} \int dp^i \ln \frac{dp^i}{dq}.$$

There is $G \subseteqq \{1, \ldots, N\}$ with $|G| > \dfrac{N}{t+1}$ s.t.

$$\int dp^i \ln \frac{dp^i}{dq} \leqq \left(1+\frac{1}{t}\right)\tilde{E} \quad \text{for} \quad i \in G.$$

Put $q' := q_1' \times \cdots \times q_t'$, where $q_v' := \dfrac{1}{|G|}\sum_{i\in G} p_v^i$. Then there is $H \subseteqq G$ with $|H| \geqq \dfrac{1}{2}|G|$ s.t. for $i \in H$

$$\sum_{v=1}^{t} \int dp_v^i \left|\ln \frac{dp_v^i}{dq_v'}\right|^{2+\delta} \leqq 2\sum_{v=1}^{t} \frac{1}{|G|}\sum_{j\in G} \int dp_v^j \left|\ln \frac{dp_v^j}{dq_v'}\right|^{2+\delta} = O(t)$$

holds. Set $q_v'' := \dfrac{1}{|H|}\sum_{i\in H} p_v^i$ $(1 \leqq v \leqq t)$, $q'' := q_1'' \times \cdots \times q_t''$,

$$\bar{q} := \bar{q}_1 \times \cdots \times \bar{q}_t \quad \text{where} \quad \bar{q}_v := \left(1-\frac{2}{t}\right)q_v + \frac{1}{t}\,q_v' + \frac{1}{t}\,q_v''.$$

For $i \in H$ holds

$$\int dp^i \ln \frac{dp^i}{d\bar{q}} \leqq \int dp^i \ln\left(\frac{dp^i}{dq}\left(1-\frac{2}{t}\right)^{-t}\right) \leqq \left(1+\frac{1}{t}\right)\tilde{E} - t\ln\left(1-\frac{2}{t}\right) \leqq \tilde{E} + O(1)$$

and

$$\sum_{v=1}^{t} \int dp_v^i \left|\ln \frac{dp_v^i}{d\bar{q}_v}\right|^{2+\delta} \leqq t\cdot C_1 + \sum_{v=1}^{t} \int dp_v^i \left|\ln\left(\frac{dp_v^i}{dq_v'}\,t\right)\right|^{2+\delta}$$

(uniformly bounding the integrals $\displaystyle\int_{\left\{\frac{dp_v^i}{d\bar{q}_v} \leqq 1\right\}} dp_v^i \left|\ln \frac{dp_v^i}{d\bar{q}_v}\right|^{2+\delta}$

as at the beginning of § 3). The right hand side of the inequality is $\leqq$

$$tC_1 + tC_2(\ln t)^{2+\delta} + O(t).$$

Hence

$$\left[\sum_{v=1}^{t} \int dp_v^i \left|\ln \frac{dp_v^i}{d\bar{q}_v}\right|^{2+\delta}\right]^{\frac{1}{2+\delta}} = O(t^{\frac{1}{2+\delta}} \ln t) \quad \text{for} \quad i \in H.$$

This together implies for $i \in H$:

$$\bar{q}(E_i) \geqq \exp\left[-\int dp^i \ln \frac{dp^i}{d\bar{q}} + \Psi_{(\varepsilon)}\left[\int dp^i \left(\ln \frac{dp^i}{d\bar{q}} - \int dp^i \ln \frac{dp^i}{d\bar{q}}\right)^2\right]^{1/2} - \right.$$

$$\left. - K_{(\varepsilon)}\left[\sum_{v=1}^{t} \int dp^i \left|\ln \frac{dp_v^i}{d\bar{q}_v}\right|^{2+\delta}\right]^{\frac{1}{2+\delta}} - O(\ln t)\right]$$

where δ is fixed $0 < \delta \leqq 1$ and $K_{(\varepsilon)} \geqq 0$ some function of ε for this fixed δ) (see f.i. [5])

$$2(t+1)\frac{1}{N} \geqq \frac{1}{|H|}\sum_{i \in H} \bar{q}(E_i) \geqq \exp\left[-\tilde{E} - O(t^{\frac{1}{2+\delta}} \ln t)\right] \cdot A$$

where

$$A := \frac{1}{|H|}\sum_{i \in H} \exp\left[\Psi_{(\varepsilon)}\left[\int dp^i \left(\ln \frac{dp^i}{d\bar{q}} - \int dp^i \ln \frac{dp^i}{d\bar{q}}\right)^2\right]^{1/2}\right].$$

Observe now, that $\exp[\Psi(\varepsilon) y^{\frac{1}{2}}]$ is a convex function of y for $y \geqq \left(\frac{1}{\Psi(\varepsilon)}\right)^2$ and that $(y_1)^{\frac{1}{2}} \geqq (y_1 + y_2)^{\frac{1}{2}} - (y_2)^{\frac{1}{2}}$ for $y_1, y_2 \geqq 0$.

Thus

$$A \geqq \exp\left[\Psi(\varepsilon)\left[B + \left(\frac{1}{\Psi(\varepsilon)}\right)^2\right]^{1/2} - \frac{1}{\Psi(\varepsilon)}\right] \geqq \exp[\Psi(\varepsilon) B^{1/2} - O(1)],$$

where

$$B := \frac{1}{|H|}\sum_{i \in H} \int dp^i \left(\ln \frac{dp^i}{d\bar{q}} - \int dp^i \ln \frac{dp^i}{d\bar{q}}\right)^2.$$

It remains to estimate B.

$$B = \sum_{v=1}^{t} \frac{1}{|H|}\sum_{i \in H} \int dp_v^i \ln^2 \frac{dp_v^i}{d\bar{q}_v} - \sum_{v=1}^{t} \frac{1}{|H|}\sum_{i \in H}\left(\int dp_v^i \ln \frac{dp_v^i}{d\bar{q}_v}\right)^2.$$

The split of

$$\ln^2 \frac{dp_v^i}{d\bar{q}_v} \quad \text{into} \quad \ln^2 \frac{dp_v^i}{dq_v''} + \ln^2 \frac{dq_v''}{d\bar{q}_v} + 2 \ln \frac{dp_v^i}{dq_v''} \ln \frac{dq_v''}{d\bar{q}_v}$$

shows:

(13)
$$\left|\frac{1}{|H|}\sum_{i \in H} \int dp_v^i \left(\ln^2 \frac{dp_v^i}{d\bar{q}_v} - \ln^2 \frac{dp_v^i}{dq_v''}\right)\right| \leqq$$

$$\leqq \int dq_v'' \ln^2 \frac{dq_v''}{d\bar{q}_v} + 2\left(\frac{1}{|H|}\sum_{i \in H} \int dp_v^i \ln^2 \frac{dp_v^i}{dq_v''}\right)^{1/2}\left(\int dq_v'' \ln^2 \frac{dq_v''}{d\bar{q}_v}\right)^{1/2} \leqq$$

$$\leqq C_3 R\left(\int dq_v'' \ln^2 \frac{dq_v''}{d\bar{q}_v}\right)$$

where

$$R(y) := y + y^{1/2}.$$

Now set $s(y):=y\ln^2 y$, $h(y):=y\ln y-y+1$. For an interval $[0, K_1]$ there is $K>0$ s.t. $Kh(y)\geqq s(y)$ for $y\in[0, K_1]$ and there is $K>0$ such that $y\ln y\leqq Kh(y)$ for all $y\geqq 2$. Because $\frac{dq_v''}{d\bar{q}_v}\leqq t$, one has K s.t.

$$\int dq_v''\ln^2\frac{dq_v''}{d\bar{q}_v}=\int dq_v s\left(\frac{dq_v''}{d\bar{q}_v}\right)\leqq\int dq_v h\left(\frac{dq_v''}{d\bar{q}_v}\right)K\ln t=\int dq_v''\ln\frac{dq_v''}{d\bar{q}_v}K\ln t.$$

With the concavity of $R(y)$, it follows

$$\left|\sum_{v=1}^{t}\frac{1}{|H|}\sum_{i\in H}\int dp_v^i\left(\ln^2\frac{dp_v^i}{d\bar{q}_v}-\ln^2\frac{dp_v^i}{dq_v''}\right)\right|\leqq tC_3R\left(\frac{1}{t}\sum_{v=1}^{t}\int dq_v''\ln^2\frac{dq_v''}{d\bar{q}_v}\right)\leqq \tag{14}$$

$$\leqq C_4 tR\left(\frac{\ln t}{t}\int dq''\ln\frac{dq''}{d\bar{q}}\right)=tO\left(\left(\int dq''\ln\frac{dq''}{d\bar{q}}\right)^{1/2}\right)t^{-1/2}\ln^{1/2}t.$$

Here

$$\int dq''\ln\frac{dq''}{d\bar{q}}=\frac{1}{|H|}\sum_{i\in H}\int dp^i\ln\frac{dp^i}{d\bar{q}}-\frac{1}{|H|}\sum_{i\in H}\int dp^i\ln\frac{dp^i}{dq''}\leqq$$

$$\leqq C(M_{[1,t]})+O(1)-\frac{1}{|H|}\sum_{i\in H}\int dp^i\ln\frac{dp^i}{dq''}$$

and therefore

$$\int dq''\ln\frac{dq''}{d\bar{q}}=O(t^{1/2}) \tag{15}$$

(because corollary 3.4 can be modified for $\frac{1}{|H|}\sum_{i\in H}\int dp^i\ln\frac{dp^i}{dq''}$). Next, we estimate $\sum_{v=1}^{t}\frac{1}{|H|}\sum_{i\in H}\left(\int dp_v^i\ln\frac{dp_v^i}{d\bar{q}_v}\right)^2$.

It similarly follows as above that

$$\left|\sum_{v=1}^{t}\frac{1}{|H|}\sum_{i\in H}\left(\left(\int dp_v^i\ln\frac{dp_v^i}{d\bar{q}_v}\right)^2-\left(\int dp_v^i\ln\frac{dp_v^i}{dq_v''}\right)^2\right)\right|= \tag{16}$$

$$=tO\left(\left(\int dq''\ln\frac{dq''}{d\bar{q}}\right)^{1/2}\right)t^{-1/2}\ln^{1/2}t.$$

Set $\bar{\mu}:=\bar{\mu}_1\times\cdots\times\bar{\mu}_t$ where $\bar{\mu}_v:=\left(1-\frac{1}{t}\right)\mu_v+\frac{1}{t}q_v''$, μ_v for M_v is the probability which has been discussed in lemma 4.4. With $\frac{dq_v''}{d\bar{\mu}_v}\leqq t$, one obtains as above

$$\left|\sum_{v=1}^{t}\frac{1}{|H|}\sum_{i\in H}\left(\left(\int dp_v^i\ln\frac{dp_v^i}{d\bar{\mu}_v}\right)^2-\left(\int dp_v^i\ln\frac{dp_v^i}{dq_v''}\right)^2\right)\right|= \tag{17}$$

$$=tO\left(\left(\int dq''\ln\frac{dq''}{d\bar{\mu}}\right)^{1/2}\right)t^{-1/2}\ln^{1/2}t,$$

again with

$$\int dq'' \ln \frac{dq''}{d\bar{\mu}} = O(t^{1/2}).$$

Finally,

$$\left| \sum_{v=1}^{t} \left(C^2(M_v) - \frac{1}{|H|} \sum_{i \in H} \left(\int dp^i \ln \frac{dp^i}{d\bar{\mu}_v} \right)^2 \right) \right| \leqq$$

$$\leqq \sum_{v=1}^{t} \left(C^2(M_v) - \frac{1}{|H|} \sum_{i \in H} \left(\int dp_v^i \ln \frac{dp_v^i}{d\mu_v} \right) \right) + O(1) = O(t^{1/2}),$$

because

$$\frac{1}{|H|} \sum_{i \in H} \int dp_v^i \ln \frac{dp_v^i}{dq_v''} \leqq \frac{1}{|H|} \sum_{i \in H} \int dp_v^i \ln \frac{dp_v^i}{d\mu_v} \leqq C(M_v)$$

and

$$\int dp_v^i \ln \frac{dp_v^i}{d\mu_v} \leqq C(M_v) \quad \text{for} \quad 1 \leqq i \leqq N.$$

Thus,

$$\frac{1}{N} \geqq \exp\left[-\tilde{E} + \Psi_{(\varepsilon)} \left[\left(\bar{B} - O(t^{3/4} \ln^{1/2} t) \right)^+ \right]^{1/2} - O\left(t^{\frac{1}{2+\delta}} \ln t \right) \right],$$

where

$$\bar{B} := \sum_{v=1}^{t} \frac{1}{|H|} \sum_{i \in H} \int dp_v^i \ln^2 \frac{dp_v^i}{dq_v''} - \sum_{v=1}^{t} \left(\frac{1}{|H|} \sum_{i \in H} \int dp_v^i \ln \frac{dp_v^i}{dq_v''} \right)^2.$$

$\bar{B} \geqq D^2_{(\varepsilon)}(M_{[1,t]}) - o(t)$ because

$$\sum_{v=1}^{t} \frac{1}{|H|} \sum_{i \in H} \int dp_v^i \ln \frac{dp_v^i}{dq_v''} \geqq C(M_{[1,t]}) - O(t^{1/2}).$$

This implies the theorem. One immediately concludes from the proof the

THEOREM 5. 2: *Let a sequence* $\{M_v\}_{v=1,2,\ldots}$ *be given and suppose* $\sup_v \sup_{\tilde{p}_v \in \tilde{M}_v} \int d\tilde{p}_v |\ln \tilde{f}_v|^{2+\delta} < \infty$ *for some* $\delta > 0$.

Furthermore, suppose that for every $d > 0$ *there is* $b > 0$ *s.t.*

$$\sum_{v=1}^{t} \int d\tilde{p}_v \ln \tilde{f}_v > C(M_{[1,t]}) - dt$$

implies

$$\sum_{v=1}^{t} \int d\tilde{p}_v \left(\ln \tilde{f}_v - \int d\tilde{p}_v \ln \tilde{f}_v \right)^2 \in [D^2_{(3/4)}(M_{[1,t]}) - dt,\ D^2_{(1/4)}(M_{[1,t]}) + dt].$$

Then

$$\left| \ln N(M_{[1,t]}, \varepsilon) - C(M_{[1,t]}) + \Psi_{(\varepsilon)} D_{(\varepsilon)}(M_{[1,t]}) \right| = o(t^{1/2}).$$

REMARK: The second supposition has to be made because we do not have from the first supposition lemma 4. 6 uniformly for all v. We obtain, furthermore, from the proof of theorem 5. 1 the

LEMMA 5. 3: *Under the supposition of theorem* 5. 2 *holds: For* ε *fixed* $0 < \varepsilon < 1$ *there is a function* $o(t^{\frac{1}{2}})$ *s.t.: If* $\{(p^i, F_i)\}_{1 \leqq i \leqq N}$ *is an* ε*-code of maximal length for* $M_{[1,t]}$, *then*

$$\left| C(M_{[1,t]}) - \frac{1}{N} \sum_{i=1}^{N} \int dp^i \ln \frac{dp^i}{dq} \right| \leqq o(t^{1/2}),$$

$$\left(\text{where } q: = \left(\frac{1}{N} \sum_{i=1}^{N} p_1^i \right) \times \cdots \times \left(\frac{1}{N} \sum_{i=1}^{N} p_t^i \right) \right).$$

If $b(d)$ of lemma 4 .6 is known for the M_v, then $o(t^{\frac{1}{2}})$ can be improved in lemma 5. 2. We obtain for example:

THEOREM 5. 4: *Let a sequence* $\{M_v\}_{v=1,2,\ldots}$ *be given, where the* M_v *are copies of each other and suppose that there is* λ, *for* M_1 *and* $K \geqq 1$ *s.t.* $\frac{dp_1}{d\lambda_1} \leqq K$ *for all* $p_1 \in M_1$. *Furthermore, suppose* $D_{(\frac{3}{4})}(M_1) > 0$ *and that, with a fixed constant* $c > 0$,

$$\int d\tilde{p}_1 \ln \tilde{f}_1 > C(M_1) - b$$

implies

$$D^2_{(3/4)}(M_1) - cb^{1/2} < \int d\tilde{p}_1 \left(\ln \tilde{f}_1 - \int d\tilde{p}_1 \ln \tilde{f}_1 \right)^2 < D^2_{(1/4)}(M_1) + cb^{1/2} \quad (b > 0).$$

Then for every u $(0 < u)$ *holds*

a)
$$|\ln N(M_{[1,t]}, \varepsilon) - tC(M_1) + \Psi_{(\varepsilon)} D(M_1) t^{1/2}| \leqq O(t^u),$$

b)
$$\left| tC(M_1) - \frac{1}{N} \sum_{i=1}^{N} \int dp^i \ln \frac{dp^i}{dq} \right| \leqq O(t^u)$$

(with the left hand side term according to lemma 5. 3*).*

PROOF: One obtains a remainder term $O(\ln^2 t)$ instead of $O(t^{\frac{1}{2+\delta}} \ln t)$ in the estimate of theorem 5. 1 using the proof of theorem 3. 6. We may forget about the additional small perturbation of probabilities which has to be used and write with the notation of the proof of theorem 5. 1 for $\varepsilon > \frac{1}{2}$ (the case $\varepsilon \leqq \frac{1}{2}$ would only be shorter):

$$\frac{1}{N} \geqq \exp \left[-\tilde{E} + \Psi_{(\varepsilon)} \left[\left(\bar{B} - (t^{1/2} \ln^{1/2} t) O\left(\left(\int dq'' \ln \frac{dq''}{d\mu} \right)^{1/2} \right) \right)^{+} \right]^{1/2} - O(\ln^2 t) \right].$$

Observing, that $\sum_{v=1}^{t} \int d\tilde{p}_v \ln \tilde{f}_v > tC(M_1) - \sum_{v=1}^{t} b_v$ implies

$$\sum_{v=1}^{t} \int d\tilde{p}_v \left(\ln \tilde{f}_v - \int d\tilde{p}_v \ln \tilde{f}_v \right)^2 > tD^2_{(\varepsilon)}(M_1) - c \sum_{v=1}^{t} b_v^{1/2} \geqq$$

$$\geqq tD^2_{(\varepsilon)}(M_1) - ct \left(\frac{1}{t} \sum_{v=1}^{t} b_v \right)^{1/2},$$

we have:

$$\frac{1}{N} \geqq \exp\left[-\tilde{E}+\Psi_{(\varepsilon)}\left[\left(tD^2_{(\varepsilon)}(M_1)-ct\left(C(M_1)-\frac{1}{t}\sum_v \int d\tilde{p}_v \ln \tilde{f}_v\right)^{1/2}-\right.\right.\right.$$

$$\left.\left.\left.-(t^{1/2}\ln^{1/2} t)\,O\left(\left(\int dq'' \ln \frac{dq''}{du}\right)^{1/2}\right)\right)^{+}\right]-O(\ln^2 t)\right],$$

where

$$\int d\tilde{p}_v \ln \tilde{f}_v := \frac{1}{|H|}\sum_{i\in H}\int dp_v^i \ln \frac{dp_v^i}{dq_v''}.$$

$$\int dq_v'' \ln \frac{dq_v''}{d\mu_v} = \frac{1}{|H|}\sum_{i\in H}\int dp_v^i\left(\ln \frac{dp_v^i}{d\mu_v}-\ln \frac{dp_v^i}{dq_v''}\right) \leqq C(M_v)-\frac{1}{|H|}\sum_{i\in H}\int dp_v^i \ln \frac{dp_v^i}{dq_v''}.$$

Together with $(1+y)^{1/2} \approx 1+\frac{1}{2}y$ for small $|y|$ follows

$$\frac{1}{N} \geqq \exp\left[-\tilde{E}+\Psi_{(\varepsilon)}D_{(\varepsilon)}(M_1)t^{1/2}\left[1-c_1\frac{\left(C(M_1)-\frac{1}{t}\sum_v\int d\tilde{p}_v\ln\tilde{f}_v\right)^{1/2}}{D^2_{(\varepsilon)}(M_1)}-\right.\right.$$

$$\left.\left.-(t^{-1/2}\ln^{1/2} t)\,O\left(\left(\int dq''\ln\frac{dq''}{d\mu}\right)^{1/2}\right)\right]-O(\ln^2 t)\right] \geqq$$

$$\geqq \exp\left[-\tilde{E}+\Psi_{(\varepsilon)}D_{(\varepsilon)}(M_1)t^{1/2}-c_2 t^{1/4}-c_3 t^{1/4}\ln^{1/2} t-O(\ln^2 t)\right]$$

for some constants $c_1, c_2, c_3 \geqq O$, because

$$\frac{1}{|H|}\sum_{i\in H}\int dp^i \ln\frac{dp^i}{dq''}$$

satisfies corollary 3.4. It follows $tC(M_1)-\tilde{E} \geqq O(t^{1/4}\ln^{1/2} t)$. Now observe that $\frac{1}{|H|}\sum_{i\in H}\int dp^i \ln\frac{dp^i}{dq''}$ belongs to a code of length $\frac{N}{2(t+1)}$ and that the same situation as for $\tilde{E}$ holds for $\frac{1}{|H|}\sum_{i\in H}\int dp^i\ln\frac{dp^i}{dq''}$. One obtains the theorem by iterating H.

REMARK: One immediately obtains nonstationary generalizations of this theorem.

We finally show, how the last result may be connected with STRASSEN's estimates (see [4], the estimate in [4] is a bit sharper):

COROLLARY 5.5: For the stationary memoryless channel with finite alphabets and with $D_{(\frac{1}{4})}(M_1)>0$ holds

$$|\ln N(M_{[1,t]},\varepsilon)-tC(M_1)+\Psi_{(\varepsilon)}D_{(\varepsilon)}(M_1)t^{1/2}| = O(t^u) \quad \text{for every} \quad u>0.$$

We sketch a proof:

1. Let $M_1 = \{p_1^{(1)}\}, \ldots, p_1^{(n)}\}$. Then the supposition of theorem 5. 4 is fulfilled:

$$\int d\tilde{p}_1 \ln \tilde{f}_1 = \int d\tilde{p}_1 \ln \frac{d\tilde{p}_1}{da_1 \times d\mu_1} - \int dq_1 \ln \frac{dq_1}{d\mu_1} =$$

$$= \sum_{j=1}^{n} a^{(j)} \int dp_1^{(j)} \ln \frac{dp_1^{(j)}}{d\mu_1} - \int dq_1 \ln \frac{dq_1}{d\mu_1}.$$

$$\int d\tilde{p}_1 \ln \tilde{f}_1 > C(M_1) - \delta \quad \text{implies} \quad \sum_{j=1}^{n} a^{(j)} \int dp_1^{(j)} \ln \frac{dp_1^{(j)}}{d\mu} >$$

$$> C(M_1) - \delta \quad \text{and} \quad \int dq_1 \ln \frac{dq_1}{d\mu_1} < \delta.$$

Furthermore, we have

$$a\left\{j : C(M_1) \geqq \int dp_1^{(j)} \ln \frac{dp_1^{(j)}}{d\mu_1} > C(M_1) - \delta\right\} > 1 - \bar{C}_1 \delta$$

(for some constant $\bar{C}_1$) and we obtain together with the inequality

$$\|q_1 - \mu_1\| \leqq C_1 \int dq_1 \ln \frac{dq_1}{d\mu_1} + C_2 \left(\int dq_1 \ln \frac{dq_1}{d\mu_1}\right)^{1/2}$$

(see proof of lemma 4. 2),
$\inf_{b \in G} \|a - b\| \leqq \bar{C}_2 \delta^{\frac{1}{4}}$ (for some constant $\bar{C}_2$) as a matter of linear inequalities in R^n, where

$$G := \left\{(b^{(1)}, \ldots, b^{(n)}) : b^{(j)} \geqq 0 \quad (1 \leqq j \leqq n), \sum b^{(j)} = 1 \text{ and} \right.$$

$$\left. \sum_{j=1}^{n} b^{(j)} \int dp_1^{(j)} \ln \frac{dp_1^{(j)}}{d\left(\sum_{k=1}^{n} b^{(k)} p^{(k)}\right)} = C(M_1)\right\};$$

for the above norm
of $a - b$ f.i. the euclidean norm in R^n may be taken.
Now take b s.t. $\|a - b\| = \inf_{b \in G} \|a - b\|$. Then

$$\left| \int d\tilde{p}_1 \ln^2 \tilde{f}_1 - \sum a^{(j)} \int dp_1^{(j)} \ln^2 \frac{dp_1^{(j)}}{d\mu_1} \right| \leqq O\left(\left(\int dq_1 \ln \frac{dq_1}{d\mu_1}\right)^{1/2}\right) = O(\delta^{1/2})$$

(see the proof of theorem 5. 1, $\ln \frac{dp^{(j)}}{d\mu_1}$ is bounded $p^{(j)}$-a.e. above and below uniformly with respect to j $(1 \leqq j \leqq n)$ because the alphabet is finite),

$$\sum_{j=1}^{n} b^{(j)} \int dp_1^{(j)} \ln^2 \frac{dp_1^{(j)}}{d\left(\sum_{k=1}^{n} b^{(k)} p_1^{(k)}\right)} = \sum_{j=1}^{n} b^{(j)} \int dp_1^{(j)} \ln^2 \frac{dp_1^{(j)}}{d\mu_1}$$

and

$$\left| \sum_{j=1}^{n} (a^{(j)} - b^{(j)}) \int dp_1^{(j)} \ln^2 \frac{dp_1^{(j)}}{d\mu_1} \right| \leqq \bar{C}_3 \bar{C}_2 \delta^{1/2}.$$

REMARK: Our main result are the theorems 5. 1, 5. 2, 5. 4 and lemma 5. 3. The sharp bounds given in theorem 3. 5 and theorem 3. 6 have to be considered as a matter of minor importance because of computational reasons and because these are bounds for the case $\varepsilon < \frac{1}{2}$, (one is interested in the case $\varepsilon > \frac{1}{2}$).

REFERENCES

[1] AUGUSTIN, U.: Gedächtnisfreie Kanäle für diskrete Zeit. *Z. Wahrscheinlichkeitstheorie verw. Geb.* **6** (1966).

[2] PINSKER, M. S.: *Arbeiten zur Informationstheorie V.* Math. Forschungsgebiete XVIII, VEB Deutscher Verlag d. Wiss. Berlin 1963.

[3] WOLFOWITZ, J.: *Coding Theorems of Information Theory.* Ergebn. d. Math. u. Grenzgebiete, Bd 31. Springer, Berlin—Göttingen—Heidelberg 1964.

[4] STRASSEN, V.: Asymptotische Abschätzungen in Shannons Informationstheorie. *Transactions of the third Prague Conference on Information Theory, Statistical Decision Functions and Random Processes* (1965).

[5] OSIPOV, L. V. and PETROV, V. V.: On the Estimation of the Remainder in the Central Limit Theorem, *Theory of Probability,* **12**, No. 2 (1967).

Universität Erlangen, Erlangen, West Germany

(Received December 18, 1968.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 127—135

A NOTE ON THE MATRIX INVERSION BY THE PARTITIONING TECHNIQUE

by

L. MIHÁLYFFY

1. Introduction. The partitioning technique for inverting large matrices is a frequently used method of numerical analysis. Its applicability is, however, rather limited by the requirement on the regularity of certain submatrices of the matrix to be inverted. In the simplest case, the invertible $n \times n$ matrix A is partitioned in the form

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{11}, A_{12}, A_{21} and A_{22} are $p \times p$, $p \times q$, $q \times p$ and $q \times q$ matrices, respectively, $p+q = n$, and X, the inverse of A, is assumed to have the same partitioned form as A. Consequently, the equations

$$\begin{aligned} A_{11}X_{11} + A_{12}X_{21} &= I \\ A_{21}X_{11} + A_{22}X_{21} &= 0 \\ A_{11}X_{12} + A_{12}X_{22} &= 0 \\ A_{21}X_{12} + A_{22}X_{22} &= I, \end{aligned}$$

with unit matrices I and zero matrices 0 of the appropriate order on the right side, must hold. In addition, the existence of the inverse matrices A_{11}^{-1} and $(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$ (or, what is essentially the same, the existence of A_{22}^{-1} and $(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$) is usually required, and then the submatrices of X can be given by the formulae

$$\begin{aligned} X_{22} &= (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \\ X_{12} &= -A_{11}^{-1}A_{12}X_{22} \\ X_{21} &= -X_{22}A_{21}A_{11}^{-1} \\ X_{11} &= A_{11}^{-1} - A_{11}^{-1}A_{12}X_{21}. \end{aligned}$$

The description of this standard technique can be found e.g. in [1]. However, as the example of the matrix

$$\begin{pmatrix} 0 & \dots & 0 & 1 \\ 0 & \dots & 1 & 0 \\ \vdots & & & \\ 1 & \dots & 0 & 0 \end{pmatrix}$$

shows, an invertible matrix need not have any regular proper principal minor. Obviously, the problem of finding a regular $p \times p$ submatrix by suitable rearranging of the rows and columns of the given invertible $n \times n$ matrix requires a considerable amount of computations in an unfavourable case, in particular, if both p and n are sufficiently large and $p \approx \frac{1}{2}n$. The purpose of this paper is to improve the partitioning technique in such a manner that it may be carried out whenever the given matrix is invertible, no matter what properties the submatrices have.

In the following Section 2. we shall fix the notations used in the paper. Section 3. contains some auxiliary results we need for developing our main object; one of them ("R-orthogonalization of matrices") is a generalization of the well known GRAM-SCHMIDT procedure for finite-dimensional Euclidean vector spaces. The main results are formulated in Sec. 4., and some conclusions concerning practical aspects can be found in Sec. 5.

2. Notations. The following conventions will be used throughout. The letters $i, j, k, m, n, N, p, q, p_i, q_j, q_{\max}, p_{\max}, a(.,.)$ and $m(.,.)$ will represent positive integers: indices, dimensions of matrices, etc.;

$\mathbf{f}$ denotes p-dimensional (column) vector;

$A, B, C, D, E, F, G, H, H', S, T, U, V, V', X, Y, Y'$ and Z with or without indices stand for complex matrices;

I denotes unit matrix of appropriate order and 0 denotes rectangular zero matrix of appropriate order.

The adjoint, i.e. the conjugate transpose of a matrix A or of a vector $\mathbf{f}$ will be denoted by A^* and $\mathbf{f}^*$, respectively. The adjoint of a column vector is a row vector and vice versa. Finally, we denote the p-dimensional complex arithmetic vector space by $\mathbf{R}^p$, and the Euclidean norm of an element $\mathbf{f} = (f_1, \ldots, f_p)$ of this space, i.e. the quantity $\sqrt{f_1^2 + f_2^2 + \cdots + f_p^2}$, by $\|\mathbf{f}\|$.

3. Preparations; R-orthogonal matrices. The main tool used in our investigations is the MOORE—PENROSE generalized inverse of matrices [2]. We recall that the generalized inverse of an arbitrary complex non-zero matrix A is the unique solution of the following four equations:

$$AXA = A \tag{1a}$$

$$XAX = X \tag{1b}$$

$$(AX)^* = AX \tag{1c}$$

$$(XA)^* = XA \tag{1d}$$

which we shall denote by A^+. In addition, for the sake of uniformity we shall consider the $n \times m$ zero matrix as the generalized inverse of the $m \times n$ zero matrix.

Now we are going to introduce the concept of the R-orthogonality of matrices. To this end we have to consider an elementary fact which is expressed in the following

LEMMA. *Let A be a non-zero $p \times m$ matrix, B a non-zero $p \times q$ matrix. Then $A^+ B = 0$ if and only if $B^+ A = 0$.*

PROOF. Form the adjoints on both sides of (1a) and combine this result with (1c); we obtain that $A^*AA^+ = A^*$. Therefore $A^+B = 0$ implies $A^*B = 0$ or $B^*A = 0$. On the other hand, $B^*BB^+ = B^*$ holds similarly, so we have $B^*BB^+A = 0$. This is equivalent to $BB^+A = 0$, which, by virtue of (1b), proves necessity. Suffiecincy follows by changing the notation. (Note that $(A^+)^+ = A$).

We remark that our assertion was not trivial since, in general, $(CD)^+ \neq D^+C^+$.

This simple observation allows us to establish the following

DEFINITION. The non-zero $p \times m$ matrix A and the non-zero $p \times q$ matrix B are called R-orthogonal to each other if $A^+B = 0$. A $p \times m$ zero matrix is R-orthogonal to every $p \times q$ matrix.[1]

Since AA^+ and BB^+ are projections onto the range of A and B, respectively, $A^+B = 0$ implies that the range of A and that of B are orthogonal subspaces, so the concept of R-orthogonality deserves its name. Moreover, since a p-dimensional column vector $\mathbf{f}$ can be regarded as a $p \times 1$ matrix whose generalized inverse is $\frac{1}{\|\mathbf{f}\|^2}\mathbf{f}^*$ we see that the concept of R-orthogonality of p-dimensional column vectors as $p \times 1$ matrices coincides with that of the usual orthogonality induced by the scalar product.

Naturally, we call a system $S_1, \ldots, S_n$ of $p \times q_i$ matrices $(i = 1, \ldots, n)$ R-orthogonal if any two elements of it are R-orthogonal to each other. The following theorem shows how an arbitrary system $T_1, \ldots, T_n$ of $p \times q_i$ matrices $(i = 1, \ldots, n)$ can be replaced by an R-orthogonal one.

THEOREM 1. (R-orthogonalization of matrices.) *Let $T_1, \ldots, T_n$ be an arbitrary system of $p \times q_i$ matrices $(i = 1, \ldots, n)$. Define the S's via induction*

$$S_1 = T_1$$

$$S_k = (I - S_1 S_1^+ - \cdots - S_{k-1} S_{k-1}^+)T_k, \quad k = 2, 3, \ldots, n. \tag{2}$$

Then

(i) *the system $S_1, \ldots, S_n$ is R-orthogonal;*

(ii) *S_i is R-orthogonal to T_j if $j < i$;*

(iii) *$S_i^+ S_i = S_i^+ T_i$ for every index i;*

(iv) *the subspace in $\mathbf{R}^p$ spanned by the ranges of the matrices $T_1, \ldots, T_n$ coincides with that spanned by the ranges of the matrices $S_1, \ldots, S_n$.*

PROOF. Assertion (i) is trivial for $k = 1$. Suppose it holds for $2, 3, \ldots, k-1$, then

$$S_i^+ S_k = (S_i^+ - S_i^+ S_i S_i^+)T_k$$

if $1 \leqq i \leqq k-1$, which, by virtue of (1b), completes the proof of (i).

A rearranging of (2)

$$T_k = S_k + (S_1 S_1^+ + \cdots + S_{k-1} S_{k-1}^+)T_k \tag{2a}$$

shows the validity of (ii) and (iii) using the R-orthogonality of the S's proved above.

[1] The concepı of R-orthogonality defined here is actually equivalent to that of *-orthogonality introduced by M. HESTENES in [4]. In HESTENES' paper, however, there is no orthogonalization procedure.

To prove (iv), first we observe that the ranges of the matrices T_1 and S_1 being equal to each other coincide. Assuming that the ranges of the matrices $T_1, \ldots, T_{k-1}$ and those of the matrices $S_1, \ldots, S_{k-1}$ span the same subspace we conclude the analogous statement for k; $2 \leqq k \leqq n$. Indeed, (2) and the induction hypothesis imply the relation "$\supseteq$", while (2a) implies the relation "$\subseteq$" between the two subspaces in question. Our theorem is thereby proved.

As we saw above, the concept of R-orthogonality is a generalization of the usual orthogonality of column vectors. The process described in Theorem 1 can be considered therefore as a generalization of the GRAM—SCHMIDT orthogonalization process for finite-dimensional Euclidean vector spaces.

Part (iv) of the above theorem can be stated more explicitly. Actually we can give the exact relation between the solutions of eq. (3) and that of eq. (5) (see below) provided that the coefficients of the latter are R-orthogonal matrices constructed from the coefficients of the former according to Theorem 1. This relation plays an important role in the next section, therefore we formulate it as

THEOREM 2. *Let $T_1, \ldots, T_n$ be given $p \times q_i$ matrices $(i = 1, 2, \ldots, n)$ and $S_1, \ldots, S_n$ be R-orthogonal matrices constructed from the T's according to Theorem 1, furthermore, let T_0 be a given $p \times q$ matrix. Then every solution $U_1, U_2, \ldots, U_n$ of the equation*

$$T_1 U_1 + T_2 U_2 + \cdots + T_n U_n = T_0 \tag{3}$$

can be written in the form

$$U_k = \sum_{j=k}^{n} H_{kj} V_j \qquad k = 1, 2, \ldots, n, \tag{4}$$

where $V_1, \ldots, V_n$ is a solution of the equation

$$S_1 V_1 + S_2 V_2 + \cdots + S_n V_n = T_0 \tag{5}$$

and the matrices H_{kj} $(k = 1, 2, \ldots, n;\ j = k,\ k+1, \ldots, n)$ depend upon $S_1, \ldots, S_n$, $T_1, \ldots, T_n$ only.

PROOF. We shall prove this theorem in two steps. First we show that the matrices $U_1, \ldots, U_n$ defined by the formulae

$$U_n = V_n$$

$$U_k = V_k - S_k^+ (T_{k+1} U_{k+1} + \cdots + T_n U_n) \qquad k = n-1, n-2, \ldots, 2, 1 \tag{6}$$

satisfy eq. (3) if and only if $V_1, \ldots, V_n$ in these formulae is a solution of eq. (5); then we point out that suitably constructing the H's we can replace (6) by (4).

For the first step we note that parts (ii) and (iii) of Theorem 1 admit us to write (6) in the form

$$U_k = V_k + S_k^+ S_k U_k - S_k^+ (T_1 U_1 + \cdots + T_n U_n). \tag{6a}$$

Multiplying this equation by S_k, taking the sum for $k = 1, 2, \ldots, n$ and applying (1a) we obtain

$$\sum_{k=1}^{n} S_k U_k = \sum_{k=1}^{n} S_k V_k + \sum_{k=1}^{n} S_k U_k - \sum_{k=1}^{n} S_k S_k^+ \sum_{j=1}^{n} T_j U_j.$$

Using (iv) of Theorem 1 and realizing that $\sum_{k=1}^{n} S_k S_k^+$, as a sum of orthogonal projections, is the projection onto the span of the ranges of the matrices S_i we see that the last term on the right side of the latter equation is equal to $\sum_{j=1}^{n} T_j U_j$.

Thus we have

$$\sum_{k=1}^{n} S_k V_k = \sum_{j=1}^{n} T_j U_j$$

and this verifies our first assertion. To prove the second one, we define the matrices H_{kj} as follows:

$$H_{jj} = I \quad (q_j\text{-dimensional})$$

$$j = 1, 2, \ldots, n; \tag{7}$$

$$H_{kj} = -S_k^+ (T_{k+1} H_{k+1,j} + T_{k+2} H_{k+2,j} + \cdots + T_{j-1} H_{j-1,j} + T_j)$$

$$k = j-1, j-2, \ldots, 2, 1.$$

We assert that (6) and (7) imply (4). The assertion is trivial for $k = n$. Assume it holds for $k+1, k+2, \ldots, n-1, n$. Then it is obvious from (6) that U_k can be written in the form

$$U_k = \sum_{j=k}^{n} H'_{kj} V_j.$$

It is sufficient to prove that $H'_{kj} = H_{kj}$ for some j, $k \leqq j \leqq n$. Using the induction hypothesis let us substitute the expressions of $U_{k+1}, U_{k+2}, \ldots, U_n$ given by (4) into (6) and notice that V_j occurs in the expressions of $U_{k+1}, U_{k+2}, \ldots, U_j$ only. Taking then the sum of the coefficients of V_j we obtain the same formula for H'_{kj} as for H_{kj} in (7).

This completes the proof of Theorem 2.

The importance of this theorem is obvious when we take in account that the general solution of (5) can be written in the following form

$$V_i = S_i^+ T_0 + (I - S_i^+ S_i) V'_i \qquad i = 1, 2, \ldots, n \tag{8}$$

where V'_i is an arbitrary $q_i \times q$ matrix; this follows immediately from the R-orthogonality of the S's and of the properties of the generalized inverse.

Summarizing our results so far we can state that the R-orthogonalization process, Theorem 2 and (8) permit us to give the general solution of an arbitrary equation of the form (3) in a rather simple way.

4. The general partitioning procedure. Now let us turn our attention to our original problem. Let us consider an invertible $N \times N$ matrix A and partition it in the following fashion:

$$A = \begin{pmatrix} A_{11} & A_{12} \ldots A_{1n} \\ A_{21} & A_{22} \ldots A_{2n} \\ \vdots & \\ A_{m1} & A_{m2} \ldots A_{mn} \end{pmatrix} \tag{9}$$

where A_{ij} is a $p_i \times q_j$ matrix, $i=1, \ldots, m$, $j=1, \ldots, n$; $p_1+p_2+\cdots+p_m = N$, $q_1+q_2+\cdots+q_n = N$. Then the corresponding partitioned form for X, the inverse of A, will be

$$X = \begin{pmatrix} X_{11} & X_{12} \ldots X_{1m} \\ X_{21} & X_{22} \ldots X_{2m} \\ \vdots & \\ X_{n1} & X_{n2} \ldots X_{nm} \end{pmatrix}$$

where X_{jk} is a $q_j \times p_k$ matrix, $j=1, \ldots, n$, $k=1, \ldots, m$; q_j and p_k are the same as above.

For fixed i we have

$$A_{i1}X_{1i}+A_{i2}X_{2i}+\cdots+A_{in}X_{ni} = I \tag{10}$$

and

$$A_{j1}X_{1i}+A_{j2}X_{2i}+\cdots+A_{jn}X_{ni} = 0 \tag{11a}$$

if $j = 1, \ldots, i-1, i+1, \ldots, m$.

Introducing the notations

$$B_{i1} = \begin{pmatrix} A_{11} \\ \vdots \\ A_{i-1,1} \\ A_{i+1,1} \\ \vdots \\ A_{m1} \end{pmatrix}, \ldots, \quad B_{in} = \begin{pmatrix} A_{1n} \\ \vdots \\ A_{i-1,n} \\ A_{i+1,n} \\ \vdots \\ A_{mn} \end{pmatrix}$$

(11a) can be written in the form

$$B_{i1}X_{1i}+B_{i2}X_{2i}+\cdots+B_{in}X_{ni} = 0 \tag{11}$$

where B_{ij} is an $(N-p_i)\times q_j$ matrix; $j=1, \ldots, n$. We shall solve the system of equations (10) and (11) by repeated application of the R-orthogonalization process. The idea of what we are going to do may be. outlined as follows. First, using (8) and (4) we determine the general solution of (10) This will contain n undetermined matrices corresponding to V_i' in (8). To compute these we substitute the general solution into (11). This step will lead to a (non-homogeneous) equation very similar to (10). Another application of formulae (8) and (4) completes the procedure. In what follows we describe the method in detail.

Let us denote the elements of the R-orthogonal system constructed from $A_{i1}, \ldots, A_{in}$ by $C_1, \ldots, C_n$, respectively. (There will be no ambiguity because of omitting the index i, since the value of i is fixed unless otherwise stated.) Let us consider the equation

$$C_1Y_1+C_2Y_2+\cdots+C_nY_n = I,$$

whose general solution is, by virtue of (8)

$$Y_j = C_j^+ + (I-C_j^+C_j)Y_j' \qquad j=1, \ldots, n,$$

where Y_j' is an undetermined $q_j \times p_i$ matrix.

Applying (4) we see that the general solution of (10) can be written in the form

$$X_{ki} = \sum_{j=k}^{n} F_{kj} Y_j = \sum_{j=k}^{n} F_{kj}(C_j^+ + (I - C_j^+ C_j) Y_j') \tag{12}$$

where, according to (7),

$$F_{jj} = I \quad (q_j\text{-dimensional}) \qquad j = 1, 2, \ldots, n;$$

$$F_{kj} = -C_k^+ (A_{i,k+1} F_{k+1,j} + A_{i,k+2} F_{k+2,j} + \cdots + A_{i,j-1} F_{j-1,j} + A_{ij})$$

$$k = j-1, j-2, \ldots, 2, 1.$$

Substituting the expressions of X_{ki} $(k = 1, 2, \ldots, n)$ into (11) we obtain

$$B_{i1}(I - C_1^+ C_1) Y_1' +$$

$$(B_{i1} F_{12} + B_{i2})(I - C_2^+ C_2) Y_2' +$$

$$\cdots$$

$$+ (B_{i1} F_{1n} + B_{i2} F_{2n} + \cdots + B_{i,n-1} F_{n-1,n} + B_{in})(I - C_n^+ C_n) Y_n' =$$

$$= -\big(B_{i1} C_1^+ + (B_{i1} F_{12} + B_{i2}) C_2^+ + \cdots +$$

$$+ (B_{i1} F_{1n} + B_{i2} F_{2n} + \cdots + B_{i,n-1} F_{n-1,n} + B_{in}) C_n^+\big),$$

or, denoting the coefficient of Y_j' by D_j, $j = 1, 2, \ldots, n$, the constant on the right side by D_0,

$$D_1 Y_1' + D_2 Y_2' + \cdots + D_n Y_n' = D_0. \tag{13}$$

Note that D_j is an $(N - p_i) \times q_j$ matrix, $j = 1, 2, \ldots, n$, and D_0 is an $(N - p_i) \times p_i$ matrix. Applying the R-orthogonalization procedure to $D_1, \ldots, D_n$ and denoting the resulting R-orthogonal matrices by $E_1, E_2, \ldots, E_n$, respectively, we obtain

$$E_1 Z_1 + E_2 Z_2 + \cdots + E_n Z_n = D_0. \tag{14}$$

The relation between a solution $Z_1, \ldots, Z_n$ of this equation and that of (13) is given, according to (4), by the formula

$$Y_k' = \sum_{j=k}^{n} G_{kj} Z_j$$

where

$$G_{jj} = I \quad (q_j\text{-dimensional}) \qquad j = 1, 2, \ldots, n$$

$$G_{kj} = -E_k^+ (D_{k+1} G_{k+1,j} + D_{k+2} G_{k+2,j} + \cdots + D_{j-1} G_{j-1,j} + D_j)$$

$$k = j-1, j-2, \ldots, 2, 1.$$

It is obvious that every solution $Y_1', \ldots, Y_n'$ of (13) will give the same result X_{ki} in (12) since we have assumed that the solution of the system (10), (11) is unique. Since it does not matter which solution of (14) we choose, we take $Z_j = E_j^+ D_0$, $j = 1, 2, \ldots, n$. So far we considered the index i as a fixed value. Carrying out the procedure described above for every value of i, $1 \leqq i \leqq m$, we obtain the inverse X of the partitioned matrix A.

5. Discussion of the method from practical view-point. First we give an estimation on the number of arithmetical operations involved in the calculations. Of course, this estimation will depend on the procedure used for evaluating of generalized inverses. Using certain generalized inversion algorithm let us denote by $a(p, q)$ and $m(p, q)$ the number of required additive and multiplicative operations, respectively, assumed that we have to do with a $p \times q$ matrix, and consider an $N \times N$ matrix partitioned in the same way as in (9).

A rather tiresome but otherwise trivial counting yields the clumsy upper bounds

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left(a(p_i, q_j)+a(N-p_i, q_j)\right)+\frac{5m+3}{2}N^3+\left(\frac{m+1}{2}q_{\max}-\frac{9m}{2}-2\right)N^2$$

for the additive operations and

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left(m(p_i, q_j)+m(N-p_i, q_j)\right)+\frac{5m+3}{2}N^3+\left(\frac{m+1}{2}q_{\max}-2m-1\right)N^2$$

for the multiplicative operations, where

$$q_{\max}=\max\{q_1, q_2, \ldots, q_n\}.$$

To give a concrete example, let us denote the maximum of $p_1, p_2, \ldots, p_m$ by $p_{\max}$, suppose that $m=n=2$, $p_{\max} \leqq 0.6N$, $q_{\max} \leqq 0.6N$ and agree that the algorithm of J. EGERVÁRY [3] is used for computing of generalized inverses. For this algorithm we have

$$a(p, q) \leqq \frac{10}{3}p^3+p^2q+pq^2+\frac{2}{3}p$$

$$m(p, q) \leqq \frac{10}{3}p^3+p^2q+pq^2+5p^2+\frac{14}{3}p.$$

In this case $11.14N^3$ and $11.14N^3+10.4N^2$ are upper bounds for all additive and multiplicative operations required by the partitioning method, respectively.

It is easy to see that any matrix equation of the form

$$AX=B$$

can be solved by a slightly modified version of the procedure described in Sections 3—4, whenever this equation is solvable. The method can also be used for computing the generalized inverse of a rectangular partitioned matrix A. This is a consequence of the following simple fact which we mention without proof: the generalized inverse A^+ of the matrix A can be written in the form

$$A^+=X_0AY_0$$

where X_0^* and Y_0 are solutions of the equations

$$AA^*X^*=A$$

and

$$A^*AY=A^*,$$

respectively.

Acknowledgement. The author expresses his grateful thanks to Professor P. Rózsa whose valuable suggestions enabled him to generalize his results. The statements of this paper are formulated in their generalized form.

REFERENCES

[1] Д. К. Фаддеев—В. Н. Фаддеева: *Вычислительные методы линейной алгебры.* Физматгиз, Москва 1963.
[2] PENROSE, R.: A generalized inverse for matrices. *Proc. Cambridge Phil. Soc.* **51** (1955) 406—413.
[3] EGERVÁRY, J.: Az inverz mátrix általánosítása. *Mat. Kut. Int. Közl.* **1** N° 3 (1956) 315—324.
[4] HESTENES, M. R.: Relative Self-Adjoint Operators in Hilbert Space, *Pacific J. of Math.* **11** (1961), 1315—1357.

Technical University, Budapest

(Received March 21, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 137—139

REMARK ON THE LAW OF THE ITERATED LOGARITHM

by

G. ALEXITS

Dedicated to the memory of A. Rényi

1. The classical law of the iterated logarithm concerning sequences of independent random variables $\{\xi_n\}$ can be generalized in one direction by substituting the condition of independence by a weaker one (P. RÉVÉSZ [1]). The system $\{\xi_n\}$ is called equinormed strongly multiplicatively orthogonal (ESMS), if the expectations $\mathsf{E}(\xi_\nu)$ satisfy the conditions $\mathsf{E}(\xi_\nu)=0$, $\mathsf{E}(\xi_\nu^2)=1$ for every ν and

$$\mathsf{E}(\xi_{n_1}^{r_1}\xi_{n_2}^{r_2}\cdots\xi_{n_k}^{r_k}) = \mathsf{E}(\xi_{n_1}^{r_1})\mathsf{E}(\xi_{n_2}^{r_2})\cdots\mathsf{E}(\xi_{n_k}^{r_k}) \tag{1}$$

for every collection of different indices $n_1, n_2, \ldots, n_k$, where $r_1, r_2, \ldots, r_k$ equal to 1 or 2. Denoting by $\mathsf{P}(A)$ the probability of the event A, RÉVÉSZ proved the following theorem:

If $\{\xi_n\}$ is a uniformly bounded ESMS, *then*

$$\mathsf{P}\left(\limsup_{n\to\infty}\frac{\xi_1+\xi_2+\cdots+\xi_n}{\sqrt{n\log\log n}}\leqq 7\right)=1.$$

This result seems to be satisfactory apart from the restrictive condition of equinorming, i.e. the left hand side of (1) shall have the value 1 for $r_1=r_2=\cdots=r_k=2$. Indeed, one can show by simple examples that there exist strongly multiplicatively orthogonal systems which can not be equinormed. This circumstance induce us to show in the following that the condition of equinorming is superfluous in the theorem of RÉVÉSZ.

Instead of (1) we introduce the following more comprehensive notion: $\{\xi_n\}$ is called *weakly quasi independent*, if for any finite collection of different indices $N=(n_1, n_2, \ldots, n_k)$ we have

$$\mathsf{E}(\xi_{n_1}^{r_1}\xi_{n_2}^{r_2}\cdots\xi_{n_k}^{r_k}) = K_N\cdot\mathsf{E}(\xi_{n_1}^{r_1})\mathsf{E}(\xi_{n_2}^{r_2})\cdots\mathsf{E}(\xi_{n_k}^{r_k}) \tag{2}$$

where the non-zero constant K_N depends only on the choice of the collection N. We shall prove the following

THEOREM: *Put $\{\zeta_n\}$ a uniformly bounded system of random variables and*

$$\xi_n=\zeta_n-\mathsf{E}(\zeta_n).$$

If $\{\xi_n\}$ is weakly quasi independent, then there exists a constant K such that

$$\mathsf{P}\left(\limsup_{n\to\infty}\frac{\xi_1+\xi_2+\cdots+\xi_n}{\sqrt{n\log\log n}}\leqq K\right)=1.$$

The proof of our theorem follows entirely the trail of RÉVÉSZ's proof. It is clear that the uniform boundedness is necessary in our theorem, if we renounce to any other norming condition. Indeed one can see easily that $\{C_n\}$ being an arbitrary slowly increasing sequence of positive numbers tending to infinite, there exist systems of independent random variables $\{\xi_n\}$ such that $|\xi_n| \leqq C_n$ and

$$\mathsf{P}\left(\limsup_{n\to\infty} \frac{\xi_1+\xi_2+\cdots+\xi_n}{\sqrt{n\log\log n}} = \infty\right) = 1.$$

To this aim it is sufficient to consider, for instance, the system $\{\xi_n\}$ with $\xi_n = C_n r_n(x)$ where $r_n(x)$ denotes the nth Rademacher function. Then our statement follows from the inverse part of the law of iterated logarithm.

2. We turn to the proof of our theorem. Put for abbreviation

$$\eta_n = \sum_{k=1}^{n} \xi_k, \quad \lambda_n = \sqrt{\frac{\log\log n}{n}}, \quad \mu_n = c_1 \log\log n,$$

where c_1 is an appropriate constant. (In the following $c_2, c_3, \ldots$ denote positive constants.) Then $|\xi_n| \leqq 2 \max |\zeta_n| = 2M$, so we have $\lambda_n|\xi_k| \leqq 1$ for all sufficiently large n, hence

$$\mathsf{E}(e^{\lambda_n\eta_n - \mu_n}) = e^{-\mu_n}\mathsf{E}\left(\prod_{k=1}^{n} e^{\lambda_n\xi_k}\right) \leqq e^{-\mu_n}\mathsf{E}\left(\prod_{k=1}^{n}(1+\lambda_n\xi_k+\lambda_n^2\xi_k^2)\right), \tag{3}$$

the last inequality being satisfied because of $e^x \leqq 1+x+x^2$, $|x| \leqq 1$. Executing the multiplication in the product on the right hand side, we get an expression of the following form:

$$\prod_{k=1}^{n}(1+\lambda_n\xi_k+\lambda_n^2\xi_k^2) = 1+S+\sum_{k=1}^{n}\lambda_n^2\xi_k^2+\sum\lambda_n^4\xi_i^2\xi_j^2+$$

$$+\sum\lambda_n^6\xi_i^2\xi_j^2\xi_k^2+\cdots+\sum\lambda_n^{2n}\xi_1^2\xi_2^2\cdots\xi_n^2,$$

where S denotes the sum of all terms containing at least one factor $\lambda_n\xi_\nu$, while in the other sums we have to sum over all terms containing two factors $\xi_i^2\xi_j^2$, three factors $\xi_i^2\xi_j^2\xi_k^2$, etc. Taking in account that $\mathsf{E}(\xi_\nu) = 0$ $(\nu = 1, 2, \ldots, n)$ and $\{\xi_n\}$ satisfies the condition (2), furthermore that $\xi_\nu^2 \leqq 4M^2$, consequently also $\mathsf{E}(\xi_\nu^2) \leqq 4M^2$, we get $\mathsf{E}(S) = 0$ and $\mathsf{E}(\xi_{\nu_1}^2\xi_{\nu_2}^2\ldots\xi_{\nu_k}^2) \leqq (4M^2)^k$, hence

$$\mathsf{E}\left(\prod_{n=1}^{n}(1+\lambda_n\xi_k+\lambda_n^2\xi_k^2)\right) \leqq 1+\binom{n}{1}4\lambda_n^2M^2+\binom{n}{2}(4\lambda_n^2M^2)^2+\cdots+$$

$$+\binom{n}{n}(4\lambda_n^2M^2)^n = (1+4\lambda_n^2M^2)^n.$$

Thus it follows by (3):

$$\mathsf{E}(e^{\lambda_n\eta_n-\mu_n}) \leqq e^{-\mu_n}(1+4\lambda_n^2M^2)^n =$$

$$= e^{-\mu_n}\left\{\left(1+\frac{4M^2\log\log n}{n}\right)^{\frac{n}{\log\log n}}\right\}^{\log\log n} \leqq e^{-\mu_n}\cdot e^{c_2\log\log n} = (\log n)^{c_2-c_1}.$$

If we choose $c_1 > c_2 + 8$, then, for $n_k = [e^{\sqrt[8]{k}}]$, we have

$$\sum_{k=1}^{\infty} \mathsf{E}\,(e^{\lambda_{n_k}\eta_{n_k} - \mu_{n_k}}) < \infty,$$

from which the convergence almost everywhere of the series $\Sigma e^{\lambda_{n_k}\eta_{n_k} - \mu_{n_k}}$ follows by the known theorem of B. LEVI. Hence $\lambda_{n_k}\eta_{n_k} - \mu_{n_k} < 0$ a.e. for sufficiently great k, or in other words

$$\frac{\eta_{n_k}}{\sqrt{n_k \log\log n_k}} \leqq c_1 \quad \text{a. e.}$$

Thus

$$\mathsf{P}\left(\limsup_{k\to\infty} \frac{\eta_{n_k}}{\sqrt{n_k \log\log n_k}} \leqq c_1\right) = 1,$$

and it remains only to show that there exists a constant c_3 such that

$$\mathsf{P}\left(\limsup_{n\to\infty} \max_{n_k < n < n_{k+1}} \frac{\eta_n}{\sqrt{n \log\log n}} \leqq c_3\right) = 1. \tag{4}$$

For the proof of (4) we have to apply a lemma of RÉVÉSZ ([1], Theorem 3. 1. 2.). This states the following: if $\alpha_1, \alpha_2, \ldots, \alpha_n$ are real numbers and $\vartheta_1, \vartheta_2, \ldots, \vartheta_n$ uniformly bounded, weakly quasi independent random variables of mean value zero, then

$$\mathsf{E}\left(\max_{1 \leqq n \leqq m} \left(\sum_{j=1}^{n} \alpha_j \vartheta_j\right)^2\right) \leqq c_4 \log^4 m \left(\sum_{j=1}^{m} \alpha_j^2\right)^2.$$

Put $m = n_{k+1} - n_k$, $\vartheta_j = \xi_{n_k + j}$, $\alpha_j = (n_k \log\log n_k)^{-\frac{1}{2}}$ for $j = n_k + 1, \ldots, n_{k+1}$, then, taking in account that the number of integers between n_k and n_{k+1} is $\leqq c_5 e^{\sqrt[8]{k}} \cdot k^{-\frac{7}{8}}$, we obtain

$$\mathsf{E}\left(\max_{n_k < n < n_{k+1}} \left(\frac{\eta_n - \eta_{n_k}}{\sqrt{n \log\log n}}\right)^4\right) \leqq c_4 \log^4 e^{\sqrt[8]{k+1}} \left(\frac{c_5}{k^{7/8} \log k}\right)^2 < \frac{c_6}{k^{5/4}}.$$

Summing over k, we get

$$\sum_{k=2}^{\infty} \mathsf{E}\left(\max_{n_k < n < n_{k+1}} \left(\frac{\eta_n - \eta_{n_k}}{\sqrt{n \log\log n}}\right)^4\right) < \infty,$$

from which follows

$$\mathsf{P}\left(\sum_{k=2}^{\infty} \max_{n_k < n < n_{k+1}} \left(\frac{\eta_n - \eta_{n_k}}{\sqrt{n \log\log n}}\right)^4 < \infty\right) = 1,$$

and this is equivalent to (4).

LITERATURE

[1] RÉVÉSZ, P.: *The laws of large numbers*, Academic Press and Akadémiai Kiadó, 1967.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received April 16, 1969)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 141—150

AN APPROXIMATION THEORETICAL STUDY OF THE STRUCTURE OF REAL FUNCTIONS

by

G. FREUD

I. Introduction

Let $\varphi(h)$ $(0<h\leqq\pi)$ be a nonvanishing, nondecreasing continuous function with $\varphi(+0)=0$ and

$$\varphi(2h)\leqq 2\varphi(h) \quad \left(0<h\leqq\frac{\pi}{2}\right). \tag{1}$$

We assume further (except in part II) that there exists a $B>1$ so that*

$$1<\Theta<\underline{\lim_{h\to 0}}\frac{\varphi(Bh)}{\varphi(h)}\leqq\overline{\lim_{h\to 0}}\frac{\varphi(Bh)}{\varphi(h)}<\Theta^{-1}B<B. \tag{2}$$

In part II we shall need the first part of this inequality only.

We consider the set $C(\varphi)$ of continuous 2π-periodic functions $f(x)$ which satisfy

$$|f(x+h)-f(x)|\leqq A(f)\varphi(|h|) \tag{3}$$

with a constant $A(f)$ depending neither on x nor on h.

We consider the following three sets characterized by the fact, that on points of these sets — roughly to speak — the local continuity behaviour of $f(x)$ is better than the global one.

$m_0(f)$ is the set of points x for which we have uniformly with respect to x

$$f(x+h)-f(x)=o\{\varphi(h)\} \quad [x\in m_0(f)] \tag{4}$$

for $h\to 0$;

$m_1(f)$ is the set of points x for which we have uniformly with respect to x

$$f(x+h)-f(x-h)=o\{\varphi(h)\} \qquad [x\in m_1(f)] \tag{5}$$

for $h\to 0$;

$m_2(f)$ is the set of points x for which we have uniformly with respect to x

$$f(x+h)+f(x-h)-2f(x)=o\{\varphi(h)\} \qquad [x\in m_2(f)]. \tag{6}$$

Clearly (4) implies both (5) and (6), so that

$$m_0(f)\subseteqq m_1(f), \quad m_0(f)\subseteqq m_2(f). \tag{7}$$

Let $\tilde{f}(x)$ be the harmonic conjugate of $f(x)$. As a consequence of a generalization of PRIVALOFF's theorem, (2), (3) and $f\in C(\varphi)$ imply that $\tilde{f}\in C(\varphi)$ (see N. K. BARI

* In parts III, IV and V all consequences of (1) which we need could be deduced of the second half of (2); in part II we make essential use of condition (1) but we do not need the second half of (2).

[2]). *We consider beside*

$$m_0(f), \quad m_1(f), \quad m_2(f)$$

the sets

$$m_0(\tilde{f}), \; m_1(\tilde{f}), \; m_2(\tilde{f})$$

and we are going to prove that *all these six sets are measurable and each two of these six sets differ by a set of measure zero at most.*

In part II we prove the measurability and equivalence of $m_0(f)$ and $m_1(f)$, in part III we prove the equivalence of $m_0(f)$ and $m_2(f)$ and in part IV the equivalence of $m_0(f)$ and $m_0(\tilde{f})$. From this clearly the equivalence of any two of the six sets $m_0(f)$, $m_1(f)$, $m_2(f)$; $m_0(\tilde{f})$, $m_1(\tilde{f})$, $m_2(\tilde{f})$ will follow as a consequence of the following pattern:

$$\begin{array}{ccc} m_0(f) & \text{———————} & m_0(\tilde{f}) \\ \swarrow \quad \searrow & & \swarrow \quad \searrow \\ m_1(f) \quad m_2(f) & & m_1(\tilde{f}) \quad m_2(\tilde{f}) \end{array}$$

Finally, in part IV we extend our result for the case that (2) is satisfied only on a subinterval $[a, b]$ of $[-\pi, +\pi]$.

II. Equivalence of the sets $m_0(f)$ and $m_1(f)$

In this section we use only the first part of condition (2).

We consider JACKSON's trigonometric approximation polynomials

$$J_n(f; x) = \int_{-\pi}^{+\pi} K_n(x-t) f(t)\, dt, \tag{8}$$

with

$$K_n(t) = \gamma_n \left(\frac{\sin \frac{nt}{2}}{\sin \frac{t}{2}} \right)^4, \quad \gamma_n^{-1} = \int_{-\pi}^{+\pi} K_n(t)\, dt. \tag{9}$$

It is well known that both $K_n(t)$ and $J_n(f; x)$ are trigonometric polynomials of order $2n$ (resp. order $2n$ at most), and

$$0 \leqq K_n(t) \leqq c_1 \min\left(n, \frac{1}{n^3 t^4}\right). \tag{10}$$

Further, an easy computation shows that*

$$|K_n'(t)| \leqq c_2 \min\left(n^2, \frac{1}{n^2 t^4}\right). \tag{11}$$

We observe that as a consequence of (1) and the monotonity of $\varphi(h)$

$$\varphi(h) \leqq \varphi(n^{-1}) + 2nh\varphi(n^{-1}) \qquad (h>0, \quad n=1, 2, \ldots) \tag{12}$$

holds true.

* We are denoting by $c_1, c_2, \ldots$ positive absolute constants.

From (8) and (9) we obtain taking into account the periodicity of K_n and f as well as the evenness of K_n

$$f(x)-J_n(f;x)=\frac{1}{2}\int_{-\pi}^{+\pi}[f(x+t)-f(x-t)-2f(x)]K_n(t)\,dt. \tag{13}$$

We obtain further

$$J_n'(f;x)=\frac{1}{2}\int_{-\pi}^{+\pi}[f(x+t)-f(x-t)]K_n'(t)\,dt. \tag{14}$$

In what follows let $f\in C(\varphi)$. We conclude from (13), (10), (3) and (12) that

$$|f(x)-J_n(f;x)| \leqq c_3 A(f)\varphi(n^{-1}) \tag{15}$$

and from (6) in place of (3) we obtain*

$$f(x)-J_n(f;x)=o_x\{\varphi(n^{-1})\} \qquad [x\in m_2(f)]. \tag{16}$$

Further, we have by (14), (11), (3) and (12) resp. (14), (11), (5) and (12)

$$|J_n'(f;x)|\leqq c_4 A(f)n\varphi(n^{-1}) \tag{17}$$

resp.

$$J_n'(f;x)=o_x\{n\varphi(n^{-1})\} \qquad [x\in m_1(f)]. \tag{18}$$

Let $M_1(f)$ be the set of all points where the o_x-relation (18) is satisfied and $M_2(f)$ be the set of all points where the o_x-relation (16) holds true. (16) and (18) imply

$$m_1(f)\subseteqq M_1(f) \quad \text{and} \quad m_2(f)\subseteqq M_2(f). \tag{19}$$

The sets $M_1(f)$ and $M_2(f)$ are as a consequence of their definitions $F_{\sigma\delta}$ sets and so they are measurable.

THEOREM 1. *Almost every point of $M_1(f)$ belongs to $m_0(f)$.*

REMARK. We know that

$$m_0(f)\subseteqq m_1(f)\subseteqq M_1(f)$$

(see (7) and (19)), so that also $m_0(f)$ and $m_1(f)$ differ in a set of measure zero at most. We recall again the fact, that only the assumption $1<\Theta<\varliminf_{h\to 0}\frac{\varphi(Bh)}{\varphi(h)}$ is used; the second part of (2) is not necessary here.

* The relation

$$\mathscr{F}(n,x)=o_x\{\mathscr{G}(n)\} \qquad (x\in E)$$

means — as usual — that for each fixed value of $x\in E$ we have

$$\lim_{n\to\infty}\frac{\mathscr{F}(n,x)}{\mathscr{G}(n)}=0$$

but this limit need not exist uniformly with respect to x.

PROOF. Let δ be an arbitrary small positive number. By EGOROFF's theorem there exists a measurable set $M_1(\delta;f)$ so that

$$M_1(\delta,f)\subseteqq M_1(f), \quad |M_1(\delta,f)| > |M_1(f)|-\delta \tag{20}$$

and there is a numerical sequence $\varepsilon_n=\varepsilon_n(\delta)\to 0$ so that

$$|J_n'(f;x)|\leqq\varepsilon_n(\delta)n\varphi(n^{-1}) \qquad [x\in M_1(\delta,f)]. \tag{21}$$

We keep from the set $M_1(\delta,f)$ only the points of density, i.e. the points x for which

$$\lim_{h\to+0}\frac{1}{2h}|[x-h,\,x+h]\cap M_1(\delta,f)|=1 \tag{22}$$

is valid. A well known theorem of H. LEBESGUE states, that the omitted point set has measure zero. In what follows we call $M_1(\delta,f)$ the remaining point set; clearly (20), (21) and (22) are satisfied for the points of this $M_1(\delta,f)$.

We fix a point $x\in M_1(\delta,f)$ and let

$$\mu(h)=[x-h,\,x+h]\setminus M_1(\delta,f) \qquad (h>0). \tag{23}$$

We conclude from (22) that for $h\leqq\delta(\varepsilon)$ we have

$$|\mu(h)|\leqq\varepsilon h. \tag{24}$$

Now let $|h|\leqq\delta(\varepsilon)$. Using (15) we obtain for $|h|\leqq\delta(\varepsilon)$, and taking (24), (17) and (21) in consideration

$$|f(x+h)-f(x)|\leqq 2c_3A(f)\varphi(n^{-1})+\left|\int_x^{x+h}J_n'(f;t)\,dt\right|\leqq$$

$$\leqq 2c_3A(f)\varphi(n^{-1})+\int_{\mu(|h|)}|J_n'(f;t)|\,dt+\int_{[x,x+h]\cap M_1(\delta,f)}|J_n'(f,t)|\,dt\leqq \tag{25}$$

$$\leqq 2c_3A(f)\varphi(n^{-1})+\varepsilon|h|\,c_4A(f)n\varphi(n^{-1})+\varepsilon_n(\delta)n\varphi(n^{-1})|h|.$$

Up to now n was an arbitrary integer, we are choosing now an n satisfying

$$1\leqq\varepsilon^{\frac{1}{2}}|h|n<2; \tag{26}$$

this is certainly possible provided that $|h|$ is small enough. We observe that for fixed ε we have $n\to\infty$ if $|h|\to 0$. Let $m(\varepsilon)$ be the integer satisfying

$$B^{m(\varepsilon)}<\varepsilon^{-\frac{1}{2}}\leqq B^{m(\varepsilon)+1}, \tag{27}$$

we observe that $m(\varepsilon)\to\infty$ if $\varepsilon\to 0$.

Let $|h|$ be so small that

$$\varphi(B\eta)>\Theta\varphi(\eta) \quad \text{for} \quad |\eta|\leqq|h|,$$

so that

$$\varphi(n^{-1})\leqq\varphi\left(\varepsilon^{\frac{1}{2}}|h|\right)\leqq\varphi\left(B^{-m(\varepsilon)}|h|\right)\leqq\Theta^{-m(\varepsilon)}\varphi(|h|). \tag{28}$$

We have as a consequence of (25), (26) and (28)

$$|f(x+h)-f(x)| \leqq [2c_3 A(f)\Theta^{-m(\varepsilon)} + 2c_4 A(f)\varepsilon^{\frac{1}{2}} + 2\varepsilon^{-\frac{1}{2}}\varepsilon_n(\delta)]\varphi(|h|)$$

for sufficiently small values of $|h|$.

The factor in the brackets can be made arbitrary small, if we choose first ε small enough and after that we choose $|h|$ so small (and consequently n so large) that $\varepsilon_n(\delta) < \varepsilon$. In this way we proved

$$f(x+h)-f(x) = o_x\{\varphi(|h|)\} \qquad [x \in M_1(\delta, f)]$$

i.e.

$$M_1(\delta, f) \subseteqq m_0(f). \tag{29}$$

Taking (20) in account, we see that the $m_0(f)$ overlaps the measurable set $M_1(\delta, f)$ and is overlapped by the measurable set $M_1(f)$, and the difference of the measure of these two sets is less than the arbitrary positive number δ. We conclude that $m_0(f)$ is measurable and

$$|m_0(f)| = |M_1(f)|. \tag{30}$$

Using (7) and (19) we have

$$m_0(f) \subseteqq m_1(f) \subseteqq M_1(f) \tag{31}$$

and from (30) and (31) follows, that $m_1(f)$ is measurable and

$$|m_1(f) \setminus m_0(f)| = 0$$

q.e.d.

III. Equivalence of $m_0(f)$ and $m_2(f)$

From now on we assume the validity of both parts of (2).

Let $M_2(f)$ be the set of all points x for which

$$f(x) - J_n(f; x) = o_x\{\varphi(n^{-1})\} \qquad [x \in M_2(f)] \tag{32}$$

is satisfied. Clearly $M_2(f)$ is an $F_{\sigma\delta}$-set, and it is a fortiori measurable. By EGOROFF's theorem there exists for each $\delta > 0$ a measurable set $M_2(\delta, f)$ so that

$$M_2(\delta, f) \subseteqq M_2(f), \quad |M_2(\delta, f)| > |M_2(f)| - \delta \tag{33}$$

and (32) holds uniformly on $M_2(\delta, f)$, i.e. there exists a sequence $\varepsilon_n \searrow 0$ for which

$$|f(x) - J_n(f; x)| < \varepsilon_n \varphi(n^{-1}) \qquad [x \in M_2(\delta, f)] \tag{34}$$

holds true. We can omit by LEBESGUE's theorem from $M_2(\delta, f)$ all points which are no points of density, so we can assume that for all points of $M_2(\delta, f)$

$$\lim_{h \to 0} \frac{1}{2h} |[x-h, x+h] \cap M_2(\delta, f)| = 1 \qquad [x \in M_2(\delta, f)] \tag{35}$$

is valid. Now let $1 > \varepsilon > 0$ arbitrary and x a fixed point of $M_2(\delta, f)$. As a consequence of (35) for sufficiently small h none of the sets $[x-|h|, x-(1-\varepsilon)|h|] \cap M_2(\delta, f)$ and $[x+(1-\varepsilon)|h|, x+|h|] \cap M_2(\delta, f)$ is empty; for $h < 0$ let t be a point of the first

of these sets and for $h>0$ let t be a point of the second mentioned non-empty set, in both cases we have

$$t\in M_2(\delta,f),\quad |x+h-t|\leqq\varepsilon|h|,\quad |x-t|\leqq h. \tag{36}$$

We obtain now using (36), (3), (34) and (17)

$$|f(x+h)-f(x)|\leqq|f(x+h)-f(t)|+|f(t)-J_n(f;t)|+|f(x)-J_n(f;x)|+ \tag{37}$$

$$+|J_n(f;t)-J_n(f;x)|\leqq A(f)\varphi(\varepsilon|h|)+2\varepsilon_n\varphi(n^{-1})+c_4A(f)n\varphi(n^{-1})|h|.$$

Now let $|h|$ be so small that there exists an integer n with

$$\varepsilon^{\frac{1}{2}}\leqq n|h|<2\varepsilon^{\frac{1}{2}}; \tag{38}$$

we obtain from (2), (27) and (28) for sufficiently small $|h|$

$$\varphi(n^{-1})\leqq\varphi\left(\frac{|h|}{\varepsilon^{1/2}}\right)\leqq\varphi\left(|h|B^{m(\varepsilon)+1}\right)\leqq B^{m(\varepsilon)+1}\Theta^{-m(\varepsilon)-1}\varphi(|h|)\leqq B\varepsilon^{-\frac{1}{2}}\Theta^{-m(\varepsilon)}\varphi(|h|)$$

further we have as in part II

$$\varphi(\varepsilon|h|)\leqq\varphi(\varepsilon^{\frac{1}{2}}|h|)<\Theta^{-m(\varepsilon)}\varphi(|h|)$$

so that we obtain from (37)

$$|f(x+h)-f(x)|\leqq[A(f)(1+2c_4B)\Theta^{-m(\varepsilon)}+2B\varepsilon^{-\frac{1}{2}}\varepsilon_n]\varphi(|h|). \tag{39}$$

Taking first ε small enough and after that n large enough (which is by (38) the same as taking $|h|$ sufficiently small) the factor on the right hand side of (39) can be made arbitrarily small, so that

$$f(x+h)-f(x)=o_x\{\varphi(|h|)\}\qquad[x\in M_2(\delta,f)] \tag{40}$$

so that $M_2(\delta,f)\subseteqq m_0(f)$ and from (13), (6) and (7)

$$m_0(f)\subseteqq m_2(f)\subseteqq M_2(f). \tag{41}$$

We obtained

$$M_2(\delta,f)\subseteqq m_0(f)\subseteqq M_2(f).$$

Comparing this with (33) and using the fact, that $\delta>0$ was arbitrary we have

$$|M_2(f)\setminus m_0(f)|=0. \tag{42}$$

We now conclude from (41) that $m_2(f)$ is measurable and

$$|m_2(f)\setminus m_0(f)|=0,$$

in this way we proved

THEOREM 2. *Almost all points of $m_2(f)$ belong to $m_0(f)$.*

IV. Equivalence of $m_0(f)$ and $m_0(\tilde{f})$

LEMMA. *The second part of* (2) *implies, that*

$$\delta \int_{\delta}^{\pi} \frac{\varphi(t)}{t^2}\, dt < \Gamma(\varphi)\varphi(\delta) \quad (\delta > 0), \tag{43}$$

where $\Gamma(\varphi) > 0$ *is dependent only on the choice of* $\varphi(\delta)$.

PROOF. Let Δ be so small that

$$\frac{\varphi(Bt)}{\varphi(t)} \leqq \Theta^{-1} B \quad (0 \leqq t \leqq B\Delta) \tag{44}$$

let further $\delta < \Delta$ and

$$B^m \delta \leqq \Delta < B^{m+1}\delta. \tag{45}$$

We conclude

$$\delta \int_{\delta}^{\Delta} \frac{\varphi(t)}{t^2}\, dt \leqq \delta \sum_{r=0}^{m} \int_{B^r\delta}^{B^{r+1}\delta} \frac{\varphi(t)}{t^2}\, dt = \delta \sum_{r=0}^{m} B^{-r} \int_{\delta}^{B\delta} \frac{\varphi(B^r h)}{h^2}\, dh \leqq$$

$$\leqq \delta \sum_{r=0}^{m} B^{-r}(B\Theta^{-1})^r \int_{\delta}^{B\delta} \frac{\varphi(h)}{h^2}\, dh \leqq \delta \sum_{r=0}^{m} \Theta^{-r}(B-1)\delta \cdot \delta^{-2}\varphi(B\delta) \leqq$$

$$\leqq \sum_{r=0}^{m} \Theta^{-r}(B-1)\Theta^{-1} B\varphi(\delta) = \frac{B(B-1)}{\Theta - 1}\varphi(\delta) = \Gamma_1(\varphi)\varphi(\delta)$$

and

$$\delta \int_{\Delta}^{\pi} \frac{\varphi(t)}{t^2}\, dt \leqq \Gamma_2(\varphi)\delta$$

so that we have for $\delta \leqq \Delta$

$$\delta \int_{\delta}^{\pi} \frac{\varphi(t)}{t^2}\, dt \leqq \Gamma_1(\varphi)\varphi(\delta) + \Gamma_2(\varphi)\delta \tag{46}$$

and choosing $\Gamma_2(\delta)$ sufficiently large, this will be valid for all $0 \leqq \delta \leqq \pi$.

For $\delta < \Delta$ we obtain from (45)

$$\varphi(\delta) \geqq \varphi(B^{-m-1}\Delta) \geqq (\Theta B^{-1})^{m+1}\varphi(\Delta) > B^{-m-1}\varphi(\Delta) \geqq B\Delta^{-1}\varphi(\Delta)\delta.$$

From this we conclude the existence of a $\Gamma_3(\varphi) > 0$ so that

$$\varphi(\delta) \geqq \Gamma_3(\varphi)\delta \qquad (0 < \delta \leqq \pi). \tag{47}$$

From (46) and (47) follows the validity of (43), q.e.d.

Let us denote by $\sigma_n(f; x)$ the FEJÉR means of the Fourier series of $f(x)$.

We have than

$$\sigma_n(f; x) - f(x) = \int_{0}^{\pi} [f(x+t) + f(x-t) - 2f(x)] K_n^{(1)}(t)\, dt \tag{48}$$

with

$$0 \leqq K_n^{(1)}(t) \leqq c_5 \min(n, n^{-1}t^{-2}). \tag{49}$$

We conclude that if (2) is satisfied, then as a consequence of (43) we have for $f \in C(\varphi)$ and $x \in m_2(f)$ [see (6)]

$$\sigma_n(f; x) - f(x) = o_x\{\varphi(n^{-1})\} \qquad [x \in m_2(f)]. \tag{50}$$

Let $\sigma_n(\tilde{f}; x)$ be the FEJÉR sum of the conjugate function $\tilde{f}(x)$.

We apply — following an idea of G. ALEXITS [1] — the identity

$$\sigma_n'(\tilde{f}; x) = n\sigma_n(f; x) - \frac{2}{n+1}\sum_{r=0}^{n-1}{}' (r+1)\sigma_r(f; x) = \tag{51}$$

$$= n[\sigma_n(f; x) - f(x)] - \frac{2}{n+1}\sum_{r=0}^{n-1}{}' (r+1)[\sigma_r(f; x) - f(x)].$$

We have by (2) $n\varphi(n^{-1}) \to \infty$, so that from (50) and (51) we obtain

$$\sigma_n'(\tilde{f}; x) = o_x(1)\left[n\varphi(n^{-1}) + \frac{1}{n}\sum_{r=0}^{n-1}{}' (r+1)\varphi(r^{-1})\right] \qquad [x \in m_2(f)].$$

We get from (12) applied for $h = r^{-1}$ $(r = 1, 2, \ldots, n-1)$

$$\frac{1}{n}\sum_{r=0}^{n-1}{}' (r+1)\varphi(r^{-1}) \leqq c_6\varphi(n^{-1})$$

so that

$$\sigma_n'(\tilde{f}; x) = o_x\{n\varphi(n^{-1})\} \qquad [x \in m_2(f)]. \tag{52}$$

As a consequence of the PRIVALOFF—BARI theorem (see N. K. BARI [2]) we have $\tilde{f} \in C(\varphi)$ so that we have from (48), (49), (43) and (3)

$$|\sigma_n(\tilde{f}; x) - \tilde{f}(x)| \leqq c_6 A(\tilde{f})\varphi(n^{-1}) \qquad (x \in [-\pi, +\pi]).$$

Let $M_1^*(\tilde{f})$ be the set of all points x for which (52) holds, so that by (52)

$$m_2(f) \subset M_1^*(\tilde{f}). \tag{53}$$

Repeating the same argument as in the proof of Theorem 1 but replacing $J_n(f; x)$ by $\sigma_n(f; x)$ we get that almost all points of $M_1^*(\tilde{f})$ belong to $m_0(\tilde{f})$. By (53) and (7) a fortiori almost all points of $m_0(f)$ belong to $m_0(\tilde{f})$, i.e.

$$|m_0(f) \setminus m_0(\tilde{f})| = 0.$$

By symmetry we have

$$|m_0(\tilde{f}) \setminus m_0(f)| = 0.$$

We proved in this way the validity of the following

THEOREM 3. *We have for $f \in C(\varphi)$ under condition* (2)

$$|[m_0(f) \setminus m_0(\tilde{f})] \cup [m_0(\tilde{f}) \setminus m_0(f)]| = 0.$$

In our last remark the localization of our results is concerned. Let us consider $f(x)$

as defined on an interval $[a, b]$, $b-a < 2\pi$ only. We extend $f(x)$ to a 2π-periodic continuous function by defining it as a linear function in $[b, a+2\pi]$.

If $f(x)$ satisfied (3) for $x, x+h \in [a, b]$ the extended function also satisfies (3) with probably an other constant $A(f)$ [see (47)]. In this way we get

THEOREM 4. *Let $f(x)$ satisfy* (3) *for $x, x+h \in [a, b]$ where $\varphi(h)$ satisfies* (2). *Denoting $m_0(f; a, b)$, $m_1(f; a, b)$ resp. $m_2(f; a, b)$ the sets of points $x \in (a, b)$ for which* (4), (5) *resp.* (6) *hold, then no two of these three sets differ by more than a set of measure zero.*

Let now $\mathcal{F}(x)$ be an arbitrary $\mathfrak{L}$-integrable 2π-periodic function which satisfies (2) for $x, x+h \in [a, b]$. We consider the 2π-periodic function $f(x)$ which coincides wiht $\mathcal{F}(x)$ on $[a, b]$, and is linear on $[b, a+2\pi]$. In the interval $[a, b]$, where $\mathcal{F}(x)-f(x)$ vanishes, the function $\tilde{\mathcal{F}}(x)-\tilde{f}(x)$ is analytic. We conclude that the sets $m_0(\tilde{f}; a, b)$ and $m_0(\tilde{\mathcal{F}}; a, b)$ coincide. From Theorem 3 we know that almost all points of $m_0(\tilde{\mathcal{F}}; a, b) = m_0(f; a, b)$ belong to $m_0(\tilde{f}; a, b) = m_0(\tilde{\mathcal{F}}; a, b)$, i.e.

$$|m_0(\mathcal{F}; a, b) \setminus m_0(\tilde{\mathcal{F}}; a, b)| = 0. \tag{54}$$

As a consequence of the PRIVALOFF—BARI theorem, we have $\tilde{f} \in C(\varphi)$, so that by $\tilde{\mathcal{F}} = \tilde{f} + (\tilde{\mathcal{F}} - \tilde{f})$ and the analyticity of $\tilde{\mathcal{F}} - \tilde{f}$ in (a, b) we have for an arbitrary small $\delta > 0$

$$|\tilde{\mathcal{F}}(x+h) - \tilde{\mathcal{F}}(x)| \leqq A(\delta; \mathcal{F})\varphi(\delta) \quad (x, x+h \in [a+\delta, b-\delta]).$$

Applying (54) in the opposite direction we obtain

$$|m_0(\tilde{\mathcal{F}}; a+\delta, b-\delta) \setminus m_0(\mathcal{F}; a+\delta, b-\delta)| =$$

$$= |m_0(\tilde{\mathcal{F}}; a+\delta, b-\delta) \setminus m_0(\mathcal{F}; a, b)| = 0 \quad (\delta > 0).$$

Inserting for δ a null-sequence, we get

$$|m_0(\tilde{\mathcal{F}}; a, b) \setminus m_0(\mathcal{F}; a, b)| = 0. \tag{55}$$

From (54) and (55) we get

THEOREM 5. *Let $\mathcal{F}(x)$ be a 2π-periodic $\mathfrak{L}$-integrable function, satisfying* (3) *for $x, x+h \in [a, b]$ with $\varphi(h)$ for which* (2) *holds; then*

$$|[m_0(\mathcal{F}; a, b) \setminus m_0(\tilde{\mathcal{F}}; a, b)] \cup [m_0(\tilde{\mathcal{F}}; a, b) \setminus m_0(\mathcal{F}; a, b)]| = 0.$$

Combining Theorems 4 and 5, we get that provided the conditions of Theorem 5 are satisfied, no two of the six sets $m_0(\mathcal{F}; a, b)$, $m_1(\mathcal{F}; a, b)$, $m_2(\mathcal{F}; a, b)$, $m_0(\tilde{\mathcal{F}}; a, b)$, $m_1(\tilde{\mathcal{F}}; a, b)$ and $m_2(\tilde{\mathcal{F}}; a, b)$ differ by more than a set of measure zero (see the pattern at the end of Chapter I).

REFERENCES

[1] ALEXITS, G.: Sur l'ordre de grandeur de l'approximation d'une fonction par les moyennes de sa série de Fourier (Hungarian with French summary). *Mat.-Fiz. Lapok* **48** (1941) 410—422.

[2] Н. К. БАРИ.: О наилучшем приближении тригонометрическими полиномами двух сопряженных функций. *Известия Академии НАУК СССР Серия математическая* **19** (1955), 285—302.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received April 16, 1969)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 151—160

ON AN ESTIMATION PROBLEM OF ALEXITS

by

F. MÓRICZ

1. Let $\{\varphi_\nu(x)\}$ $(\nu = 0, 1, \ldots)$ be an orthonormal system in $[0, 1]$. We shall consider orthogonal series

$$\sum_{\nu=0}^{\infty} c_\nu \varphi_\nu(x), \tag{1}$$

where $\{c_\nu\}$ is a sequence of real numbers satisfying

$$\sum_{\nu=0}^{\infty} c_\nu^2 < \infty. \tag{2}$$

It is known that the nth *Euler mean* of a sequence $\{s_k\}$ is defined as follows:

$$t_n = \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} s_k \qquad (n = 0, 1, \ldots).$$

(In detail, see, for example, KNOPP [4].) A series $\sum_0^\infty c_\nu$ with the kth partial sums s_k is called *summable by the Euler method* (in abbreviation "*E-summable*") to s if $\lim_{n\to\infty} t_n = s$. The Euler summation procedure is permanent, i.e., the (ordinary) convergence of $\{s_k\}$ always implies the Euler summability of $\{s_k\}$ to the same limit.

MEDER (see [5], Theorem 6., p. 146.) gave necessary and sufficient conditions for the almost everywhere (in abbreviation "a.e.") Euler summability of the series (1). More exactly, he proved the following theorem:

The orthonormal series (1) *satisfying* (2) *is E-summable a.e. if and only if*
(i) *the series* (1) *is* $(C, 1)$*-summable a.e., and*

$$\text{(ii)} \qquad \lim_{n\to\infty} \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} \frac{1}{k+1} \sum_{\nu=0}^{k} \nu c_\nu \varphi_\nu(x) = 0 \text{ [1]} \quad a.e.$$

2. ALEXITS raised the following problem: does, under the condition (2), the latter relation hold a.e. or not? We answer this question negatively by giving a counter example.

Our result reads as follows:

[1] The sum on the left-hand side is the nth Euler mean of the $(C, 1)$-means of the terms $\nu c_\nu \varphi_\nu(x)$.

THEOREM. *There exist an orthonormal system* $\{\varphi_\nu(x)\}$ *in* $[0, 1]$ *and a sequence* $\{c_\nu\}$ *of real numbers such that*

$$\sum_{\nu=2}^{\infty} c_\nu^2 (\log\log \nu)^2 < \infty, \tag{3}$$

and such that

$$\overline{\lim_{n\to\infty}} \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} \frac{1}{k+1} \sum_{\nu=0}^{k} \nu c_\nu \varphi_\nu(x) = +\infty \tag{4}$$

holds a.e.

We note that the orthogonal series occurring in our theorem is, on the one hand, $(C, 1)$-summable a.e., and on the other hand, not E-summable a.e. The former assertion follows from (3) applying a theorem of MENCHOFF—KACZMARZ [2], and MEDER's theorem cited above, owing to (4), implies the latter one.

3. Before proving our theorem, let us recall some properties of binomial coefficients. Among others, we shall use the following: if the positive integer N is fixed then

$$\lim_{n\to\infty} \frac{1}{2^n} \sum_{k=0}^{N} \binom{n}{k} = 0, \tag{5}$$

and

$$\lim_{n\to\infty} \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} \frac{1}{k+1} = 0. \tag{6}$$

To show these, it is sufficient to observe that the sums in question are the Euler means of sequences tending (in the sense of ordinary convergence) to 0.

We also need a deeper result that is known in probability theory as the MOIVRE—LAPLACE formula. We state it in the form of a lemma. (See, e.g., [7], p. 129.)

LEMMA 1. *Let* a, b, p *and* q *be real numbers,* $a<b$, $0<p<1$, $q = 1-p$. *Then*

$$\lim_{n\to\infty} \sum \binom{n}{k} p^k q^{n-k} = G(b) - G(a),$$

where the sum Σ *is extended over every integer* k *for which*

$$a \leqq \frac{k-np}{\sqrt{npq}} \leqq b,$$

and

$$G(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{x^2}{2}}\, dx \qquad (-\infty < y < \infty).^2$$

Choosing, in particular, $p = \frac{1}{2}$ we obtain that

$$\lim_{n\to\infty} \frac{1}{2^n} \sum_{n+a\sqrt{n} \leqq 2k \leqq n+b\sqrt{n}} \binom{n}{k} = G(b) - G(a). \tag{7}$$

[2] I.e. $G(y)$ denotes the (normalized) Gaussian distribution function.

4. To construct our counter example we shall use the so-called MENCHOFF's functions [6] which are very common in the investigation of the convergence behaviour of orthogonal series.

LEMMA 2. *Let $p \geqq 2$ be an integer and $I=[u, v]$ an arbitrary interval. Then there exists in $[u, v]$ a system $\{f_l(p, I; x)\}$ of $2p$ step functions, orthogonal to one another, and with the following properties:*

1° $$\int_u^v f_l^2(p, I; x)\, dx = v-u.$$

2° *For every point x of the half open interval*

$$F(I) = \left[u+\frac{2(v-u)}{5}, u+\frac{3(v-u)}{5}\right)$$

there exists an index $m(x)$, $p \leqq m(x) < 2p$, depending on x such that the functions

$$f_1(p, I; x), \ldots, f_{m(x)}(p, I; x)$$

are positive and the inequalities

$$\sum_{l=1}^{m(x)-\mu} f_l(p, I; x) \geqq K\sqrt{p}\log p \tag{8}$$

hold for $\mu = 0, 1, \ldots, [\sqrt{p}]$ [3], *where K is a positive constant.*

In the sequel $K, K_1, K_2, \ldots$ will denote positive constants.

PROOF of Lemma 2. This lemma, with $\sum_{l=1}^{m(x)}$ instead of $\sum_{l=1}^{m(x)-\mu}$ in (8), is due to KACZMARZ [3] who refined MENCHOFF's original proof [6]. That is why we only sketch the proof.

First of all let us choose $I=[0, 5]$ and define in $[0, 4)$ a system of $2p$ functions as follows: for $l=1, 2, \ldots, 2p$ set

$$f_{p,l}(t) = \frac{1}{k-p-l-\frac{1}{2}} \quad \text{if} \quad t \in \left[\frac{k-1}{p}, \frac{k}{p}\right),$$

where k assumes the values $1, 2, \ldots, 4p$. The functions $f_{p,l}(t)$ possess the following property: for every point $x \in [2, 3)$ there exists an integer $p \leqq m = m(x) < 2p$ such that

$$\sum_{i=1}^{m-\mu} f_{p,l}(t) \geqq K_1 \log p, \tag{9}$$

[3] $[\alpha]$ denotes the greatest integer not exceeding α.

where μ takes the values $0, 1, \ldots, [\sqrt{p}]$. In fact, there is an m such that $t \in \left[\frac{p+m}{p}, \frac{p+m+1}{p}\right)$ and then it follows from the definition that

$$\sum_{l=1}^{m-\mu} f_{p,l}(t) = \sum_{l=1}^{m-\mu} \frac{1}{p+m+1-p-l-\frac{1}{2}} =$$

$$= \sum_{i=\mu+1}^{m} \frac{1}{i-\frac{1}{2}} \geqq \sum_{i=[\sqrt{p}]+1}^{p} \frac{1}{i-\frac{1}{2}} \geqq K_1 \log p.$$

Now we put

$$\delta_{i,l} = \delta_{l,i} = \int_0^4 f_{p,i}(t) f_{p,l}(t)\, dt,$$

and then continue $f_{p,l}(t)$ on $[0, 5]$ so that we divide $[4, 5]$ into $2p(2p-1)$ equal parts, denote the single subintervals by $J_{i,l}$ $(1 \leqq i, l \leqq 2p, i \neq l)$ and define for $t \in [4, 5]$ the values of $f_{p,l}(t)$ as follows:

$$f_{r,l}(t) = \begin{cases} \sqrt{p(2p-1)|\delta_{i,l}|} & \text{if} \quad t \in J_{l,i}\,, \\ -\sqrt{p(2p-1)|\delta_{i,l}|}\,\operatorname{sign}\delta_{i,l} & \text{if} \quad t \in J_{i,l}\,, \\ 0 \text{ at the remaining points of } [4,5]. \end{cases}$$

It is clear that the functions $f_{p,l}(t)$ are orthogonal to each other in $[0, 5]$, and a simple calculation shows that

$$\frac{K_2}{p} \leqq \int_0^5 f_{p,l}^2(t)\, dt \leqq \frac{K_3}{p} \qquad (l = 1, 2, \ldots, 2p). \tag{10}$$

By norming the functions $f_{p,l}(t)$ and denoting the normed functions by $\bar{f}_{p,l}(t)$ we obtain from (9) and (10) that for $t \in [2, 3)$

$$\sum_{l=1}^{m-\mu} \bar{f}_{p,l}(t) \geqq K_4 \sqrt{p} \log p \tag{11}$$

holds for $\mu = 0, 1, \ldots, [\sqrt{p}]$.

Now we proceed from the interval $[0, 5]$ to the interval $I = [u, v]$ by means of the linear transformation $x = u + t(v-u)$ and put

$$f_l(p, I; x) = \sqrt{5}\, \bar{f}_{p,l}\left(5\,\frac{x-u}{v-u}\right) \qquad (l = 1, 2, \ldots, 2p).$$

Then the functions $f_l(p, I; x)$ are obviously orthogonal to each other in $[u, v]$, and 1° is fulfilled. Moreover, by virtue of (11) the relations

$$\sum_{l=1}^{m-\mu} f_l(p, I; x) \geqq \sqrt{5} K_4 \sqrt{p} \log p \qquad (\mu = 0, 1, \ldots, [\sqrt{p}])$$

hold in the interval $F(I)$ into which the interval $[2, 3)$ is transferred by the linear transformation; this proves 2°.

5. It is clear from the definition of the functions $f_l(p, I; x)$ in Lemma 2 that

$$|f_l(p, I; x)| \leqq K_5 \sqrt{p} \qquad (l = 1, 2, \ldots, 2p). \tag{12}$$

We are going to define two monotone increasing sequences $\{N_r\}$ and $\{p_r\}$ of integers by recurrence with respect to r. Set $N_1 = -1$ and $p_1 = 2$. Now for arbitrary $r \geqq 1$ we suppose that N_ϱ and p_ϱ are already defined for $\varrho = 1, 2, \ldots, r$. Let ur chose N_{r+1} so large that the conditions

$$N_{r+1} \geqq 6p_r,$$

$$2K_5 \sum_{\varrho=1}^{r} \frac{p_\varrho}{\sqrt[4]{\log^3 p_\varrho}} \cdot \frac{1}{2^n} \sum_{k=0}^{N_r+2p_r} \binom{n}{k} \leqq 1, \tag{13}$$

$$2K_5 \sum_{\varrho=1}^{r} \frac{3p_\varrho^2}{\sqrt[4]{\log^3 p_\varrho}} \cdot \frac{1}{2^n} \sum_{k=N_r+2p_r+1}^{n} \binom{n}{k} \frac{1}{k+1} \leqq 1$$

are satisfied for every $n \geqq N_{r+1}$. On account of (5) and (6) this choice is possible. Then let us put $p_{r+1} = N_{r+1}$.

6. After these preliminaries we are able to construct a simple set [4] F_r for every value of r, and a system $\{\varphi_\nu(x)\}$ of orthonormal step functions in [0, 1] with following properties:

a) The sets $F_1, F_2, \ldots$ are stochastically independent, and

$$|F_1| = \frac{1}{5}, \quad |F_r| = \frac{1}{10} \ [5] \qquad (r = 2, 3, \ldots). \tag{14}$$

b) For every $x \in F_r$ there exists an index $m_r(x)$ with $p_r \leqq m_r(x) < 2p_r$ such that on the one hand the functions

$$\varphi_{N_r+1}(x), \ \ldots, \ \varphi_{N_r+m_r(x)}(x)$$

are positive, and on the other hand the inequality

$$\sum_{\nu=N_r+1}^{N_r+m_r(x)-\mu} \varphi_\nu(x) \geqq K\sqrt{p_r} \log p_r \tag{15}$$

holds for $\mu = 0, 1, \ldots, [\sqrt{p_r}]$.

To prove this we apply induction with respect to r. First of all let $r = 1$. We apply Lemma 2 with $p = p_1 = 2$ and $I = [0, 1]$. Putting $F_1 = F(I)$ we can see that (14) is fulfilled for $r = 1$. Now let

$$\varphi_{l-1}(x) = f_l(p_1, I; x) \qquad (l = 1, \ldots, 4).$$

According to Lemma 2 the functions $\varphi_\nu(x)$ form an orthonormal system with $2p_1$ terms. For $r = 1$ (15) follows from (8). Finally, we have to define $\varphi_\nu(x)$ also for the indices $2p_1 \leqq \nu \leqq N_2$. This is to be done in the following way: There exists a

[4] A set F is said to be simple if it is the union of finitely many non-overlapping intervals.
[5] $|F|$ denotes the Lebesgue measure of the set F.

division of [0, 1] into a finite number of subintervals $J_1, J_2, \ldots, J_s$ such that all the hitherto defined functions $\varphi_\nu(x)$ with $0 \leqq \nu < 2p_1$ remain constant in each single J_σ. Let us divide J_σ into its two halves J'_σ, J''_σ and set for $l = 0, 1, \ldots, N_2 - 2p_1$

$$\varphi_{2p_1+l}(x) = \sum_{\sigma=1}^{s} r_l(J'_\sigma, x) - \sum_{\sigma=1}^{s} r_l(J''_\sigma, x).$$

Here $r_l(I, x)$ denotes the lth Rademacher function [6], formed in the interval $I = [u, v]$, i.e.

$$r_l(I, x) = \begin{cases} r_l\left(\dfrac{x-u}{v-u}\right) & \text{for} \quad u \leqq x \leqq v, \\ 0 & \text{otherwise.} \end{cases}$$

It is obvious that the step functions $\varphi_\nu(x)$ with $0 \leqq \nu \leqq N_2$ are orthogonal to each other and also normed in [0, 1].

Now for arbitrary $r \geqq 2$ we assume that the assertion is already proved for each integer $\leqq r-1$. Then we can divide [0, 1] into a finite number of subintervals $I_1, I_2, \ldots, I_t$ so that every function $\varphi_\nu(x)$ with $\nu = 0, 1, \ldots, N_r$ remains constant and every set F_ϱ with $\varrho = 1, 2, \ldots, r-1$ is the union of some intervals I_τ. Let I'_τ, I''_τ denote the two halves of I_τ. Now let us apply Lemma 2 with $p = p_r$ and $I = I'_\tau$, $I = I''_\tau$ $(\tau = 1, 2, \ldots, t)$, and let us put

$$F_r = \bigcup_{\tau=1}^{t} F(I'_\tau),$$

and

$$\varphi_{N_r+l}(x) = \sum_{\tau=1}^{t} f_l(p_r, I'_\tau; x) - \sum_{\tau=1}^{t} f_l(p_r, I''_\tau; x)$$

for $l = 1, 2, \ldots, 2p_r$. And let us define the functions $\varphi_\nu(x)$ for $N_r + 2p_r < \nu \leqq N_{r+1}$ just as in the case of $r = 1$ by means of Rademacher functions. By a simple calculation we can see that $\{\varphi_\nu(x)\}$ $(0 \leqq \nu \leqq N_{r+1})$ is an orthonormal system. (In detail, see ALEXITS [1], pp. 94—95.) Furthermore, by virtue of (8) in Lemma 2 we obtain the required properties a) and b) for r as well. This completes the scheme of induction.

7. Now let us put

$$c_\nu = \begin{cases} C_r = \dfrac{1}{\sqrt[4]{p_r^2 \log^3 p_r}} & \text{for} \quad N_r < \nu < N_r + p_r, \\ C_r - \dfrac{C_r}{2p_r}(\nu - 2p_r) & \text{for} \quad N_r + p_r \leqq \nu \leqq N_r + 2p_r, \\ 0 & \text{otherwise} \quad (r = 1, 2, \ldots). \end{cases}$$

[6] The lth Rademacher function is defined as follows: $r_l(x) = \operatorname{sign} \sin 2^l \pi x$ $(0 \leqq x \leqq 1;\ l = 0, 1, \ldots)$. (See ALEXITS [1], p. 51.)

By (13) we find that $p_r \geqq 2^r$ $(r=1, 2, \ldots)$. It then follows that

$$\sum_{\nu=2}^{\infty} c_\nu^2 (\log\log \nu)^2 = \sum_{r=1}^{\infty} \sum_{\nu=N_r+1}^{N_r+2p_r} c_\nu^2 (\log\log \nu)^2 \leqq$$

$$\leqq 2 \sum_{r=1}^{\infty} p_r C_r^2 (\log\log 3p_r)^2 \leqq 2 \sum_{r=1}^{\infty} \frac{(\log\log 3p_r)^2}{\sqrt[4]{\log^3 p_r}} < \infty,$$ [7]

i.e. our requirement (3) concerning the coefficients c_ν is fulfilled.

By virtue of (15) it is clear that for every $x \in F_r$ and for $\mu = 0, 1, \ldots, [\sqrt{p_r}]$ we have

$$\sum_{\nu=N_r+1}^{N_r+m_r(x)-\mu} c_\nu \varphi_\nu(x) \geqq \frac{K}{2} \sqrt[4]{\log p_r} \qquad (r = 1, 2, \ldots). \tag{16}$$

8. Finally, we consider the Euler means

$$\pi_n(x) = \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} \frac{1}{k+1} \sum_{\nu=0}^{k} \nu c_\nu \varphi_\nu(x)$$

occurring in (4) of our theorem. If $N_r < n \leqq N_{r+1}$ $(r \geqq 1)$, we write $\pi_n(x)$ in the following form:

$$\pi_n(x) = \frac{1}{2^n} \sum_{k=0}^{N_{r-1}+2p_{r-1}} \binom{n}{k} \frac{1}{k+1} \sum_{\nu=0}^{k} \nu c_\nu \varphi_\nu(x) +$$

$$+ \frac{1}{2^n} \sum_{k=N_{r-1}+2p_{r-1}+1}^{N_r} \binom{n}{k} \frac{1}{k+1} \sum_{\nu=0}^{N_{r-1}+2p_{r-1}} \nu c_\nu \varphi_\nu(x) +$$

$$+ \frac{1}{2^n} \sum_{k=N_r+1}^{n} \binom{n}{k} \frac{1}{k+1} \sum_{\nu=0}^{N_{r-1}+2p_{r-1}} \nu c_\nu \varphi_\nu(x) +$$

$$+ \frac{1}{2^n} \sum_{k=N_r+1}^{n} \binom{n}{k} \frac{1}{k+1} \sum_{\nu=N_r+1}^{k} \nu c_\nu \varphi_\nu(x) = P_1 + P_2 + P_3 + P_4.$$

(In case $r=1$ we have $\pi_n(x) = P_4$.) We show that the first three sums on the right-hand side remain under a common bound for every n and every x. In fact, on account of (12) and (13) this follows from the estimates

$$|P_1| \leqq \max_{0 \leqq k \leqq N_{r-1}+2p_{r-1}} \frac{1}{k+1} \left| \sum_{\nu=0}^{k} \nu c_\nu \varphi_\nu(x) \right| \cdot \frac{1}{2^n} \sum_{k=0}^{N_{r-1}+2p_{r-1}} \binom{n}{k} \leqq$$

$$\leqq 2K_5 \sum_{\varrho=1}^{r-1} \frac{p_\varrho}{\sqrt[4]{\log^3 p_\varrho}} \cdot \frac{1}{2^n} \sum_{k=0}^{N_{r-1}+2p_{r-1}} \binom{n}{k} \leqq 1,$$

[7] This immediately follows if we write the general term of the last sum into the form

$$\frac{(\log\log 3p_r)^2}{\sqrt[4]{\log p_r}} \cdot \frac{}{\sqrt[4]{\log^5 p_r}},$$

which does not exceed $K_6/\sqrt[4]{r^5}$ if r is large enough.

and

$$|P_2+P_3| \leqq \left| \sum_{v=0}^{N_{r-1}+2p_{r-1}} vc_v\varphi_v(x) \right| \cdot \frac{1}{2^n} \sum_{k=N_{r-1}+2p_{r-1}+1}^{n} \binom{n}{k} \frac{1}{k+1} \leqq 1.$$

Thus we obtain for $N_r < n \leqq N_{r+1}$ and for every x that

$$\pi_n(x) \geqq \frac{1}{2^n} \sum_{k=N_r+1}^{n} \binom{n}{k} \frac{1}{k+1} \sum_{v=N_r+1}^{k} vc_v\varphi_v(x) - 2. \tag{17}$$

Denoting by $\varrho_n(x)$ the sum on the right-hand side, it remains to estimale merely $\varrho_n(x)$. In the particular case $N_r+2p_r < n \leqq N_{r+1}$ and $x \in F_r$ we consider the following decomposition:

$$\varrho_n(x) = \frac{1}{2^n} \sum_{k=N_r+1}^{N_r+m_r(x)} \binom{n}{k} \frac{1}{k+1} \sum_{v=N_r+1}^{k} vc_v\varphi_v(x) +$$

$$+ \frac{1}{2^n} \sum_{k=N_r+m_r(x)+1}^{N_r+2p_r} \binom{n}{k} \frac{1}{k+1} \sum_{v=N_r+1}^{k} vc_v\varphi_v(x) +$$

$$+ \frac{1}{2^n} \sum_{k=N_r+2p_r+1}^{n} \binom{n}{k} \frac{1}{k+1} \sum_{v=N_r+1}^{N_r+2p_r} vc_v\varphi_v(x) = R_1 + R_2 \quad R_3,$$

where $m_r(x)$ is defined in (15). As for R_2 and R_3, a simple caleulation shows that

$$v' c_{v'} \geqq v'' c_{v''}$$

where $v' = N_r + m_r(x) - \varkappa$ and $v'' = N_r + m_r(x) + \varkappa + 1$ $(\varkappa = 0, 1, \ldots, 2p_r - m_r(x) - 1)$; on the other hand

$$\varphi_{v'}(x) = -\varphi_{v''}(x)$$

at the points of F_r, with the above v', v''. Summing up these results we can see that

$$\sum_{v=N_r+1}^{k} vc_v\varphi_v(x) \geqq 0 \qquad (N_r < k \leqq N_r + 2p_r,\ x \in F_r).$$

Hence we deduce that

$$\varrho_n(x) \geqq \frac{1}{2^n} \sum_{k=N_r+1}^{N_r+m_r(x)} \binom{n}{k} \frac{1}{k+1} \sum_{v=N_r+1}^{k} vc_v\varphi_v(x),$$

if $N_r+2p_r < n \leqq N_{r+1}$ and $x \in F_r$.

Thus our problem has been reduced to showing that the sum on the right-hand side of the last inequality increases unboundedly with n. Taking into account that

$$\min_{N_r < v \leqq k \leqq N_r + m_r(x)} \frac{v}{k+1} \geqq \frac{N_r}{N_r+2p_r} = \frac{1}{3},$$

we also have

$$\varrho_n(x) \geqq \frac{1}{3\cdot 2^n} \sum_{k=N_r+1}^{N_r+m_r(x)}{}' \binom{n}{k} \sum_{\nu=N_r+1}^{k}{}' c_\nu \varphi_\nu(x) \geqq \tag{18}$$

$$\geqq \frac{1}{3\cdot 2^n} \sum_{k=N_r+m_r(x)-[\sqrt{p_r}]}^{N_r+m_r(x)}{}' \binom{n}{k} \sum_{\nu=N_r+1}^{k}{}' c_\nu \varphi_\nu(x).$$

Now for every $x \in F_r$ let us put $n_r(x) = 2(N_r + m_r(x))$. Then

$$N_r + 2p_r < 4p_r \leqq n_r(x) \leqq 6p_r \leqq N_{r+1}. \tag{19}$$

By virtue of (16) and (18) we conclude that in the points of F_r the inequality

$$\varrho_{n_r(x)}(x) \geqq \frac{K}{6} \sqrt[4]{\log p_r} \frac{1}{2^{n_r(x)}} \sum_{k=N_r+m_r(x)-[\sqrt{p_r}]}^{N_r+m_r(x)}{}' \binom{n_r(x)}{k}$$

holds. Let us observe that (19) implies

$$N_r + m_r(x) - [\sqrt{p_r}] \leqq \frac{n_r(x)}{2} - \sqrt{\frac{2}{3}} \frac{\sqrt{n_r(x)}}{2},$$

so that we can apply Lemma 1 in the form (7) with $a = -\sqrt{\frac{2}{3}}$ and $b = 0$. Thus we get

$$\frac{1}{2^{n_r(x)}} \sum_{k=N_r+m_r(x)-[\sqrt{p_r}]}^{N_r+m_r(x)}{}' \binom{n_r(x)}{k} \geqq \frac{1}{2} \left\{ G(0) - G\left(-\sqrt{\frac{2}{3}}\right) \right\}, \tag{20}$$

if r is large enough.

Collecting results (17), (18) and (20) we come to the following conclusion: if $x \in F_r$ and r is sufficiently large then

$$\pi_n(x) \geqq \frac{K}{12} \left\{ G(0) - G\left(-\sqrt{\frac{2}{3}}\right) \right\} \sqrt[4]{\log p_r} - 2 \geqq K_7 \sqrt[4]{r}. \tag{21}$$

Making use of BOREL—CANTELLI lemma we obtain by virtue of (14) that

$$\left| \overline{\lim_{r\to\infty}} F_r \right| = 1.$$

If $x \in \overline{\lim_{r\to\infty}} F_r$ then the inequality (21) is satisfied for infinitely many values of r, and this proves the desired assertion (4).

We have thus completed the proof of our theorem.

REFERENCES

[1] ALEXITS, G.: *Convergence Problems of Orthogonal Series* (Budapest, 1961).
[2] KACZMARZ, S.: Sur la convergence et sommabilité des développements orthogonaux, *Studia Math.*, **1** (1929), 81—121.
[3] KACZMARZ, S.: Notes on orthogonal series. II, *Studia Math.*, **5** (1934), 103—106.
[4] KNOPP, K.: Über das Eulerische Summierungsverfahren. II, *Math. Zeit.*, **18** (1923), 125—156.
[5] MEDER, J.: On the summability almost everywhere of orthonormal series by the method of Euler-Knopp, *Ann. Polon. Math.*, **5** (1958—59), 135—148.
[6] MENCHOFF, D.: Sur les séries de fonctions orthogonales. I, *Fund. Math.*, **4** (1923), 82—105.
[7] RÉNYI, A.: *Wahrscheinlichkeitsrechnung* (Berlin, 1962).

J. Bolyai Mathematical Institute, Szeged

(Received March 20, 1969*)*

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 161—171

НЕКОТОРЫЕ ВОПРОСЫ, СВЯЗАННЫЕ С АППРОКСИМАЦИЕЙ СПЛАЙН-ФУНКЦИЯМИ И МНОГОЧЛЕНАМИ

Г. ФРАЙД и В. А. ПОПОВ

В этой работе будет развит метод, позволяющий получать из оценок, в которых участвует вариация k-той производной функции $f(x)$, другие оценки, в которых участвуют разные модули непрерывности функции $f(x)$. Часть результатов были сообщены на Колоквиуме по Конструктивной теории функций в Будапеште, 1969 г.

Мы будем в дальнейшем интересоваться наилучшими приближениями функции $f(x)$ сплайн-функциями и многочленами в разных метриках. Напомним определение сплайн-функций [1]:

Будем говорить, что функция $f(x)$, заданная на отрезке $[0, 1]$, является сплайн-функцией (k, n)-ного порядка, если $f(x) \in C^{k-1}_{[0,1]}$ и существуют $n+1$ точек интервала $[0, 1]$: $0 = x_0 \leqq x_1 \leqq \ldots \leqq x_n = 1$ такие, что в интервале $[x_{i-1}, x_i]$, $i = 1, \ldots, n$, функция $f(x)$ является многочленом k-той степени.

Класс всех сплайн-функций (k, n)-ного порядка будем обозначать через $S[k, n]$. Точки x_i, $i = 0, \ldots, n$, будем называть узлами сплайн-функции $f(x)$.

Кроме сплайн-функций класса $S[k, n]$, которые будем называть сплайн-функциями с подвижными узлами, мы будем рассматривать сплайн-функции с фиксированными узлами. Пусть $\Sigma_n = \{x_i,\ i = 0, \ldots, n,\ 0 = x_0 \leqq \ldots \leqq x_n = 1\}$ — заданная система $n+1$ точек интервала $[0, 1]$. Будем обозначать через $S[k, \Sigma_n]$ класс всех сплайн-функций (k, n)-ного порядка с узлами в точках Σ_n. Обозначим через $\mathscr{E}^k_{n,R}(f)$ и $\mathscr{E}^k_{\Sigma_n,R}(f)$ наилучшие приближения функции $f(x)$ функциями из $S[k, n]$ и $S[k, \Sigma_n]$ относительно некоторого расстояния $R(f, g)$ на отрезке $[0, 1]$:

$$\mathscr{E}^k_{n,R}(f) = \inf_{s \in S[k,n]} R(f, s)$$

$$\mathscr{E}^k_{\Sigma_n,R}(f) = \inf_{s \in S[k,\Sigma_n]} R(f, s).$$

В качестве расстояния $R(f, g)$ будем брать равномерное расстояние $\varrho(f, g)$ между функциями $f(x)$ и $g(x)$:

$$\varrho(f, g) = \sup_{x \in [0,1]} |f(x) - g(x)|$$

или интегральное расстояние в L_p: $\left\{\int_0^1 |f(x) - g(x)|^p \, dx\right\}^{\frac{1}{p}}$.

Когда рассматриваем равномерное наилучшее приближение функции $f(x)$, мы иногда будем опускать индекс R, например $\mathscr{E}^k_n(f)$.

Мы будем тоже рассматривать наилучшее одностороннее приближение функции $f(x)$ сплайн-функциями из $S[k, \Sigma_n]$ и многочленами в L_1:

$$\tilde{\mathscr{E}}^k_{\Sigma_n}(f) = \inf_{\substack{p(x) \leqq f(x) \leqq q(x) \\ p \in S[k, \Sigma_n];\ q \in S[k, \Sigma_n]}} \int_0^1 \{q(x) - p(x)\}\, dx$$

и

$$\tilde{E}_n(f) = \inf_{\substack{p(x) \leqq f(x) \leqq q(x) \\ p \in H_n,\ q \in H_n}} \int_0^1 \{q(x) - p(x)\}\, dx$$

где H_n — совокупность всех многочленов n-ной степени.

В первой части работы будут получены некоторые вспомогательные леммы и оценки для разных наилучших приближений функции $f(x)$ через вариацию ее k-той производной. Некоторые из них — известны, другие являются новыми.

Во второй части вводится модифицированная функция Стеклова для функции $f(x)$ и для нее доказываются ряд свойств. В частности дается оценка ее k-той производной через модуль непрерывности в L_1 $k-1$-той производной функции $f(x)$.

Третья часть работы содержит основные теоремы. Они связывают наилучшее приближение данной функции $f(x)$ с ее модулями непрерывности разных видов. Эти теоремы обобщают некоторые известные результаты. Например, Корнейчук [2] показал, что если $f(x) \in \mathrm{Lip}\, 1$, то тогда $f(x)$ можно равномерно приближать ломанными с n узлами с точностью $o\left(\frac{1}{n}\right)$, т. е. если $f(x) \in \mathrm{Lip}\, 1$, то $\mathscr{E}^1_n(f) = o\left(\frac{1}{n}\right)$. В настоящей работе получается более общая теорема (Теорема 3), из которой следует аналогичный результат, когда рассматриваем $\mathscr{E}^k_n(f)$ и предполагаем, что $k-1$-вая производная функции $f(x)$ является абсолютно-непрерывной и $f^{(k-1)}(x)$ принадлежит к классу Зигмунда.

Вообще говоря, доказанные теоремы дают порядок $o\left(\frac{1}{n^k}\right)$ для наилучших приближений функции $f(x)$ сплайн-функциями с подвижными узлами или для односторонных наилучших приближений сплайн-функциями с равноотстоящими узлами или многочленами, если $f^{(k-1)}(x)$ является абсолютно-непрерывной или принадлежит к классу Зигмунда.

§ 1

Докажем сначала некоторые вспомогательные результаты.

Лемма 1. *Пусть $f(x)$ задана на отрезке* $[0, 1]$ *и* $s(x) \in S[k-1, \Sigma_n]$, $\Sigma_n = \{x_i\}_0^n$, $x_i \neq x_j$ *если* $i \neq j$, *такая, что*

$$\varrho(f, s) \leqq \eta.$$

Тогда существует функция $s^*(x) \in S[k, \Sigma_n]$ *такая, что*

$$\varrho\left(\int_0^x f(t)\,dt, s^*(x)\right) \leqq (k+1)\eta\Delta_n$$

где $\Delta_n = \max_i |x_i - x_{i-1}|$.

Доказательство. Рассмотрим следующие сплайн-функции $\varphi_i(x) \in S[k-1, \Sigma_n]$:

$$\varphi_i(x) = \sum_{j=1}^{i+k} \frac{k(x_j - x)_+^{k-1}}{\omega_i'(x_j)}; \qquad i = 0, \ldots, n-k$$

где

$$\omega_i(x) = \prod_{j=i}^{i+k} (x - x_j), \quad (x-t)_+^{k-1} = \begin{cases} (x-t)^{k-1} & \text{если} \quad x \geqq t \\ 0 & \text{если} \quad x \leqq t. \end{cases}$$

Известно [3], что $\varphi_i(x) > 0$ для $x \in (x_i, x_{i+k})$, $\varphi_i(x) = 0$ если $x \overline{\in} (x_i, x_{i+k})$ и кроме того

$$\int_{-\infty}^{\infty} \varphi_i(x)\,dx = 1.$$

Положим:

$$A_i = \int_{x_i}^{x_{i+1}} \{f(x) - s(x)\}\,dx.$$

Рассмотрим теперь сплайн-функцию $s^*(x) \in S[k, \Sigma_n]$:

$$s^*(x) = \int_0^x s(t)\,dt - \sum_{i=0}^{n-k} A_i \int_0^x \varphi_i(t)\,dt. \tag{1}$$

Тогда

$$\int_0^x f(t)\,dt - s^*(x) = \int_0^x \{f(t) - s(t)\}\,dt - \sum_{i=0}^{n-k} A_i \int_0^x \varphi_i(t)\,dt. \tag{2}$$

Пусть $x_{i_0} \leqq x \geqq x_{i_0+1}$. Тогда из (2) следует

$$\left| \int_0^x f(t)\,dt - s^*(x) \right| \geqq \int_{x_{i_0-k+1}}^{x_{i_0+1}} |f(t) - s(t)|\,dt \leqq (k+1)\eta\Delta_n$$

(x_i для отрицательных i полагаем равными нулю) и лемма доказана.

Лемма 2. *Пусть* $f(x)$ *задана на отрезке* [0, 1] *и* $\mathscr{E}_n^{k-1}(f) = O(\varphi(n))$, *где функция* $\varphi(n)$ *такая, что* $\varphi(mn) \geqq C_m \varphi(n)$. *Тогда*

$$\mathscr{E}_n^k\left(\int_0^x f(t)\,dt\right) = O\left(\frac{\varphi(n)}{n}\right).$$

Доказательство. Если $\mathscr{E}_n^{k-1}(f) = O(\varphi(n))$, то для каждого натурального n существует сплайн-функция $S_n(x) \in S[k-1, n]$ такая, что

а) Если x_i, $i=0, \ldots, n$, являются узлами сплайн-функции $S_n(x)$, то $\max_i |x_i - x_{i-1}| \leqq \frac{N}{n}$, где константа N не зависит от n.

б) $\varrho(f, S_n) \leqq M\varphi(n)$ (здесь используется $\varphi(mn) \leqq C_m \varphi(n)$).

Если приложить лемму 1 к функциям $S_n(x)$, то получаем утверждении леммы.

Лемма 3. *Пусть* $f(x)$ *задана на отрезке* $[0, 1]$ *и* $S(x) \in S[k-1, \Sigma_n]$, $\Sigma_n = \{x_i\}_0^n$, $x_i \neq x_j$ *если* $i \neq j$, *такая, что* $f(x) \geqq S(x)$ *и*

$$\int_0^1 \{f(x) - s(x)\}\, dx \leqq \eta.$$

Тогда существует функция $S^*(x) \in S[k, \Sigma_n]$, $S^*(x) \leqq \int_0^x f(t)dt$ *такая, что*

$$\int_0^1 \left[\int_0^x f(t)\, dt - s^*(x)\right] dx \leqq (k+1)\eta\Delta_n$$

где $\Delta_n = \max_i |x_i - x_{i-1}|$.

Доказательство этой леммы повторяет доказательство леммы 1. Снова рассмотрим функцию $S^*(x)$, заданную формулой (1). Заметим, что

$$\int_0^x f(t)\, dt - s^*(x) \geqq 0. \tag{3}$$

Действительно, так как $f(x) \geqq S(x)$, то $(x_{i_0} \leqq x \leqq x_{i_0+1})$

$$\int_0^x f(t)\, dt - s^*(x) = \sum_{i=i_0-k}^{i_0} \alpha_i \int_{x_i}^{x_{i+1}} \{f(t) - s(t)\}\, dt \tag{4}$$

где $0 \leqq \alpha_i \leqq 1$. (Если $i_0 - k < 0$, то сумма берется с нуля.) Из (4) следует очевидно (3).

Оценим сейчас выражение (4). Имеем:

$$\int_0^1 \left[\int_0^x f(t)\, dt - s^*(x)\right] dx \leqq \int_0^1 \left[\int_{x-(k+1)\Delta_n}^{x} \{f(t) - s(t)\}\, dt\right] dx =$$

$$= \int_0^1 \left[\int_t^{t+(k+1)\Delta_n} \{f(t) - s(t)\}\, dx\right] dt \leqq (k+1)\eta\Delta_n$$

(если $x - (k+1)A_n < 0$ или $t + (k+1)\Delta_n > 1$, то интегралы берутся с 0, соответсно до 1).

Лемма доказана.

Лемма 4. *Пусть* $f(x)$ — *ограниченная функция, задана на* $[0, 1]$, $s(x) \in S[k, n]$ *и* $\int_0^1 |f(t)-s(t)|\,dt \leqq \eta$. *Тогда существует* $s^*(x) \in S[k+1, 3n]$ *такая, что*

$$\varrho\left(\int_0^x f(t)\,dt,\ s^*(x)\right) \leqq c\,\frac{\eta}{n}.$$

Доказательство. Прибавим к узлам функции $s(x)$ $2n$ узлов так, что если x_i, $i=0, \ldots, 3n$, являются новыми узлами, то выполнено

а) $$|x_i - x_{i-1}| \leqq \frac{1}{n};$$

б) $$\int_{x_i}^{x_{i+1}} |f(t)-s(t)|\,dt \leqq \frac{\eta}{n}.$$

Рассматривая снова функцию $s^*(x)$ из (1), пользуясь а) и б) получаем утверждения леммы.

Эта лемма позволяет из оценок для одностороннего приближения сплайн-функциям получать утверждения для равномерной аппроксимации функции $f(x)$ сплайн-функциями с подвижными узлами.

Получим теперь оценки для наилучших приближений функции $f(x)$ через вариацию ее k-той производной.

В [4] доказано, что если функция $f(x)$ выпукла на отрезке $[0, 1]$ и имеет в точках 0 и 1 соответственно правую производную $f'(0_{+0}) > -\infty$ и левую производную $f'(1_{-0}) < \infty$, то

(5) $$\mathscr{E}_n^1(f) \leqq \frac{f'(1_{-0})-f'(0_{+0})}{8n^2}.$$

Из (5) немедленно следует, что если функция $f(x)$ является интегралом функции $f'(x)$ ограниченной вариации, то

(6) $$\mathscr{E}_n^1(f) = O\left\{\frac{V_0^1(f')}{n^2}\right\}$$

где $V_0^1(g)$ обозначает вариацию функции $g(x)$ на отрезке $[0, 1]$.

Из (6) и леммы 2 следует, что если $f(x)$ имеет $k-1$-вую производную, которая является интегралом функции $f^{(k)}(x)$ ограниченной вариации, то

(7) $$\mathscr{E}_n^k(f) = O\left\{\frac{V_0^1(f^{(k)})}{n^{k+1}}\right\}.$$

В [5] доказано, что

(8) $$\tilde{\mathscr{E}}^1_{\Sigma_n}(f) = O(V_0^1(f')\Delta_n^2).$$

Применяя к (8) лемму 3 получаем

(9) $$\tilde{\mathscr{E}}^k_{\Sigma_n}(f) = O(V_0^1(f^{(k)})\Delta_n^{k+1}).$$

Заметим, что из (7) следует непосредственно

$$\mathcal{E}^k_{n,L_p}(f) = O\left(\frac{V_0^1(f^{(k)})}{n^{k+1}}\right). \tag{10}$$

Кроме того в [6] показано, что

$$\tilde{E}_n(f) = O\left(\frac{V_0^1(f^{(k)})}{n^{k+1}}\right). \tag{11}$$

Объединяя (7), (9), (10) и (11) получаем

Теорема 1. *Пусть $f(x)$ обладает $k-1$-вой производной, которая является интегралом функции $f^{(k)}(x)$ ограниченной вариации. Тогда*

$$\mathcal{E}^k_n(f) = O\left(\frac{V_0^1(f^{(k)})}{n^{k+1}}\right)$$

$$\mathcal{E}^k_{n,L_p}(f) = O\left(\frac{V_0^1(f^{(k)})}{n^{k+1}}\right)$$

$$\tilde{\mathcal{E}}^k_{\Sigma_n}(f) = O(V_0^1(f^{(k)})\Delta_n^{k+1})$$

$$\tilde{E}_n(f) = O\left(\frac{V_0^1(f^{(k)})}{n^{k+1}}\right).$$

§ 2

Рассмотрим следующую функцию, которая является некоторой модификацией функции Стеклова:

$$\hat{f}_{k,h}(x) = \tag{12}$$

$$= \frac{(-1)^{k-1}}{h^k}\int_0^h\cdots\int_0^h\left[f(x+t_1+\cdots+t_k)-\binom{k}{1}f\left(x+\frac{k-1}{k}(t_1+\cdots+t_k)\right)+\right.$$

$$\left.+\cdots+(-1)^{k-1}\binom{k}{k-1}f\left(x+\frac{t_1+\cdots+t_k}{k}\right)\right]dt_1\ldots dt_k.$$

Если функция $f(x)$ задана на отрезке $[0, 1]$, то функция $\hat{f}_{k,h}(x)$ задана на отрезке $[0, 1-kh]$. Следовательно чтобы получить $\hat{f}_{k,h}(x)$ на всем отрезке $[0, 1]$, надо продолжить $f(x)$ подходящим образом на отрезке $[0, 1+kh]$. Мы хотим, чтобы продолженная функция имела модуль непрерывности k-того порядка, который мажорируется модулем непрерывности $\omega_k(f; h)$ k-того порядка, умноженный на константу. Для этого воспользуемся интерполяционным многочленом Лагранжа $P_h(x)$ для функции $f(x)$ построенным по узлам $1-ih$, $i=0, \ldots, k$. Известно [7], что

$$\|f(x)-P_h(x)\|_{C[1-kh,1]} \leqq C_1\,\omega_k(f; h) \tag{13}$$

где C_1 от h не зависит.

Если мы продолжим $f(x)$ на отрезке $[1, 1+kh]$, полагая $f(x)=P_h(x)$, то получаем, в силу (13), что новая функция $\tilde{f}_h$ обладает модулем непрерывности k-того порядка $\omega_k(\tilde{f}_h; h)$ для которого

$$\omega_k(\tilde{f}_h; h) \leqq C_2\,\omega_k(f; h) \tag{14}$$

где C_2 не зависит от h.

Мы будем по-прежнему обозначать новую продолженную функцию через $f(x)$, так как для нас будет важно только, что выполнено (14).

Для функции $\hat{f}_{k,h}(x)$ очевидно имеем:

$$\|\hat{f}_{k,h}(x)-f(x)\|_{C[0,1]} \leqq C_3\,\omega_k(f; h) \tag{15}$$

где C_3 не зависит от h.

Кроме того функция $\hat{f}_{k,h}(x)$ является k-раз дифференцируемой и

$$\hat{f}_{k,h}^{(k)}(x) = (-1)^{k-1}h^{-k}\left[\Delta_h^k f(x)-\binom{k}{1}\Delta^k_{\frac{(k-1)h}{k}} f(x)+\cdots+ \right. \tag{16}$$

$$\left. +(-1)^k\binom{k}{k-1}\Delta^k_{\frac{h}{k}} f(x)\right].$$

Отметим, что из (16) следует

$$|\hat{f}_{k,h}^{(k)}(x)| \leqq 2^k h^{-k}\omega_k(f; h). \tag{17}$$

Из (17) получаем

$$V_0^1\left(\hat{f}_{k+1,h}^{(k)}(x)\right) \leqq 2^{k+1}h^{-k-1}\omega_{k+1}(f; h). \tag{18}$$

Чтобы получить аналог (15) в метрике L_p, всюду в дальнейшем, когда рассматриваем $\omega_k(f; h)_{L_p}$, будем предполагать, что функция $f(x)$ задана на отрезке, более большом чем $[0, 1]$ (скажем $[0, 2]$) и $\omega_k(f; h)_{L_p}$ будет обозначать ее L_p-модуль непрерывности k-того порядка на этом отрезке.

При этих предположениях из определения (12) сразу следует

$$\|\hat{f}_{k,h}(x)-f(x)\|_{L_p} \leqq C_4\,\omega_k(f; h)_{L_p}. \tag{19}$$

Кроме того

$$\|\hat{f}_{k,h}^{(k)}(x)\|_{L_p} \leqq C_5 h^{-k}\omega_k(f; h)_{L_p}. \tag{20}$$

Из (15), (18), (19), (20) и теорема 1 следует

Теорема 2. *Для каждой ограниченной функции выполнено*

$$\mathcal{E}_n^k(f) = O\left(\omega_{k+1}\left(f; \frac{1}{n}\right)\right)$$

$$\mathcal{E}_{n,L_p}^k(f) = O\left(\omega_{k+1}\left(f; \frac{1}{n}\right)_{L_p}\right)$$

$$\tilde{\mathcal{E}}_{\Sigma_n}^k(f) = O\left(\omega_{k+1}(f; \Delta_n)\right)$$

$$\tilde{E}_n(f) = O\left(\omega_{k+1}\left(f; \frac{1}{n}\right)\right).$$

Мы вычислим теперь $V_0^1(\hat{f}_{k+1,h}^{(k)})$ еще другим способом, в случае, когда $f(x)$ имеет абсолютно-непрерывную производную $k-1$-го порядка.

Так как $\hat{f}_{k,h}(x)$ является конечной линейной комбинацией функций вида

$$f_{k,\delta}(x) = \frac{1}{\delta^k}\int_0^\delta \cdots \int_0^\delta f(x+t_1+\cdots+t_k)\,dt_1\cdots dt_k$$

для некоторых $\delta \leqq h$, то достаточно вычислить вариацию только функции $f_{k+1,h}^{(k)}(x)$.

Имеем:

$$V_0^1\left(f_{k+1,h}^{(k)}(x)\right) = V_0^{1-(k+1)h}(f_{k+1,h}^{(k)}) + V_{1-(k+1)h}^1(f_{k+1,h}^{(k)}).$$

Для $V_{1-(k+1)h}^1(f_{k+1,h}^{(k)})$, используя (17), (20) получаем:

$$V_{1-(k+1)h}^1(f_{k+1,h}^{(k)}) \leqq C_6 h^{-k}\omega_{k+1}(f;h)$$

$$V_{1-(k+1)h}^1(f_{k+1,h}^{(k)}) \leqq C_7 h^{-k}\omega_{k+1}(f;h)_{L_p}.$$

Для $V_0^{1-(k+1)h}(f_{k+1,h}^{(k)})$ получаем $(\alpha = 1+(1-(k+1)h)m)$:

$$V_0^{1-(k+1)h}(f_{k+1,h}^{(k)}) = \lim_{m\to\infty}\sum_{i=1}^{\alpha}\left|f_{k+1,h}^{(k)}\left(\frac{i+1}{m}\right)-f_{k+1,h}^{(k)}\left(\frac{i}{m}\right)\right| =$$

$$= \lim_{m\to\infty}\sum_{i=1}^{\alpha}\left|h^{-k}\Delta_h^k\left[f_{1,h}\left(\frac{i+1}{m}\right)-f_{1,h}\left(\frac{i}{m}\right)\right]\right| =$$

$$= \lim_{m\to\infty}\sum_{i=1}^{\alpha}\left|h^{-k}\int_0^h\cdots\int_0^h\left[f_{1,h}^{(k)}\left(\frac{i+1}{m}+\sum_{j=1}^k t_j\right)-f_{1,h}^{(k)}\left(\frac{i}{m}+\sum_{j=1}^k t_j\right)\right]dt_1\ldots dt_k\right| =$$

$$= \lim_{m\to\infty}\sum_{i=1}^{\alpha}\left|h^{-k-1}\int_0^h\cdots\int_0^h\int_{\frac{i}{m}}^{\frac{i+1}{m}}\left[f^{(k)}\left(x+\sum_{j=1}^k t_j+h\right)-f^{(k)}\left(x+\sum_{j=1}^k t_j\right)\right]dx\,dt_1\ldots dt_k\right| \leqq$$

$$\leqq \lim_{m\to\infty}\sum_{i=1}^{\alpha}h^{-k-1}\int_0^h\cdots\int_0^h\int_{\frac{i}{m}}^{\frac{i+1}{m}}\left|f^{(k)}\left(x+\sum_{j=1}^k t_j+h\right)-f^{(k)}\left(x+\sum_{j=1}^k t_j\right)\right|dx\,dt_1\cdots dt_k \leqq$$

$$\leqq \lim_{m\to\infty}h^{-k-1}\int_0^h\cdots\int_0^h\int_0^{1-(k+1)h}\left|f^{(k)}\left(x+\sum_{j=1}^k t_j+h\right)-f^{(k)}\left(x+\sum_{j=1}^k t_j\right)\right|dx\,dt_1\ldots dt_k \leqq$$

$$\leqq \frac{1}{h}\omega_1(f^{(k)};h)_{L_1}.$$

Окончательно

$$V_0^1(\hat{f}_{k+1,h}^{(k)}) \leqq C_8[h^{-1}\omega_1(f^{(k)};h)_{L_1}+h^{-k}\omega_{k+1}(f;h)] \tag{21}$$

и

$$V_0^1(\hat{f}_{k+1,h}^{(k)}) \leqq C_9[h^{-1}\omega_1(f^{(k)},h)_{L_1}+h^{-k}\omega_{k+1}(f;h)_{L_p}]. \tag{22}$$

§ 3

Используя теперь результаты первых двух параграфов, получим некоторые оценки.

Пусть $f(x)$ имеет $k-1$-вую абсолютно непрерывную производную. Из (16), (21) и теоремы 1 получаем

$$\mathcal{E}_n^k(f) \leqq C_{10}\left[\omega_{k+1}(f;h)+\frac{h^{-k}\omega_{k+1}(f;h)+h^{-1}\omega_1(f^{(k)};h)_{L_1}}{n^{k+1}}\right].$$

Из (19), (22) и теоремы 1 получаем

$$\mathcal{E}_{n,L_p}^k(f) \leqq C_{11}\left[\omega_{k+1}(f;h)_{L_p}+\frac{h^{-k}\omega_{k+1}(f;h)_{L_p}+h^{-1}\omega_1(f^{(k)};h)_{L_1}}{n^{k+1}}\right].$$

Вполне аналогично из (16), (21), и теоремы 1

$$\tilde{\mathcal{E}}_{\Sigma_n}^k(f) \leqq C_{12}\left[\omega_{k+1}(f;h)+\left(h^{-k}\omega_{k+1}(f;h)+h^{-1}\omega_1(f^{(k)};h)_{L_1}\right)\Delta_n^{k+1}\right]$$

и наконец

$$\tilde{E}_n(f) \leqq C_{13}\left[\omega_{k+1}(f;h)+\frac{h^{-k}\omega_{k+1}(f;h)+h^{-1}\omega_1(f^{(k)};h)_{L_1}}{n^{k+1}}\right].$$

Если выбрать h_n так, что $h_n n \to 0$, $h_n^k n^{k+1} \underset{n\to\infty}{\to} \infty$ и $(h_n^{-1}\,\omega_1(f^{(k)};h_n)_{L_1})n^{-1} \underset{n\to\infty}{\to} 0$ (что возможно ввиду абсолютной непрерывности функции $f^{(k-1)}(x)$), получаем

Теорема 3. *Пусть $f(x)$ имеет $k-1$-вую производную, которая является абсолютно непрерывной. Тогда*

$$\mathcal{E}_n^k(f) = O\left(\omega_{k+1}\left(f;\frac{1}{m_n}\right)+\frac{1}{m_n^k}\right)$$

$$\mathcal{E}_{n,L_p}^k(f) = O\left(\omega_{k+1}\left(f;\frac{1}{m_n}\right)_{L_p}+\frac{1}{m_n^k}\right)$$

$$\tilde{E}_n(f) = O\left(\omega_{k+1}\left(f;\frac{1}{m_n}\right)+\frac{1}{m_n^k}\right)$$

где $n^{-1}m_n \to \infty$, если $n \to \infty$.

Кроме того, если $\Sigma_n = \left\{\frac{i}{n},\ i=0,\ \ldots,\ n\right\}$ — система равноотстоящих узлов, то

$$\tilde{\mathcal{E}}_{\Sigma_n}^k(f) = O\left(\omega_{k+1}\left(f;\frac{1}{m_n}\right)+\frac{1}{m_n^k}\right)$$

где $n^{-1}m_n \to \infty$, если $n \to \infty$.

Следствие. Если $f^{(k-1)}(x)$ — абсолютно-непрерывна, то

$$\mathscr{E}_n^k(f) = o\left(n^{-k-1}\omega_2\left(f; \frac{1}{n}\right)\right)$$

$$\mathscr{E}_{n,L_p}^k(f) = o\left(n^{-k-1}\omega_2\left(f; \frac{1}{n}\right)_{L_p}\right)$$

$$\tilde{E}_n(f) = o\left(n^{-k-1}\omega_2\left(f; \frac{1}{n}\right)\right).$$

Если $f^{(k-1)}(x)$ — абсолютно-непрерывна и принадлежит к классу Зигмунда, то

$$\mathscr{E}_n^k(f) = o\left(\frac{1}{n^k}\right)$$

$$\mathscr{E}_{n,L_p}^k(f) = o\left(\frac{1}{n^k}\right)$$

$$\tilde{E}_n(f) = o\left(\frac{1}{n^k}\right)$$

$$\tilde{\mathscr{E}}_{\Sigma_n}^k(f) = o\left(\frac{1}{n^k}\right), \quad \Sigma_n = \left\{\frac{i}{n}, \quad i = 0, \ldots, n\right\}.$$

Это следствие обобщает результат Корнейчука тоже в случае $k=1$, так как существует абсолютно-непрерывная функция $f(x)$, которая удовлетворяет условию Зигмунда, но не удовлетворяет условию Липшица.

Рассмотрим функцию:

$$f(x) = \sum_{k=0}^{\infty} \frac{\sin 2^k x}{k 2^k}. \tag{23}$$

Эта функция удовлетворяет условию Зигмунда, так как ее приближение частичными суммами порядка n ряда (23) имеет порядок $O\left(\frac{1}{n}\right)$. Из теоремы Рисса—Фишера следует, что $f(x)$ также абсолютно-непрерывна:

$$f'(x) \sim \sum_{k=0}^{\infty} \frac{\cos 2^k x}{k} \in \mathscr{L}_2. \tag{24}$$

Но в точке $x=0$ арифметические средние от (24) стремятся к ∞, следовательно (см. [8], теорема IV, 4.2) $f(x) \notin \mathrm{Lip}\, 1$.

ЛИТЕРАТУРА

[1] SCHOENBERG, I. J.: Spline functions, convex curves and mechanical quadrature. *Bull. Amer. Math. Soc.* **44** (1958) 352—357.
[2] Корнейчук, Н. П.—Половина, А. И.: О приближении непрерывных и дифференцируемых функций алгебраическими многочленами на отрезке, *Доклады А. Н. СССР* **166** (1966), 281—283.
[3] RICE, J. R.: The Approximation of Functions v. 2, 1969
[4] Попов, В. А.: Аппроксимация выпуклых функций полигонами, *Доклады Болг. АН,* **22** (1969).
[5] MEIR, A. and SHARMA, A.: One-sided spline approximation. *Studia Sci. Math. Hungar.* **3** (1968) 211—218.
[6] FREUD, G.: Über einseitige Approximation durch Polynome I. *Acta Sci. Math. (Szeged)* **16** (1955), 12—18.
[7] WHITNEY, H.: On functions with bounded n^{th} differences. *Jour. de Math. pures et appl.* **36** (1957), 67—95.
[8] ZYGMUND, A.: Trigonometric series I, II.

Институт Математики Академии Наук Венгрии, Будапешт; София

(Поступила 19-ого мая 1969 г.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 173—180

ÜBER EINE AFFININVARIANTE MASSZAHL BEI EIPOLYEDERN

von

L. FEJES TÓTH

§ 1. Newtonsche Zahlen

Es seien A und B zwei eigentliche konvexe Körper des k-dimensionalen euklidischen Raumes. Ferner sei $N(A, B)$ die Maximalzahl der kongruenten, nicht übereinandergreifenden Exemplare von A, die sich mit B in Berührung bringen lassen, die also je einen gemeinsamen Randpunkt mit B, aber weder mit B noch miteinander einen gemeinsamen inneren Punkt haben. Wir nennen $N(A, A)$ die Newtonsche Zahl von A [1] und $N(A, B)$ die Newtonsche Zahl von A bezüglich B. Diese Benennungen weisen auf die Streitfrage zwischen NEWTON und D. GREGORY bezüglich der Newtonschen Zahl einer gewöhnlichen Kugel hin, die 180 Jahre später zu Gunsten von NEWTON entschieden wurde [2]. Es ist bekannt, daß die Newtonsche Zahl eines regulären Drei- bzw. Vierecks 12 bzw. 8 beträgt [3]. Weitere Ergebnisse bezüglich Newtonscher Zahlen finden sich in [4], [5] und [6].

§ 2. Hadwigersche Zahlen

Ganz ähnlich wie $N(A, B)$ definieren wir die Zahl $H(A, B)$, indem wir in der obigen Definition von $N(A, B)$ statt kongruenten Exemplaren von A verschobene Exemplare betrachten. Nach einem schönen Satz von Hadwiger [7] gilt die genaue Abschätzung

$$H(A, A) \leqq 3^k - 1,$$

in der Gleichheit nur für Parallelotope besteht [8, 15]. Wir wollen $H(A, A)$ die Hadwigersche Zahl von A und $H(A, B)$ die Hadwigersche Zahl von A bezüglich B nennen.

In üblicher Weise bezeichnen wir mit λA einen zu A homothetischen Körper, dessen lineare Masse zu denen von A im Verhältnis $\lambda:1$ stehen. Wir stellen uns die Aufgabe, das asymptotische Verhalten von $H(\lambda A, A) = H\left(A, \frac{1}{\lambda} A\right)$ für kleine Werte von λ zu untersuchen.

Es sei A ein konvexes Polyeder, f eine $(k-1)$-dimensionale Fläche von A und F ein zu f paralleler Maximalschnitt von A, d.h. ein Schnitt mit maximalem $(k-1)$-dimensionalem Inhalt. Wir betrachten eine dichteste $(k-1)$-dimensionale Packung zu der Ebene von F parallel verschobener Exemplare von A. Dann bilden natürlich auch die entsprechend verschobenen Exemplare von F eine Packung, deren Dichte mit s bezeichnet werden soll. Ist ferner $n(\lambda)$ die Maximalzahl der verschobenen, nicht

übereinandergreifenden Exemplaren von λA, die A in der Fläche f berühren, so gilt die anschaulich einleuchtende Beziehung

$$\lim_{\lambda \to 0} \frac{n(\lambda)\lambda^{k-1}F}{f} = s,$$

die sich mit Hilfe allgemeiner Sätze [9] auch streng begründen läßt.

Nun gilt aber offensichtlich

$$\lim_{\lambda \to 0} \lambda^{k-1} H(\lambda A, A) = \lim_{\lambda \to 0} \lambda^{k-1} \sum n(\lambda),$$

wo sich die Summation auf sämtliche Flächen von A erstreckt. Folglich haben wir

$$\lim_{\lambda \to 0} \lambda^{k-1} H(\lambda A, A) = \sum \frac{sf}{F}.$$

Als Beispiel betrachten wir den Fall, daß $k=3$ und A ein Tetraeder T ist Jetzt gilt für alle Flächen $f=F$, und s stimmt mit der Dichte der dichtesten Packung von verschobenen Exemplaren eines Dreiecks überein. Diese Packung ist nach einem allgemeinen Satz von ROGERS [10] gitterartig, und es folgt leicht, daß $s=2/3$ ist. Folglich haben wir für große Werte von μ

$$H(T, \mu T) \sim \frac{8}{3}\mu^2.$$

Als zweites Beispiel sei wiederum $k=3$ und A ein Kuboktaeder K, d.h. die konvexe Hülle der Kantenmittelpunkte eines Würfels (oder eines Oktaeders). K ist ein quasiregulärer Körper [11] mit 8 Dreiecksflächen und 6 Viereckflächen. Die zu den Flächen parallele Maximalschnitte sind reguläre Sechsecke bzw. Quadrate vom Inhalt $F=6f$ bzw. $F=2f$. Ferner gilt für sämtliche Flächen $s=1$. Wir haben also $\Sigma sf/F = 8/6+6/2 = 13/3$ und folglich

$$H(K, \mu K) \sim \frac{13}{3}\mu^2.$$

Da die Packungsdichten s stets $\leqq 1$ sind, gilt die Ungleichung

$$\lim_{\lambda \to 0} \lambda^{k-1} H(\lambda A, A) \leqq \sum \frac{f}{F}.$$

Gleichheit besteht dann und nur dann, wenn für jede Fläche der parallele Maximalschnitt ein $(k-1)$-dimensionaler translatorischer Pflasterkörper ist, und außerdem die entsprechend verschobenen Exemplare von A nicht übereinandergreifen. Diese Bedingungen sind für $k=2$ automatisch erfüllt. Folglich gilt für ein konvexes Polygon P

$$\lim_{\lambda \to 0} \lambda H(\lambda P, P) = \sum \frac{f}{F}.$$

§ 3. Die Affininvariante $M(A)$

Im Zusammenhang mit der affininvarianten Größe $M(A) = \Sigma f/F$ erheben sich verschiedene ungelöste Probleme, die unabhängig von den obigen Überlegungen, an und für sich beachtenswert sind. Es handelt sich um die Bestimmung von $\sup M(A)$ und $\inf M(A)$ erstreckt über sämtliche eigentliche k-dimensionale konvexe Polyeder. Im Folgenden wollen wir uns in diesen Problemen orientieren.

Es sei bemerkt, daß sich die Definition von $M(A)$ durch das Oberflächenintegral $M(A) = \int \frac{df}{F}$ auf beliebige Eikörper verallgemeinern läßt. Hier bedeutet df das Flächendifferential und F das Schnittmaß [12] in der Normalenrichtung des Flächenelements. Um aber unsere Betrachtungen möglichst elementar zu halten, wollen wir uns weiterhin auf konvexe Polyeder beschränken.

Es sei hier gezeigt: *Ist A ein k-dimensionales zentralsymmetrisches konvexes Polyeder, so gilt $M(A) \leqq 2k$, mit Gleichheit für und nur für Parallelotope.*

Es sei A ein beliebiges konvexes Polyeder, f eine Fläche von A, S ein zu f paralleler Schnitt von A in der Tiefe x, $S(x)$ der Inhalt von S und b die Breite von A in der zu f senkrechten Richtung. Wir können $b > 0$ voraussetzen, da sonst $M(A) = 2$ wäre. Im Einklang mit unseren obigen Bezeichnungen setzen wir $f = S(0)$ und $F = \max_{0 \leqq x \leqq b} S(x)$. Ist V das Volumen von A, so haben wir

$$V = \int_0^b S(x)\,dx \leqq bF.$$

Gleichheit gilt nur im Falle, wenn $S(x)$ eine Konstante ist. Dies trifft nach dem BRUNN—MINKOWSKIschen Satz [12] nur dann zu, wenn A ein Prisma mit der Grundfläche f ist.

Schreiben wir die obige Ungleichung in der Form $f/F \leqq fb/V$ und summieren über sämtliche Flächen von A, so ergibt sich

$$M(A) \leqq \frac{1}{V} \Sigma bf.$$

Gleichheit gilt dann und nur dann, wenn A bezüglich aller seiner Flächen als Grundflächen ein Prisma, und folglich ein Parallelotop ist.

Ist A symmetrisch in Bezug auf einen Punkt O, so ist $\frac{1}{k}\frac{b}{2}f$ das Volumen der von O und f bestimmten Pyramide. Wir haben also

$$\frac{1}{2k} \Sigma bf = V$$

und folglich $M(A) \leqq 2k$, wobei Gleichheit nur für Parallelotope gilt.

§ 4. Polygone

Wir beweisen hier folgende Sätze:

SATZ 1. *Ist P ein konvexes Polygon, so gilt*

$$M(P) \leqq 4$$

und Gleichheit gilt nur für ein Parallelogramm.

SATZ 2 *Ist P ein eigentliches konvexes Polygon, so gilt*

$$M(P) > \sqrt{8}.$$

Die letzte Abschätzung ist nicht genau. Vermutlich gilt

$$M(P) \geqq 3$$

mit Gleichheit für und nur für den Dreieckstumpf. Der Dreieckstumpf entsteht, wenn man von einem Dreieck durch zu den Seiten parallele Geraden drei kongruente Dreiecke abschneidet (Fig. 1). Als Spezialfälle seien das Dreieck und das affin reguläre Sechseck erwähnt.

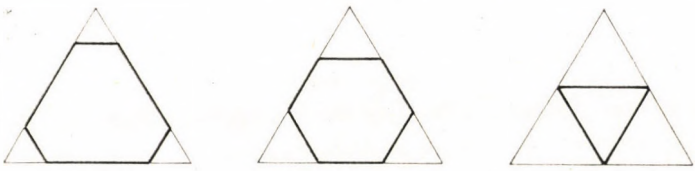


Fig. 1

Das Haupthilfsmittel beim Beweis der obigen Sätze ist die Symmetrisierung bezüglich eines Punktes [13, 7]. Es sei A ein konvexer Körper, $\bar{A}$ sein Spiegelbild bezüglich eines Punktes und B die Minkowskische Summe [12, 14] von $\frac{1}{2}A$ und $\frac{1}{2}\bar{A}$. Wir nennen diese Konstruktion, die aus A den zentralsymmetrischen konvexen Körper B erzeugt, Zentralsymmetrisierung.

Es sei P ein eigentliches konvexes Polygon, u eine Gerade, a_u^P die Vereinigung der beiden Stützmengen von P, die in zu u parallelen Stützgeraden liegen, A_u^P ein zu u paralleler Maximalschnitt von P und schließlich $a(u, P)$ und $A(u, P)$ das Maß von a_u^P und A_u^P. Wir stellen $M(P)$ in der From

$$M(P) = \sum \frac{a(u, P)}{A(u, P)}$$

dar, wo sich die Summation auf sämtliche Richtungen von u erstreckt. Natürlich ist diese Summe nur formal unendlich, weil $a(u, P)$ nur für die Seitenrichtungen von P von Null verschieden ist.

Wir erinnern jetzt an die Tatsache [14], daß sich die Stützmenge einer Minkowskischen Summe bezüglich einer orientierten Stützebene durch Minkowskische Addition der entsprechenden Stützmengen der Komponente ergibt. Ist also Q das Polygon, das aus P durch Zentralsymmetrisierung entsteht, so haben wir $a(u, P) = a(u, \bar{P}) =$

$=a(u, Q)$.Ferner gilt offenkundig $A(u, P)=A(u, Q)$, woraus $M(P)=M(Q)$ folgt. *Die Größe $M(P)$ ist* also *gegenüber* der *Zentralsymmetrisierung invariant.*

Da die Ungleichung $M(P)\leqq 4$ für zentralsymmetrische Polygone vorher bewiesen wurde, ist jetzt ihre Gültigkeit auch für beliebige konvexe Polygone dargetan. Da ferner unter der Bedingung der Zentralsymmetrie Gleichheit nur für Parallelogramme besteht, und da aus einem nicht zentralsymmetrischen Polygon durch Zentralsymmetrisierung kein Parallelogramm entstehen kann, werden durch die Gleichung $M(P)=4$ auch unter sämtlichen konvexen Polygonen die Parallelogramme charakterisiert.

Wir wenden uns jetzt dem Beweis der Ungleichung $M(P)>\sqrt{8}$ zu.

Wiederum setzen wir voraus, daß P zentralsymmetrisch ist. Unter den affin äquivalenten Exemplaren von P, die in einem Einheitskreis E enthalten sind, sei T dasjenige vom größtmöglichen Inhalt. Wir wollen zeigen, daß der Umfang U von T wenigstens $4\sqrt{2}$ beträgt.

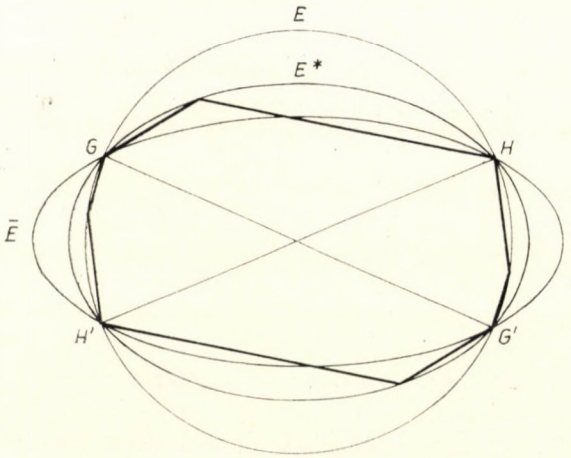


Fig. 2

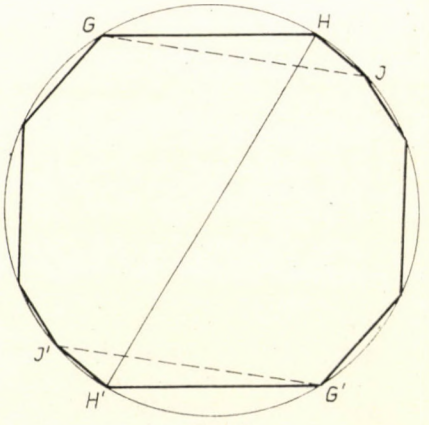


Fig. 3

Es sei b die Länge eines größten Bogens auf dem Rand von E, der außer seinen Endpunkten G und H keine Ecke von T enthält. Wir behaupten, daß $b\leqq\pi/2$ ist. Um dies einzusehen betrachten wir die G und H diametral gegenüberliegenden Punkte G' und H', sowie die Schar der durch G, H, G' und H' hindurchgehenden Ellipsen. Bezüglich der Ellipse $\bar{E}$ vom kleinstmöglichen Inhalt sind GG' und HH' konjugierte Durchmesser. Wäre nun $b>\pi/2$, so gäbe es in der E und $\bar{E}$ verbindenden Teilschar eine T enthaltende Ellipse E^* (Fig. 2). Da aber der Inhalt von E^* kleiner ist als π, würde die Affinität, die E^* in E überführt, T in ein in E liegendes Polygon von grösserem Inhalt überführen, was der Definition von T wiederspricht.

Wir ersetzen jetzt T durch die konvexe Hülle seiner auf dem Rand von E liegenden Ecken. Dadurch nimmt U offensichtlich nicht zu. Ist $b=\pi/2$, so ist $GHG'H'$ ein Quadrat von der Seitenlänge $\sqrt{2}$; folglich ist $U\geqq 4\sqrt{2}$. Ist dagegen $b<\pi/2$, so bewegen wir die Ecke H, zusammen mit der antipodischen Ecke H', so daß b zunimmt, bis entweder $b=\pi/2$ wird oder H und H' mit den nächsten Ecken J bzw. J' von T zusammenfallen (Fig. 3). Wegen $GH\geqq HJ$ nimmt bei dieser Bewegung die Länge des Streckenzuges GHJ zusammen mit U ab. Ist für das neue Polygon

T wieder $b > \pi/2$, so wiederholen wir die obige Operation sukzessiv, bis $b = \pi/2$ wird. Da in jedem Schritt U abnimmt und in der Endlage $U \geqq 4\sqrt{2}$ ausfällt, ist der Beweis der Ungleichung $U \geqq 4\sqrt{2}$ erbracht.

Nun haben wir für jede Gerade u $A(u, T) \leqq 2$, also

$$M(P) = M(T) = \sum \frac{a(u, T)}{A(u, T)} \geqq \frac{1}{2} \sum a(u, T) = \frac{1}{2} U \geqq \frac{1}{2} \cdot 4\sqrt{2}.$$

Aus den obigen Überlegungen folgt leicht, daß in der Ungleichung $U \geqq 4\sqrt{2}$ Gleichheit nur für ein Quadrat gilt. Ist aber T ein Quadrat, so ist $M(T) = 4$. Deshalb kann in der Ungleichung $M(P) \geqq \sqrt{8}$ Gleichheit nicht bestehen.

Damit sind die Sätze 1 und 2 bewiesen.

§ 5. Polyeder

Aus der Invarianz von $M(P)$ für Polygone gegenüber Zentralsymmetrisierung folgt unmittelbar die entsprechende Invarianz für dreidimensionale Prismen. Im allgemeinen ist aber $M(A)$ für $k \geqq 3$ gegenüber Zentralsymmetrisierung nicht invariant. Z. B. geht das Tetraeder T durch Symmetrisierung in bezug auf einen Punkt in das Kuboktaeder K über. Nun gilt aber $M(T) = 4 < 13/4 = M(K)$. Als zweites Beispiel betrachten wir eine regelmäßige dreiseitige Doppelpyramide D, die durch Zentralsymmetrisierung in eine sechsseitige Doppelpyramide S übergeht. Nach einer einfachen Rechnung gilt jetzt $M(D) = 11/2 > 24/5 = M(S)$. Diese Beispiele zeigen, daß die Methode der Zentralsymmetrisierung bei unseren Problemen für $k > 2$ versagt.

Obwohl wir in dieser Hinsicht weder über genügende Erfahrung noch über einen Beweisansatz verfügen, scheint es zweckmäßig zu sein, als Arbeitshypothese die Vermutung auszusprechen, daß für einen k-dimensionalen konvexen Polyeder A stets $M(A) \leqq 2k$ gilt und Gleichheit nur für ein Parallelotop erreicht wird.

Bezüglich $\inf M(A)$, erstreckt über die eigentlichen k-dimensionalen konvexen Polyeder, verfügen wir für $k > 2$ nicht einmal über eine schwache Vermutung. Für einen beliebigen k-dimensionalen konvexen Körper A gilt natürlich die triviale Ungleichung $M(A) \geqq 2$, wobei Gleichheit nur für entartete Körper gilt. Wir sahen aber, daß für eigentliche zweidimensionale Körper $\inf M > 2$ ist, und vermutlich gilt dies auch in höheren Dimensionen.

Es erhebt sich hier die naheliegende Frage, ob nicht das Simplex als extremales Polyeder in Betracht kommt, ob also nicht für einen eigentlichen k-dimensionalen Eipolyeder A stets $M(A) \geqq k+1$ ausfällt. Die Gültigkeit dieser Ungleichung wird aber sofort zweifelhaft, wenn man bedenkt, daß der Wert von M außer einem Tetraeder auch für eine dreidimensionale Kugel 4 beträgt. Bezeichnen wir den Wert von M für eine k-dimensionale Kugel mit M_k, so gilt die Rekursionsformel

$$M_{k+1} M_k = 2\pi k$$

mit dem Anfangswert $M_1 = 2$ (s. z. B. [2], [11] oder [12]). Hieraus ergibt sich $M_2 = \pi$, $M_3 = 4$, $M_4 = \frac{3}{2}\pi \approx 4{,}7$, $M_5 = \frac{16}{3} \approx 5{,}3$, $M_6 = \frac{15}{8}\pi \approx 5{,}9, \ldots$. Diese Werte lassen vermuten, dass für $k > 3$ $M_k < k+1$ ausfällt, und es ist nicht schwer, dies auch

streng nachzuweisen. Hieraus folgt sofort durch polyedrische Approximation, daß für $k>3$ die Ungleichung $M(A) \geqq k+1$ auch innerhalb der Klasse der eigentlichen Eipolyeder nicht allgemein gelten kann. Wir wollen jetzt dasselbe auch für $k=3$ zeigen.

Nach mehreren erfolglosen Versuchen habe ich es aufgegeben, unter den Platonischen, Archimedischen und anderen bekannten Körpern ein konvexer Polyeder P mit $M(P)<4$ zu finden, oder einen solchen Polyeder zu konstruieren. (Es stellte sich dabei u.a. heraus, daß der Wert von M auch für den Rhombendodekaeder genau 4 und für den Rhombentriakontaeder [11] $\dfrac{30}{2\sqrt{5}+3} \approx 4{,}015$ beträgt.) Wir zeigen hier die Existenz eines konvexen Polyeders P mit $M(P)<4$ auf indirekte Weise, indem wir einen konvexen Rotationskörper R mit $M(R)<4$ konstruieren.

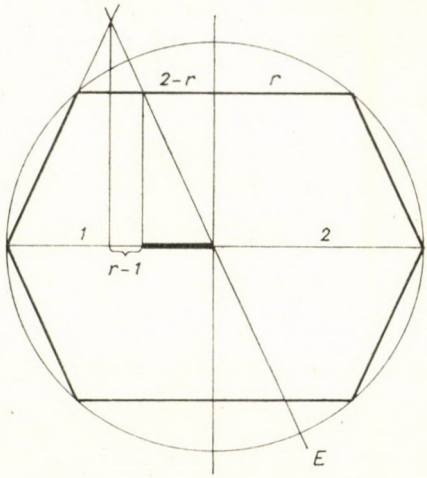


Fig. 4

R sei ein Doppelkegelstumpf, der in einer Kugel vom Radius 2 durch den Äquator und zwei kongruente Breitenkreise vom Radius r aufgespannt wird. Wir legen durch den Kugelmittelpunkt eine zu einer Erzeugenden des nördlichen Kegels parallele Ebene E. Diese schneidet den nördlichen Kegel in einer Parabel, deren senkrechte Projektion auf die Äquatorebene sich mit der Gleichung $y=2\sqrt{x}$ darstellen läßt. Wir betrachten dasjenige Segment dieser Parabel, das sich durch Projektion des Durchschnittes von E mit dem nördlichen Teil von R ergibt (Fig. 4). Ist $r \geqq 1$, so ist der Inhalt dieses Segments

$$2\int_{r-1}^{1} 2\sqrt{x}\,dx = \frac{8}{3}[1-(r-1)^{3/2}].$$

Da ferner die Projektion des Mantels von R ein Kreisring vom Inhalt $\pi(4-r^2)$ ist, beträgt der Wert des Integrals $\int \dfrac{df}{F}$, erstreckt über den ganzen Mantel von R

$$\frac{\pi(4-r^2)}{\frac{8}{3}[1-(r-1)^{3/2}]}.$$

Nehmen wir noch die beiden Kreisflächen von R in Betracht, so ergibt sich

$$M(R) = \frac{1}{2}r^2 + \frac{3\pi(4-r^2)}{8[1-(r-1)^{3/2}]}, \quad 1 \leqq r < 2.$$

Nun hat aber die rechtsstehende Funktion von r in der Nähe von $r=1{,}085$ ein Minimum, dessen Näherungswert 3,9987 beträgt. Damit ist die gewünschte Konstruktion beendet.

Eine zu $M(A)$ analoge Größe ist die Summe $S(A)=\Sigma l/L$, wo l die Länge einer Kante des Polyeders A und L die Länge der zu dieser Kante parallelen größten Sehne bedeutet und die Summation sich auf sämtliche Kanten von A erstreckt. Zusammen mit $M(A)$ und $S(A)$ können wir für einen k-dimensionalen Polyeder $k-1$ ähnliche affininvariante Größen definieren, indem wir die i-dimensionalen Zellen und die parallelen i-dimensionalen Maximalschnitte von A in Betracht ziehen $(1 \leqq i \leqq k-1)$. Vermutlich gilt für ein dreidimensionales Eipolyeder A stets $S(A) \leqq 6$ mit Gleichheit nur für einen Tetraeder. Dagegen zeigt das Beispiel der Prismen, daß für $S(A)$ eine universale obere Schranke nicht existiert.

Wir sprechen noch eine weitere Vermutung bezüglich einer Extremaleigenschaft des Tetraeders aus: Mit den Bezeichnungen von § 2 gilt für einen dreidimensionalen Eipolyeder A $\Sigma sf/F \geqq 8/3$. Zusammen mit unserer obigen Vermutung bezüglich der oberen Schranke von $M(A)$ würde hieraus folgen, daß

$$\frac{8}{3} \leqq \lim_{\lambda \to 0} \lambda^2 H(\lambda A, A) \leqq 6.$$

Zum Schluß verweisen wir auf einen Aufsatz von BANTEGNIE [16], der mit der vorliegenden Arbeit gewisse Berührungspunkte aufweist.

LITERATURVERZEICHNIS

[1] FEJES TÓTH, L.: Remarks on a theorem of R. M. Robinson, *Stud. Sci. Math. Hungar.* **4** (1969) 441—445.

[2] FEJES TÓTH, L.: *Reguläre Figuren*, Akadémiai Kiadó, Budapest, Teubner, Leipzig 1965.

[3] BÖRÖCZKY, K.: Über die Newtonsche Zahl regelmäßiger Vielecke, *Stud. Sci. Math. Hung.*

[4] 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) 53—71.

[5] FEJES TÓTH, L.: On the number of equal discs that can touch another of the same kind, *Stud. Sci. Math. Hungar.* **2** (1967) 363—367.

[6] SCHOPP, J.: Über die Newtonsche Zahl von Bereichen konstanter Breite, *Stud. Sci. Math. Hungar.* **5** (1970)

[7] HADWIGER, H. und DEBRUNNER, H.: Kombinatorische Geometrie in der Ebene (Genève 1959), Problem 43.

[8] GRÜNBAUM, B.: On a conjecture of Hadwiger, *Pacific J. Math.* **11** (1961) 215—219.

[9] GROEMER, H.: Existenzsätze für Lagerungen im Euklidischen Raum, *Math. Zeitschrift* **81** (1963) 260—278.

[10] ROGERS, C. A.: The closest packing of convex two-dimensional domains, *Acta Math.* **86** (1951) 309—321.

[11] COXETER, H. S. M.: *Regular Polytopes* (Macmillan, New York, Collier-Macmillan London, 1962. Second edition).

[12] HADWIGER, H.: *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie* (Springer, Berlin—Göttingen—Heidelberg, 1957).

[13] JAGLOM, I. M. und BOLTJANSKI, V. G.: *Konvexe Figuren* (VEB Deutscher Verlag der Wiss. Berlin, 1956).

[14] HADWIGER, H.: *Altes und Neues über konvexe Körper* Birkhäuser Basel-Stuttgart, 1955).

[15] GROEMER, H.: Abschätzung für die Anzahl der konvexen Körper die einen konvexen Körper berühren, *Monatsh. Math.* **65** (1961) 74—81.

[16] BANTEGNIE, R.: Sur les configurations de Hadwiger, *Archiv Math.* **18** (1967) 534—538.

Mathematisches Institut der Ungarischen Akademie der Wissenschaften, Budapest

(Eingegangen: 3. März 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 181—195

GRAPHS IN WHICH EACH PAIR OF VERTICES IS ADJACENT TO THE SAME NUMBER d OF OTHER VERTICES

by

R. C. BOSE and S. S. SHRIKHANDE

1. Introduction. A strongly regular graph may be defined as a finite regular graph with v vertices and of valence n_1, such that each pair of adjacent vertices is adjacent to exactly p_{11}^1 other vertices, and each pair of non-adjacent vertices is adjacent to exactly p_{11}^2 vertices. This paper deals with some methods of constructing strongly regular graphs from other strongly regular graphs. In particular we consider graphs for which $p_{11}^1 = p_{11}^2 = d$, so that each pair of vertices (whether adjacent or not) is adjacent to exactly d other vertices. The problem of investigating such graphs has recently been proposed by SZEKERES. ERDŐS, RÉNYI and SÓS [9] have shown that if there is a graph G (not necessarily regular) with the property that every pair of vertices is adjacent to exactly one other vertex then G must consist of n triangles with a common vertex. G has thus $2n+1$ vertices which may be labeled by the integers $0, 1, \ldots, 2n$ such that the edges are the pairs $(0, 1), (0, 2), \ldots, (0, 2n), (1, 2), (3, 4), \ldots, (2n-1, 2n)$. Also it has been shown by ERDŐS (in a private communication) that if $d > 1$, and every pair of vertices of a graph G is adjacent to exactly d other vertices, then G must be regular, and hence strongly regular. We shall denote such a graph by $G_2(d)$. If the number of vertices is v and the valence is n_1, we say that v, n_1, d are the parameters of the $G_2(d)$ graph. We prove:

THEOREM (1. 1). *If $G_2(d)$ is a finite graph without loops or multiple edges, in which each pair of distinct vertices is adjacent to exactly d other vertices, $d \geqq 2$, then $G_2(d)$ is regular of valence n_1 such that $v-1 = n_1(n_1-1)/d$ where v is the number of vertices and there exists a positive integer m, such that*

(i) $n_1 = d+m^2$

(ii) $\dfrac{d}{m}$ *is an integer, with the same parity as* $v-1-m$.

We then consider methods of obtaining such graphs. A number of examples can readily be obtained from known configurations, and graphs of certain partial geometrics [2], but to obtain others it is necessary to use the principle of switching which was first introduced by SEIDEL [14]. In particular we show the existence of $G_2(d)$ graphs with parameters

(i) $\qquad v = 4r^2, \quad n_1 = r(2r-1), \quad d = r(r-1)$

(ii) $\qquad v = 4r^2, \quad n_1 = r(2r+1), \quad d = r(r+1)$

(iii) $\qquad v = 4r^2-1, \quad n_1 = 2r^2, \quad d = r^2$

for all $r = 3^m \cdot 2^{m+n-1}$, where m and n are any non-negative integers, $(m, n) \neq (0, 0)$.

2. Regularity of $G_2(d)$. The regularity of $G_2(d)$ was proved by ERDŐS, but a slightly different proof is included here for completeness.

Suppose $G_2(d)$ has v vertices, $1, 2, \ldots, v$. Let S_i be the set of vertices adjacent to the vertex i, and let $\mathcal{S}=\{S_1, S_2, \ldots, S_v\}$ be the class of sets S_i. Then any two distinct sets of $\mathcal{S}$ must have exactly d vertices in common, and conversely every pair of distinct vertices appears in exactly d distinct sets of $\mathcal{S}$. Consider a particular set S_i of $\mathcal{S}$. Let $|S_i|=n_1$. Let us count the number of symbols $\{(l, m), S_u\}$, (l, m) being an unordered pair of vertices belonging to both S_i and S_u, where S_u belongs to $\mathcal{S}$, $u\neq i$. We can choose (l, m) belonging to S_i in $n_1(n_1-1)/2$ ways. Now each (l, m) must occur in exactly $d-1$ of the sets of $\mathcal{S}$ other than S_i. Hence the number of symbols is $(d-1)(n_1-1)n_1/2$. Again since S_i intersects any S_u $(u\neq i)$ in exactly d elements, for any fixed S_u we get $d(d-1)/2$ symbols. Hence the number of symbols is $(v-1)d(d-1)/2$. Thus for $d\geqq 2$ we have $v-1=n_1(n_1-1)/d$. This uniquely determines n_1. Hence $G_2(d)$ is regular. This proves the first part of Theorem (1. 1).

3. Strongly regular graphs. A finite connected graph G (without loops or multiple edges) is called strongly regular [1], if it is regular of valence n_1, and any pair of adjacent vertices is adjacent to exactly p_{11}^1 other vertices, and any pair of non-adjacent vertices is adjacent to exactly p_{11}^2 vertices. If v is the number of vertices, we shall call $(v, n_1, p_{11}^1, p_{12}^2)$ the parameters of the strongly regular graph G.

The complementary graph $\bar{G}$ of G is regular of valence $n_2=v-n_1-1$. Now let us call two vertices of G first associates if they are adjacent and second associates if they are non-adjacent. If the vertices l, m of G are i-th associates let $p_{jk}^i(l, m)$ denote the number of vertices which are simultaneously j-th associates of l and k-th associates of m $(i, j, k=1, 2)$. Easy counting arguments show [3], [6] that if G is strongly regular with parameters $(v, n_1, p_{11}^1, p_{11}^2)$, then all the numbers $p_{jk}^i(l, m)$ are constants independent of l, m and that the following relations hold:

$$p_{12}^1=p_{21}^1, \quad p_{12}^2=p_{21}^2, \tag{3. 1}$$

$$p_{11}^1+p_{12}^1=n_1-1, \quad p_{21}^1+p_{22}^1=n_2, \quad p_{11}^2+p_{12}^2=n_1, \quad p_{21}^2+p_{22}^2=n_2-1, \tag{3. 2}$$

$$n_1p_{12}^1=n_2p_{11}^2, \quad n_1p_{22}^1=n_2p_{12}^2. \tag{3. 3}$$

It is clear that $\bar{G}$ is strongly regular with parameters $(v, n_2, p_{22}^2, p_{22}^1)$. The graphs $G_2(d)$ which we want to investigate are strongly regular with parameters (v, n_1, d, d).

4. PROOF of theorem (1. 1). The adjacency matrix of a finite graph G with v vertices is a $v\times v$ matrix $A=(a_{ij})$, $i, j=1, 2, \ldots, v$, such that $a_{ij}=1$ if $i\neq j$ and i and j are adjacent, and $a_{ij}=0$ otherwise. In particular $a_{ii}=0$.

Let A be the adjacency matrix of a $G_2(d)$ graph with v vertices, and valence n_1. Let J be the $v\times v$ matrix for which each element is unity. Then it is readily seen that

$$A^2=(n_1-d)I+dJ$$

Hence

$$A^3=(n_1-d)A+n_1dJ$$

$$A^3-n_1A^2-(n_1-d)A+n_1(n_1-d)I=0.$$

Thus A has only three distinct eigenvalues n_1 and $\pm(n_1-d)^{\frac{1}{2}}$. Now from the regularity of $G_2(d)$ it follows that $A^*=A/n_1$ is a stochastic matrix, and since

$G_2(d)$ is connected A^* is irreducible. It follows [8] that unity is a simple root of A^*, so that n_1 is a simple root of A. Let α_1, α_2 be the multiplicities of the roots $\theta_1 = (n_1-d)^{\frac{1}{2}}$ and $\theta_2 = -(n_1-d)^{\frac{1}{2}}$. Then

$$|A-I\theta| = (n_1-\theta)(\theta_1-\theta)^{\alpha_1}(\theta_2-\theta)^{\alpha_2}.$$

To determine α_1 and α_2 we note that

$$\operatorname{Tr} I = 1+\alpha_1+\alpha_2 = v,$$

$$\operatorname{Tr} A = n_1+\alpha_1\theta_1+\alpha_2\theta_2 = 0.$$

$$\therefore \alpha_1 = \frac{v-1}{2} - \frac{n_1}{2(n_1-d)^{1/2}}, \quad \alpha_2 = \frac{v-1}{2} + \frac{n_1}{2(n_1-d)^{1/2}}.$$

Since the multiplicities are necessarily integral, there must exist an integer m such that $n_1 = d+m^2$. Also since

$$2\alpha_2 = (v-1) - \left(m+\frac{d}{m}\right)$$

d/m must be integral, and the integers $v-1-m$ and d/m must have the same parity. This completes the proof of Theorem (1. 1).

5. Some examples of $G_2(d)$ graphs. (i) If we take $m=1$, in Theorem (1. 1), then $n_1 = d+1$, $v = d+2$. Then $G_2(d)$ is obviously the complete graph with $d+2$ vertices.

(ii) If we take $d=2$, $m\neq 1$, then since d is divisible by m, the only possible value of m is 2. This gives $n_1=6$, $v=16$. It is known that there are only two non-isomorphic strongly regular graphs with parameters (16, 6, 2, 2), [14] and [15]. They are given as follows:

(a) Arrange the integers 1 through 16 corresponding to the vertices of the graph in a 4×4 scheme.

(5. 1)

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Then any two vertices are adjacent if and only if the corresponding integers are in the same row or same column.

(b) Arrange the integers corresponding to the vertices in a 4×4 scheme, and impose on it a cyclic Latin square. We thus obtain

(5. 2)

1A	2B	3C	4D
5D	6A	7B	8C
9C	10D	11A	12B
13B	14C	15D	16A

Two vertices are adjacent if they do not occur in the same row or in the same column, and do not come together with the same letter of the cyclic square.

(iii) Consider the finite projective space $PG(3, 2)$ of three dimensions based on the Galois field of order 2. This space has 15 points and 35 lines. We may take the lines to correspond to the vertices of a graph where two vertices are adjacent if and only if the corresponding lines intersect. It is readily seen that each line is intersected by 18 lines, and any pair of lines is intersected by 9 lines. Hence our graph is $G_2(d)$ with $v=35$, $n_1=18$, $d=9$.

6. $G_2(d)$ graphs derived from partial geometries. A partial geometry (r, k, t) is a system of points and lines satisfying the following axioms:

*A*1. Any two points are not incident with more than one line.

*A*2. Each point is incident with r lines.

*A*3. Each line is incident with k points.

*A*4. If the point P is not incident with the line l, there are exactly t lines through P intersecting l.

The graph of a partial geometry (r, k, t) is obtained by taking the vertices of the graph to correspond to the points of the partial geometry, and taking two vertices to be adjacent if and only if they are incident with the same line of the geometry. Partial geometries have been discussed in [2]. It was shown that the graph of a partial geometry is strongly regular with parameters

$$v = k[(r-1)(k-1)+t]/t, \quad n_1 = r(k-1), \tag{6. 1}$$

$$p_{11}^1 = (t-1)(r-1)+k-2, \quad p_{11}^2 = rt. \tag{6. 2}$$

We will call it a geometric graph $G(r, k, t)$.

A strongly regular graph G with the parameters (6. 1), (6. 2) for some positive integral values of r, k, t is defined to be pseudo-geometric. Of course it may or not be the graph of a partial geometry. It will be called a pseudo-geometric graph $G(r, k, t)$. If for a pseudo-geometric graph $G(r, k, t)$ the condition $k = r+t+1$ is satisfied then $p_{11}^1 = p_{11}^2 = rt$ and we will have an example of the kind desired. Many examples of partial geometries are known satisfying the required condition.

(a) Consider an elliptic non-degenerate quadric in the finite projective space $PG(5, q)$ where q is a prime power. It is known [11], [13] that this quadric is ruled by straight lines called generators, but contains no planes. The number of points on the surface is $(q^3+1)(q^2+1)$. It was shown in [2], that the points and generators can be regarded as the points and lines of a partial geometry $(q^2+1, q+1, 1)$. The dual of this partial geometry is obtained by taking the points and lines of the dual to be the lines and points of the original geometry. Thus the dual partial geometry has the parameters $(q+1, q^2+1, 1)$. Another way to obtain a partial geometry with the same parameters is to take the surface $x_0^{q+1}+x_1^{q+1}+x_2^{q+1}+x_3^{q+1} = 0$ in $PG(3, q^2)$. It has been shown in [4], that this surface contains $(q^2+1)(q^3+1)$ points, and $(q+1)(q^3+1)$ generators which may be taken to be points and lines of a partial geometry $(q+1, q^2+1, 1)$. If we take $q=2$, the condition $p_{11}^1 = p_{11}^2$ is satisfied. The graph of this partial geometry is strongly regular with parameters (45, 12, 3, 3). Thus we get a $G_2(d)$ with $v=45$, $n_1=12$, $d=3$. Here $m=3$.

(b) A net of degree r and order k is a system of undefined points and lines together with an incidence relation subject to the following axioms (i) Each line

is incident with k points, $k \geqq 1$ (ii) The lines of the net can be partitioned into r disjoint, non-empty "parallel classes" such that each point of the net is incident with exactly one line of each class, (iii) Given two lines belonging to distinct classes there is exactly one point of the net which is incident with both lines. Then it readily follows that the number of points is k^2, the number of lines is rk, which fall into r parallel classes of k lines each. It is well known that the existence of a net of degree r and order k is equivalent to the existence of $(r-2)$ mutually orthogonal Latin squares of order k. If we superpose the Latin squares then each cell contains $r-2$ symbols, belonging in order to the different Latin squares. The k^2 cells can now be identified with the k^2 points of the net. Points belonging to the same row give one set of k parallel lines. Points belonging to the same column give another set of k parallel lines. Cells which contain the same symbol of the i-th Latin square give a set of parallel lines for each value of i ($i = 1, 2, \ldots, r-2$). We thus get r classes of parallel lines.

The points and lines of a net of degree r and order k satisfy the axioms of a partial geometry with parameters $(r, k, r-1)$. The graph of this partial geometry is strongly regular with parameters

$$v = k^2, \quad n_1 = r(k-1), \tag{6. 3}$$

$$p_{11}^1 = (r-2)(r-1)+(k-2), \quad p_{11}^2 = r(r-1). \tag{6. 4}$$

Such a graph will be called a net graph $L_r(k)$. Any strongly regular graph (not necessarily the graph of a net) with the parameters (6. 3), (6. 4) will be called a pseudo net graph, $L_r(k)$. The condition $k = r+t+1$ reduces to $k = 2r$. Thus a pseudo net graph $L_r(2r)$ is a $G_2(d)$ graph with $v = 4r^2$, $n_1 = r(2r-1)$, $d = r(r-1)$. We can thus get $G_2(d)$ graphs for all values of r for which there exist $r-2$ mutually orthogonal Latin squares of order $2r$. This is always true if $r = 2^m$ where m is a non-negative integer. We can therefore obtain a $G_2(d)$ graph with $v = 2^{2m+2}$, $n_1 = 2^m(2^{m+1}-1)$, $d = 2^m(2^m-1)$. The example (ii) (a) of paragraph 6, is a special case of this. Again since there exists a Latin square of order 6, we get a $G_2(d)$ graph with $v = 36$, $n_1 = 15$, $d = 6$. Also the existence of 5 mutually orthogonal squares of order 12 is known [5]. By taking 4 mutually orthogonal squares of order 12, we can get a $G_2(d)$ graph with $v = 144$, $n_1 = 66$, $d = 30$.

(c) A balanced incomplete block design (BIB) is an arrangement of a set of v_0 objects or treatments in b_0 sets or blocks, such that (i) each block contains k_0 distinct treatments (ii) each treatment is contained in r_0 blocks (iii) each pair of distinct treatments is contained in λ_0 blocks. We say that the design has parameters $(v_0, b_0, r_0, k_0, \lambda_0)$. The dual of a design is defined as a new design whose traetments and blocks are in (1, 1) correspondence with the blocks and treatments of the original design, and incidence is preserved (a block and treatment are incident if the treatment is contained in the block and non-incident otherwise). It is known [2] that the dual of a BIB design with $\lambda_0 = 1$ can be regarded as a partial geometry (r, k, r) where $r = k_0$, and $k = r_0$, the lines of the partial geometry being the blocks of the dual design.

Such a dual design (or partial geometry) may be called a linked block design (or geometry). The corresponding strongly regular graph has the parameters

$$v = k(kr-k+1)/r, \quad n_1 = r(k-1), \tag{6. 5}$$

$$p_{11}^1 = (k-2)+(r-1)^2, \quad p_{11}^2 = r^2. \tag{6. 6}$$

It will be called a linked block graph and will be denoted by $LB_r(k)$. Any strongly regular graph (not necessarily the graph of a linked block design) will be called a pseudo linked block graph $LB_r(k)$. The condition $k = r+t+1$ now reduces to $k = 2r+1$. Thus a pseudo linked block graph $LB_r(2r+1)$ is a $G_2(d)$ graph with $v = 4r^2-1$, $n_1 = 2r^2$, $d = r^2$. Now BIB designs with parameters $v_0 = 2^{m-1}(2^m-1)$, $b_0 = 2^{2m}-1$, $r_0 = 2^m+1$, $k_0 = 2^{m-1}$, $\lambda_0 = 1$ are known [7] for every integral value of m. We can therefore get a corresponding $G_2(d)$ graph with $v = 2^{2m}-1$, $n_1 = 2^{2m-1}$, $d = 2^{2m-2}$ for all integral m. Also BIB designs with parameters

$$(6.7) \qquad v_0 = r(2r-1), \quad b_0 = 4r^2-1, \quad r_0 = 2r+1, \quad k_0 = r, \quad \lambda_0 = 1$$

are known for values of $k_0 = 2, 3, 4, 5$ and 7 [1], [12]. Hence the corresponding $G_2(d)$ graphs with parameters $v = 4r^2-1$, $n_1 = 2r^2$, $d = r^2$ can be constructed for $r = 2, 3, 4, 5,$ and 7.

We shall now show that for an infinity of values of r pseudo net graphs $L_r(2r)$, and pseudo linked block graphs $LB_r(2r+1)$ can be obtained even when $r-2$ orthogonal Latin squares of order $2r$, or a BIB design with parameters (6.7) is unknown. However to do this we first need the concept of Seidel equivalence of strongly regular graphs, and some theorems relating to this equivalence. As a preliminary to this we give some definitions and notations in the next paragraph.

7. Notations and definitions. If M is a 0, 1 matrix we shall denote by $\overline{M}$ the matrix obtained from M by interchanging zeros and ones. In particular M may be a vector.

For any 0, 1 vector, $\alpha = (a_1, a_2, \ldots, a_n)$ we shall denote by $w(\alpha)$ the number of non-zero coordinates in α and call this number the weight of α. The weight of the i-th row of M will be denoted by $w_i[M]$. Clearly

$$(7.1) \qquad w(\alpha)+w(\bar{\alpha}) = n, \quad w_i[M]+w_i[\overline{M}] = n$$

where M is an $m \times n$ matrix.

Given two 0, 1 vectors $\alpha = (a_1, a_2, \ldots, a_n)$, $\beta = (b_1, b_2, \ldots, b_n)$ the Hamming distance $\delta(\alpha, \beta)$ between α and β is defined to be the number of coordinates in which α and β disagree. Clearly

$$(7.2) \qquad \delta(\bar{\alpha}, \bar{\beta}) = \delta(\alpha, \beta), \quad \delta(\alpha, \bar{\beta}) = \delta(\bar{\alpha}, \beta) = n - \delta(\alpha, \beta).$$

Again if $(\alpha \cdot \beta)$ denotes the scalar product of α and β

$$(7.3) \qquad \delta(\alpha, \beta) = w(\alpha)+w(\beta)-2(\alpha \cdot \beta).$$

We shall denote by $\sigma_{ij}[M]$ the scalar product of the i-th and j-th rows of M, and by $\delta_{ij}[M]$ the Hamming distance between the i-th and j-th rows of M. Then

$$(7.4) \qquad \delta_{ij}[M] = w_i[M]+w_j[M]-2\sigma_{ij}[M].$$

Again if M_1 and M_2 are two 0, 1 matrices each with the same number of rows, but with possibly different number of columns, then

$$(7.5) \qquad w_i[M_1, M_2] = w_i[M_1]+w_i[M_2], \quad \delta_{ij}[M_1, M_2] = \delta_{ij}[M_1]+\delta_{ij}[M_2].$$

8. Seidel equivalence of strongly regular graphs. Let G be a strongly regular graph with parameters

$$v, n_1, p_{11}^1, p_{11}^2.$$

We can obtain another graph G^* from it by the following process: Let the set of vertices V of G be divided into disjoint subsets V_1 and V_2, $V = V_1 \cup V_2$. G^* has the same set of vertices as G. Two vertices of G^* both of which belong to V_1 or to V_2 are adjacent or non-adjacent in G^* according as they are adjacent or non-adjacent in G. Two vertices of G^* one of which belongs to V_1 and the other to V_2 are adjacent in G^* if they are non-adjacent in G, and non-adjacent in G^* if they are adjacent in G. Then G^* may be said to be derived from G by complementation with respect to V_1 and V_2. If G^* is strongly regular it is defined to be Seidel equivalent to G, or more briefly S-equivalent to G.

Let $|V_1| = v_1$, $|V_2| = v_2$, then $v = v_1 + v_2$. In writing down the adjacency matrix of G, we may take the first v_1 rows (columns) to correspond to the vertices in V_1 and the last v_2 rows (columns) to correspond to the vertices in V_2. Then we can write the adjacency matrix of G as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{8. 1}$$

where A_{11} and A_{22} are square matrices of order v_1 and v_2 respectively, A_{12} is a $v_1 \times v_2$ matrix, and $A_{21} = A_{12}'$. Then clearly the adjacency matrix of G^* is

$$A^* = \begin{bmatrix} A_{11} & \bar{A}_{12} \\ \bar{A}_{21} & A_{22} \end{bmatrix}. \tag{8. 2}$$

We wish to investigate the conditions under which G^* is strongly regular and therefore by definition Seidel equivalent to G. Let $1 \leqq i \leqq v_1$, $1 \leqq j \leqq v_2$.

$$w_i[\bar{A}_{12}] = v_2 - w_i[A_{12}], \quad w_j[\bar{A}_{21}] = v_1 - w_j[A_{21}].$$

Again since G is regular and of degree n_1,

$$w_i[A_{12}] = n_1 - w_i[A_{11}], \quad w_j[A_{21}] = n_1 - w_j[A_{22}].$$

Hence for $1 \leqq i \leqq v_1$,

$$w_i[A^*] = w_i[A_{11}] + w_i[\bar{A}_{12}] = w_i[A_{11}] + v_2 - w_i[A_{12}] = 2w_i[A_{11}] + v_2 - n_1;$$

and for $1 \leqq j \leqq v_2$,

$$\begin{aligned} w_{v_1+j}[A^*] &= w_j[\bar{A}_{21}] + w_j[A_{22}] \\ &= v_1 - w_j[A_{21}] + w_j[A_{22}] \\ &= v_1 - n_1 + 2w_j[A_{22}]. \end{aligned}$$

If G^* is regular then $w_i[A^*]$ is independent of i. Hence each row of A_{11} has the same weight, say w_1, i.e., $w_i[A_{11}] = w_1$. Similarly $w_{v_1+j}[A^*]$ is independent of j. Hence each row of A_{22} has the same weight, say w_2, i.e., $w_j[A_{22}] = w_2$. Moreover $w_i[A^*] = w_{v_1+j}[A^*]$. Hence

$$w_1 - w_2 = (v_1 - v_2)/2. \tag{8. 3}$$

It is also readily seen that these conditions are sufficient for the regularity of G^*. Hence we state:

For G^* to be regular it is necessary and sufficient that each row of A_{11} has the same weight w_1, each row of A_{22} has the same weight w_2 and $w_1 - w_2 = (v_1 - v_2)/2$.

Then the degree n_1^* of G^* is given by

$$n_1^* = 2w_1 + v_2 - n_1 = 2w_2 + v_1 - n_1. \tag{8. 4}$$

Note that $w_i[A_{11}] = w_1$, means that each vertex of G belonging to V_1 is adjacent to w_1 vertices in V_1, and $n_1 - w_1$ vertices in V_2. In the same way $w_j[A_{22}] = w_2$ means that each vertex of G belonging to V_2 is adjacent to w_2 vertices in V_2 and to $n_1 - w_2$ vertices in V_1.

Now supposing G^* to be regular and of degree n_1^* let us investigate the further conditions which must be satisfied in order to make G^* strongly regular.

Let $x_1, x_2, \ldots, x_{v_1}, x_{v_1+1}, \ldots, x_v$ be the vertices of G, the first v_1 belonging to V_1 and the last v_2 belonging to V_2, where $v = v_1 + v_2$. Now from (7. 4)

$$2\sigma_{ij}[A] = w_i[A] + w_j[A] - \delta_{ij}[A] = 2n_1 - \delta_{ij}[A], \tag{8. 5}$$

$$2\sigma_{ij}[A^*] = w_i[A^*] + w_j[A^*] - \delta_{ij}[A^*] = 2n_1^* - \delta_{ij}[A^*]. \tag{8. 6}$$

If x_i and x_j both belong to V_1 or both belong to V_2, i.e., $1 \leqq i,\ j \leqq v_1$ or $v_1 + 1 \leqq i,\ j \leqq v$, we call it case I. From (7. 2) and (7. 5)

$$\delta_{ij}[A] = \delta_{ij}[A^*].$$

Hence from (8. 5) and (8. 6) it follows that in case I

$$\sigma_{ij}[A^*] = n_1^* - n_1 + \sigma_{ij}[A].$$

Again if x_i belongs to V_1 and x_j belongs to V_2, i.e., $1 \leqq i \leqq v_1$, $v_1 + 1 \leqq j \leqq v$ we will call it case II. Then from (7. 2) and (7. 5)

$$\delta_{ij}[A] = v - \delta_{ij}[A^*].$$

Hence from (8. 5) and (8. 6) it follows that in case II

$$\sigma_{ij}[A^*] = n_1^* + n_1 - \sigma_{ij}[A] - \frac{v}{2}.$$

Case I can be sub-divided into cases I(a) and I(b) according as x_i and x_j are adjacent or non-adjacent in G. Similarly case II can be sub-divided into cases II(a) and II(b) according as x_i and x_j are adjacent or non-adjacent in G. Then x_i and x_j will be adjacent in G^* in cases I(a) and II(b) and will be non-adjacent in G^* in cases I(b) and II(a). Now $\sigma_{ij}[A] = p_{11}^1$ or p_{11}^2 according as x_i and x_j are adjacent or non-adjacent in G. Hence

$$\sigma_{ij}[A^*] = n_1^* - n_1 + p_{11}^1 \text{ in case I(a)},$$

$$\sigma_{ij}[A^*] = n_1^* - n_1 + p_{11}^2 \text{ in case I(b)},$$

$$\sigma_{ij}[A^*] = n_1^* + n_1 - p_{11}^1 - \frac{v}{2} \text{ in case II(a)},$$

$$\sigma_{ij}[A^*] = n_1^* + n_1 - p_{11}^2 - \frac{v}{2} \text{ in case II(b)}.$$

The necessary and sufficient condition for G^* to be strongly regular, in addition to the condition for regularity already obtained, is that $\sigma_{ij}[A^*]$ has a fixed value say p_{11}^{1*} in cases I(a) and II(b), and similarly another fixed value say p_{11}^{2*} in cases I(b) and II(a). It follows that the parameters of G must satisfy the condition

$$p_{11}^1+p_{11}^2 = 2n_1-\frac{v}{2}. \tag{8.7}$$

When this condition is satisfied the parameters p_{11}^{1*} and p_{11}^{2*} of G^* are given by

$$p_{11}^{1*} = n_1^*-n_1+p_{11}^1, \quad p_{11}^{2*} = n_1^*-n_1+p_{11}^2.$$

We can thus state the following theorem:

THEOREM (8. 1). *Let G be a strongly regular graph with parameters*

$$v, n_1, p_{11}^1, p_{11}^2.$$

If the vertices of G are divided into two disjoint subsets V_1 and V_2, where $|V_1|=v_1$, $|V_2|=v_2$, then the necessary and sufficient conditions for the graph G^* derived from G by complementation with respect to V_1, V_2, to be strongly regular are

(a) *In G each vertex in V_1 is adjacent to w_1 vertices in V_1 (and therefore n_1-w_1 vertices in V_2); also each vertex in V_2 is adjacent to w_2 vertices in V_2 (and therefore to n_1-w_2 vertices in V_1), where*

$$w_1-w_2 = \frac{v_1-v_2}{2}$$

(b)
$$p_{11}^1+p_{11}^2 = 2n_1-\frac{v}{2}. \tag{b}$$

When these conditions are satisfied the parameters of G^* are given by

$$v^*=v, \quad n_1^* = 2w_1+v_2-n_1 = 2w_2+v_1-n_1, \tag{8.8}$$

$$p_{11}^{1*} = n_1^*-n_1+p_{11}^1, \quad p_{11}^{2*} = n^*-n_1+p_{11}^2. \tag{8.9}$$

If the graph G^* is required to have the same parameters as G, then $n_1^*=n_1$. This automatically ensures that $p_{11}^{1*}=p_{11}^1$ and $p_{11}^{2*}=p_{11}^2$. Also $n_1-w_1 = \frac{v_2}{2}$, $n_1-w_2 = \frac{v_1}{2}$, i.e., in G each vertex of V_1 is adjacent to exactly half the vertices in V_2, and each vertex in V_2 is adjacent to exactly half the vertices in V_1. We therefore have

THEOREM (8. 2). *Let G be a strongly regular graph with parameters*

$$v, n_1, p_{11}^1, p_{11}^2.$$

If the vertices of G are divided into two disjoint subsets V_1 and V_2, then the necessary and sufficient conditions for the graph G^* derived from G by complementation with respect to V_1 and V_2, to be strongly regular with the same parameters as G are

(a) *In G each vertex in V_1 is adjacent to exactly half the vertices in V_2, and each vertex in V_2 is adjacent to exactly half the vertices in V_1.*

(b) $$p_{11}^1+p_{11}^2 = 2n_1-\frac{v}{2}.$$

9. $G_2(d)$ graphs of the pseudo net and negative latin square types. We can define after MESNER [10] a negative Latin square graph $NL_r(k)$ as a strongly regular graph with parameters,

$$v=k^2, \quad n_1 = r(k+1),$$

$$p_{11}^1 = (r+1)(r+2)-(k+2), \quad p_{11}^2 = r(r+1).$$

If $k=2r$, then $p_{11}^1 = p_{11}^2 = r(r+1)$. Hence a negative Latin square graph $NL_r(2r)$ is a $G_2(d)$ graph with parameters $v=4r^2$, $n_1 = r(2r+1)$, $d = r(r+1)$.

(a) We shall now show that a negative Latin square graph $NL_r(2r)$ can be derived from a net graph $L_r(2r)$. A net graph $L_r(2r)$ is the graph of a net of degree r and order $2r$. Take any class C of parallel lines in the net, and divide them into groups of r lines each. Let V_1 be the set of vertices corresponding to the $2r^2$ points on lines of the first group, and V_2 be the set of vertices corresponding to the $2r^2$ points on the lines of the second group. If P is a point on a line l of the first group, then the vertex x corresponding to P is adjacent to the $2r-1$ vertices corresponding to the other points on l. Also through P there pass $r-1$ lines other than l (one belonging to each of the parallel classes other than C). Each of these lines intersects each line of the first group (other than l) in a single point. The vertices corresponding to these $(r-1)^2$ intersections are adjacent to the vertex x. We thus get $w_1=r^2$ vertices in V_1 adjacent to x. It is clear that these are all the vertices in V_1 adjacent to x. Similarly each vertex in V_2 is adjacent to exactly $w_2=r^2$ vertices in V_2. Again for the net graph $L_r(2r)$, $2n_1-\frac{v}{2} = p_{11}^1+p_{11}^2 = 2r(r-1)$. Thus the condition (b) of Theorem (8. 1) is satisfied. Hence by complementation with respect to V_1 and V_2 we obtain a strongly regular graph with parameters

$$v^* = 4r^2, \quad n_1^* = r(2r+1),$$

$$p_{11}^{1*} = r(r+1) = p_{11}^{2*}.$$

This is a negative Latin square graph $NL_r(2r)$ by definition. We thus have

THEOREM (9. 1). *A negative Latin square graph $NL_r(2r)$ exists whenever a net graph $L_r(2r)$ exists. In particular a negative Latin square graph $NL_r(2r)$ exists for $r=3, 6$ and 2^{n-1} where $n\geqq 1$.*

(b) A balanced incomplete block design with parameters $(v_0, b_0, r_0, k_0, \lambda_0)$ is said to be symmetric if $v_0=b_0$ and $r_0=k_0$. We shall then write the parameters as (v_0, k_0, λ_0). The incidence matrix N of the design is defined to be a $v_0\times v_0$ matrix $[n_{ij}]$ where $n_{ij}=1$ or 0 according as the treatment i occurs or does not occur in the block j. The adjacency matrix A of a $G_2(d)$ graph with parameters v, n_1, d is the incidence matrix of a symmetric BIB design with parameters (v, n_1, d). But the incidence matrix N of a symmetric BIB design (v_0, k_0, λ_0) is not always the adjacency matrix of a $G_2(d)$ graph. For this the necessary and sufficient conditions are $n_{ii}=0$,

$n_{ij}=n_{ji}$. Thus N must be symmetric and the main diagonal must consist of zeroes. When these conditions are satisfied we shall say that the incidence matrix is of type I.

Let J denote the $v_0 \times v_0$ matrix all of whose elements are unity. If N is the incidence matrix of a symmetric BIB design (v_0, k_0, λ_0), and $\overline{N} = J-N$ then $\overline{N}$ is the incidence matrix of a symmetric BIB design $(v_0, v_0-k_0, v_0-2k_0+\lambda_0)$. If N is of type I, then $\overline{N}$ is symmetric, and each element of the main diagonal is unity. The incidence matrix of a symmetric BIB will be defined to be of type II when it satisfies these conditions.

Consider a class Ω of symmetric BIB designs (v_0, k_0, λ_0) for which the condition $v_0/4 = k_0-\lambda_0$ is satisfied. Then the following result which we state in the form of a lemma is known [16].

LEMMA (9. 1). *If N_1 and N_2 are the incidence matrices of two symmetric BIB designs (v_1, k_1, λ_1) and (v_2, k_2, λ_2) belonging to the class Ω, then*

$$N = N_1 \times N_2 + \overline{N}_1 \times \overline{N}_2$$

is the incidence matrix of a symmetric BIB design (v_0, k_0, λ_0) belonging to Ω where

$$v_0 = v_1 v_2, \quad k_0 = k_1 k_2 + (v_1 - k_1)(v_2 - k_2), \quad \lambda_0 = k_0 - \frac{v_0}{4},$$

and $\times$ denotes the Kronecker product.

We also note that if N_1 is of type II and N_2 is of type I, then N is of type I.

If the adjacency matrix of a $G_2(d)$ graph with parameters v, n_1, d is the incidence matrix of a BIB design of the class Ω then $v = 4(n_1-d)$. Since $v-1 = n_1(n_1-1)/d$, it follows that $n_1 = \frac{1}{2}[(4d+1) \pm (4d+1)^{\frac{1}{2}}]$, which shows that $n_1 = r(2r+1)$ or $n_1 = r(2r-1)$ for some positive integer r. Thus $G_2(d)$ must be either a pseudo net graph $L_r(2r)$ or a negative Latin square graph $NL_r(2r)$.

THEOREM (9. 2). *The existence of pseudo net graphs $L_{r_1}(2r_1)$ and $L_{r_2}(2r_2)$* implies *the existence of a pseudo net graph $L_r(2r)$ with $r=2r_1r_2$.*

Let N_1 be the adjacency matrix of the pseudo net graph $L_{r_1}(2r_1)$ and N_2 the adjacency matrix of the pseudo net graph $L_{r_2}(2r_2)$. Then N_1 is the incidence matrix of a symmetric BIB $[4r_1^2, r_1(2r_1-1), r_1(r_1-1)]$ and $\overline{N}_2$ is the adjacency matrix of a symmetric BIB $[4r_2^2, r_2(2r_2+1), r_2(r_2+1)]$, where N_1 is of type I and $\overline{N}_2$ is of type II. From Lemma (9. 1)

$$N = \overline{N}_2 \times N_1 + N_2 \times \overline{N}_1$$

is the adjacency matrix of a symmetric BIB (v_0, k_0, λ_0), belonging to Ω, where

$$v_0 = 16r_1^2r_2^2, \quad k_0 = 2r_1r_2(4r_1r_2-1),$$

$$\lambda_0 = (2r_1r_2-1).$$

Since N is of type I, it follows that it is the adjacency matrix of a pseudo net graph $L_r(2r)$, where $r=2r_1r_2$.

THEOREM (9. 3). *A pseudo net graph $L_r(2r)$ exists for all $r = 3^m \cdot 2^{m+n-1}$, where m and n are non-negative integers, $(m, n) \neq (0, 0)$.*

A net graph is also a pseudo net graph. Hence a pseudo net graph $L_r(2r)$ exists for $r=2^{n-1}$, $n\geqq 1$. Hence the theorem is true for $m=0$, $n\geqq 1$. Again a net graph $L_r(2r)$ exists for $r = 3\cdot 2^0$. Assuming the existence of a pseudo net graph $L_r(2r)$ for $r = 3^{m-1}\cdot 2^{m-2}$, the existence of a pseudo net graph $L_r(2r)$ for $r = 3^m\cdot 2^{m-1}$ follows from Theorem (9. 2) by choosing $r_1 = 3^{m-1}\cdot 2^{m-2}$, $r_2=3$. Hence by induction a pseudo net graph $L_r(2r)$ exists for $r = 3^m\cdot 2^{m-1}$. Thus the theorem holds for $m\geqq 1$, $n=0$.

Finally the existence of the pseudo net graph $L_r(2r)$ for $r = 3^m\cdot 2^{m+n-1}$ follows from Theorem (9. 2) by choosing $r_1 = 3^m\cdot 2^{m-1}$ and $r_2=2^{n-1}$, where $m\geqq 1$, $n\geqq 1$. Hence the theorem holds for $m\geqq 1$, $n\geqq 1$.

COROLLARY. $G_2(d)$ graphs with parameters $v=4r^2$, $n_1 = r(2r-1)$, $d = r(r-1)$ exist for all $r = 3^m\cdot 2^{m+n-1}$ where m and n are non-negative integers and $(m, n)\neq$ $\neq(0, 0)$.

THEOREM (9. 4). *The existence of a pseudo net graph $L_{r_1}(2r_1)$ and a negative Latin square graph $NL_{r_2}(2r_2)$ implies the existence of a negative Latin square graph $NL_r(2r)$, where $r=2r_1r_2$.*

Let N_1 be the adjacency matrix of the pseudo net graph $L_{r_1}(2r_1)$ and N_2 the adjacency matrix of a negative Latin square graph $NL_{r_2}(2r_2)$. Then $\bar{N}_2$ is the incidence matrix of a symmetric BIB design $[4r_2^2, r_2(2r_2-1), r_2(r_2-1)]$ and is of type II, where N_1 is as in the proof of Theorem (9. 2). Hence from Lemma (9. 1)

$$N = \bar{N}_2\times N_1 + N_2\times\bar{N}_1$$

is the adjacency matrix of a symmetric BIB (v_0, k_0, λ_0) belonging to the class Ω, where

$$v_0 = 16r_1^2r_2^2, \quad k_0 = 2r_1r_2(4r_1r_2+1), \quad \lambda_0 = (2r_1r_2+1).$$

Therefore N is of type I and is the adjacency matrix of a negative Latin square graph $NL_r(2r)$, where $r=2r_1r_2$.

THEOREM (9. 5). *A negative Latin square graph $NL_r(2r)$ exists for all $r =$ $= 3^m\cdot 2^{m+n-1}$, where m and n are non-negative integers, $(m, n)\neq(0, 0)$.*

The existence of a negative Latin square graph $NL_r(2r)$ for $r=2^{n-1}$, $n\geqq 1$ has already been proved in Theorem (9. 1). Hence the required theorem holds for $m=0$, $n\geqq 1$.

Again a pseudo net graph $L_r(2r)$ exists for $r = 3^{m-1}\cdot 2^{m-2}$, $m\geqq 2$ by Theorem (9. 2), and a negative Latin square graph $NL_r(2r)$ exists for $r=3$ by Theorem (9. 1). Putting $r_1 = 3^{m-1}\cdot 2^{m-2}$, $r_2=3$ in Theorem (9. 4), we get the existence of a negative Latin square graph $NL_r(2r)$ for $r = 3^m\cdot 2^{m-1}$. Hence the theorem holds for $m\geqq 1$, $n=0$.

Finally taking $r_1 = 3^m\cdot 2^{m-1}$, $m\geqq 1$ and $r=2^{n-1}$, $n\geqq 1$ in Theorem (9. 4), we get the existence of a negative Latin square graph $NL_r(2r)$ for $r = 3^m\cdot 2^{m+n-1}$. Hence the theorem holds for $m\geqq 1$, $n\geqq 1$.

COROLLARY. $G_2(d)$ graphs with parameters $v=4r^2$, $n_1 = r(2r+1)$, $d = r(r+1)$ exist for all $r = 3^m\cdot 2^{m+n-1}$ where m and n are non-negative integers and $(m, n)\neq$ $\neq(0, 0)$.

10. Descendant of a strongly regular graph. Let G be a strongly regular graph with parameters

$$v, n_1, p_{11}^1, p_{11}^2$$

and vertex set

$$x_0, x_1, \ldots, x_{n_1}, x_{n_1+1}, \ldots, v_{v-1}$$

where $x_1, x_2, \ldots, x_{n_1}$ are adjacent to x_0, and the remaining $n_2 = v-n_1-1$ vertices $x_{n_1+1}, \ldots, x_{v-1}$ are non-adjacent to x_0. Let G_0 be the subgraph of G obtained by deleting the vertex x_0 and edges incident with x_0. Then the adjacency matrix of G can be written as

$$A = \begin{bmatrix} 0 & \mathbf{1}' & \mathbf{0}' \\ \mathbf{1} & A_{11} & A_{12} \\ \mathbf{0} & A_{21} & A_{22} \end{bmatrix} \tag{10.1}$$

where $\mathbf{1}'$ is a row-vector of order n_1 with all its elements unity, and $\mathbf{0}'$ is a row-vector of order n_2 with all its elements 0.

The adjacency matrix of G_0 is

$$A_0 = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. \tag{10.2}$$

Let V_1 be the set of vertices $x_1, x_2, \ldots, x_{n_1}$ and V_2 the set of vertices $x_{n_1+1}, \ldots, x_n$. Let G_* be obtained from G_0 by complementation with respect to V_1 and V_2, i.e., the vertices of G_* are the same as those of G_0. Two vertices both belonging to V_1 or V_2 are adjacent or non-adjacent in G_* according as they are adjacent or non-adjacent in G_0. Two vertices one of which belongs to V_1 and the other to V_2 are adjacent in G_* if they are non-adjacent in G_0 and are non-adjacent in G_* if they are adjacent in G_0. Given G and the vertex x_0, G_* is completely determined. We may define G_* to be the descendent of G with respect to the vertex x_0. The adjacency matrix of G_* is

$$A_* = \begin{bmatrix} A_{11} & \bar{A}_{12} \\ \bar{A}_{21} & A_{22} \end{bmatrix}. \tag{10.3}$$

We wish to investigate the conditions under which G_* is strongly regular. Since G is strongly regular with parameters $v, n_1, p_{11}^1, p_{11}^2$, the weight of any row of A is n_1 and the scalar product of any two rows is p_{11}^1 or p_{11}^2 according as the rows correspond to a pair of adjacent or non-adjacent vertices. Hence for $1 \leqq i \leqq n_1$, $1 \leqq j \leqq n_2 = v-n_1-1$,

$$w_i(A_{11}) = p_{11}^1, \quad 1 + w_i(A_{11}) + w_i(A_{12}) = n_1.$$

$$w_i(A_{12}) = n_1 - 1 - p_{11}^1.$$

$$w_j(A_{21}) = p_{11}^2, \quad w_j(A_{21}) + w_j(A_{22}) = n_1.$$

$$w_j(A_{22}) = n_1 - p_{11}^2.$$

Again from (7. 1)

$$w_i(\bar{A}_{12}) = n_2 - w_i(A_{12}) = n_2 - n_1 + 1 + p_{11}^1,$$

$$w_j(\bar{A}_{21}) = n_1 - w_j(A_{21}) = n_1 - p_{11}^2.$$

$$w_i(A_*) = w_i(A_{11}) = w_i(\bar{A}_{12})$$

$$= 2p_{11}^1 + n_2 - n_1 + 1.$$

$$w_{n_1+j}(A_*) = w_j(\bar{A}_{21}) + w_j(A_{22})$$

$$= 2n_1 - 2p_{11}^2.$$

Hence necessary and sufficient for G_* to be regular is that

$$2p_{11}^1 + n_2 - n_1 + 1 = 2n_1 - 2p_{11}^2$$

or

$$p_{11}^1 + p_{11}^2 = 2n_1 - \frac{v}{2}. \tag{10. 4}$$

This is the same as the condition (b) of Theorem (8. 1). When this condition is satisfied we can prove as in Theorem (8. 1) that G_* is strongly regular with parameters

$$v_* = v - 1, \quad n_{1*} = 2n_1 - 2p_{11}^2, \tag{10. 5}$$

$$p_{11*}^1 = n_{1*} - n_1 + p_{11}^1, \quad p_{11*}^2 = n_{1*} - n_1 + p_{11}^2.$$

THEOREM (10. 1). *If G is a strongly regular graph with parameters $v, n_1, p_{11}^1, p_{11}^2$, the necessary and sufficient condition for the descendant G_* of G (with respect to any vertex x_0) to be strongly regular is*

$$p_{11}^1 + p_{11}^2 = 2n_1 - \frac{v}{2}.$$

When this condition is satisfied the parameters of G_* are given by (10. 5).

The condition (8. 7) or (10. 4) which appears in Theorems (8. 1), (8. 2) and (10. 1) will be called the *switching condition.*

11. $G_2(d)$ graphs derivable as descendants. Consider a strongly regular graph G with $(v, n_1, p_{11}^1, p_{11}^2)$, for which the switching condition (10.4) is satisfied. Then its descendant G_* has the parameters (10. 5). If G_* is a $G_2(d_*)$ graph, $p_{11*}^1 = p_{11*}^2$. Hence $p_{11}^1 = p_{11}^2$. Thus G itself must be a $G_2(d)$ graph, for which $p_{11}^1 = p_{11}^2 = d$. Substituting in (10. 4) we have $v = 4(n_1 - d)$. It follows as in paragraph 9(b), that G must either be a pseudo-net $L_r(2r)$ graph or a negative Latin square $L_r(2r)$ graph. We are therefore lead to studying descendants of such graphs.

THEOREM (11. 1). *The descendant of a pseudo- net graph $L_r(2r)$ or a negative Latin square graph $L_r(2r)$ is a pseudo linked block graph $LB_r(2r+1)$.*

A pseudo linked block graph $LB_r(2r+1)$ exists for all $r = 3^m \cdot 2^{m+n-1}$ where m and n are non-negative integers $(m, n) \neq (0, 0)$.

Let the parameters of a pseudo- net graph $L_r(2r)$ be $v=4r^2$, $n_1=r(2r-1)$, $p_{11}^1=p_{11}^2=r(r-1)$. The switching condition (10. 4) is satisfied and the parameters of the descendant graph G_* given by (10. 5) are

$$v_*=4r^2-1, \quad n_{1*}=2r^2, \quad p_{11*}^1=p_{11*}^2=r^2. \tag{11. 1}$$

Thus G_* is by definition a pseudo linked block graph $LB_r(2r+1)$.

It can be proved exactly in the same way that the descendant of a negative Latin square graph $L_r(2r)$, has precisely the parameters (11. 1).

COROLLARY. $G_2(d)$ graphs with parameters $v=4r^2-1$, $n_1=2r^2$, $d=r^2$ exist for all $r=3^m\cdot 2^{m+n-1}$ where m and n are non-negative integers $(m, n)\neq(0, 0)$.

REFERENCES

[1] BOSE, R. C.: On the construction of balanced incomplete block designs, *Annals of Eugenics* (London), **9** (1938), 358—399.

[2] BOSE, R. C.: Strongly regular graphs, partial geometries and partially balanced designs, *Pacific J. Math.*, **13** (1963), 389—418.

[3] BOSE, R. C. and CLATWORTHY, W. H.: Some classes of partially balanced designs, *Ann. Math. Statist.*, **26** (1955), 212—232.

[4] BOSE, R. C. and CHAKRAVARTI, I. M.: Hermitian varieties in a finite projective space PG(N, q^2), *Can. J. Math.*, **18** (1966), 1161—1182.

[5] BOSE, R. C., CHAKRAVARTI, I. M. and KNUTH, D. K.: On methods of constructing sets of mutually orthogonal Latin squares using a computer, I., *Technometrics*, **2** (1960), 507—516.

[6] BOSE, R. C. and MESNER, D. M.: On linear associative algebras corresponding to association schemes of partially balanced designs, *Ann. Math. Statist.*, **30** (1959), 21—38.

[7] BOSE, R. C. and SHRIKHANDE, S. S.: On the construction of sets of mutually orthogonal Latin squares and the falsity of a conjecture of Euler, *Trans. Amer. Math. Soc.*, **95** (1960), 191—209.

[8] BRAUER, A.: Limits for the characteristic roots of a matrix IV: Applications to stochastic matrices. *Duke Math. J.*, **9** (1952), 75—91.

[9] ERDŐS, P., RÉNYI, A. and SÓS, V. T.: On a problem of graph theory. *Studia Scientiarum Mathematicarum Hungarica*, **1** (1966), 215—235.

[10] MESNER, D. M.: A new family of partially balanced designs with some Latin square design properties, *Ann. Math. Statist.* **38** (1967), 571—581.

[11] PRIMROSE, E. J. F.: Quadratics in finite geometries, *Proc. Camb. Phil. Soc.* **47** (1951), 299—304.

[12] RAO, C. R.: A study of BIB designs with replications 11 to 15, *Sankhya series A* **23** (1961), 117—127.

[13] RAY CHAUDHURY, D. K.: Some results on quadrics in finite projective geometry, *Can. J. Math.* **14** (1962), 129—138.

[14] SEIDEL, J. J.: Strongly regular graphs of L_2 type and triangular type. *Koninkl Nederl. Akademie Van Wetenschappen-Amsterdam Proceedings, series A* **70**, and *Indag Math.*, **29** (1967), 188—196.

[15] SHRIKHANDE, S. S.: The uniqueness of the L_2 association scheme, *Ann. Math. Statist.* **30** (1959), 781—798.

[16] SHRIKHANDE, S. S.: On a two parameter family of balanced incomplete block designs, *Sankhya series A*, **24** (1962), 33—40.

University of North Carolina, University of Bombay

(Received: May 23, 1969)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 197—210

ÜBER DIE NUMERISCHE LÖSUNG EINES WÄRMELEITUNGSPROBLEMS MIT ANWENDUNG VON HYPERMATRIZEN

von

P. KOSIK

In einer früheren Arbeit befaßten wir uns mit der numerischen Lösung des folgenden Wärmeleitungsproblems. Betrachten wir in der Ebene zwei Wärmeleitungsmedien D_1 und D_2, deren Anordnung auf der Abbildung sichtbar ist. Also bestehe D_1 aus dem Bereich $0 \leqq x \leqq d$; $0 \leqq y \leqq a$ und D_2 aus dem Bereich $0 \leqq x \leqq d$, $-b \leqq y \leqq 0$. Nehmen wir an, daß am Rand $x = d$ der beiden Medien die Temperatur $A \cos \omega t$ oder $A \sin \omega t$ beträgt und daß die übrigen nicht gemeinsamen Grenzen der beiden Bereiche wärmeisoliert sind, also hier keine Wärmemenge verlorengeht. Ferner stimme die Temperatur beider Medien an der gemeinsamen Grenze überein, und die hier aus dem einen Medium ausströmende Wärmemenge sei gleich der in das andere Medium hineinströmenden Wärmemenge. Zur Lösung dieser Aufgabe wendeten wir die Gittermethode an, und zwar derart, daß im Inneneren der Bereiche die Differentialgleichung mit einem Fehler $O(h^2)$ angenähert wurde, und die Randbedingungen mit einem Fehler $O(h)$ erfüllt wurden.

In der gegenwärtigen Arbeit geben wir eine genauere Methode zur numerischen Lösung dieser Aufgabe an, die am Rand der beiden Bereiche die Randbedingungen ebenfalls mit einem Fehler $O(h^2)$ annähert. Wir zeigen, daß die in [1] angegebene Methode auch jetzt im wesentlichen unverändert gebraucht werden kann, nur müssen gewisse Blöcke der Koeffizientenmatrix des mit der Gittermethode erhaltenen algebraischen Gleichungssystems modifiziert werden. Diese Modifikation läßt jedoch jene Eigenschaft der Koeffizientenmatrix unverändert, daß ihre Blöcke bezüglich der Multiplikation vertauschbar sind, was die Anwendbarkeit der angegebenen Methode sichert.

Die mathematische Formulierung des Problems ist die folgende. Die Bereiche D_1 und D_2 seien in einem rechtwinkligen Koordinatensystem derart angeordnet, daß ihre gemeinsame Grenze $0 \leqq x \leqq d$ sei. Bezeichne a_1^2 den Wärmeleitungskoeffizienten des Mediums D_1 und a_2^2 den Wärmeleitungskoeffizienten des Mediums D_2. Die Temperatur von D_1 sei u_1, die von D_2 sei u_2.

Wir suchen jenes Lösungssystem der ebenen Differentialgleichungen

$$\begin{aligned} u_{1_{xx}} + u_{2_{xx}} &= a_1^2 u_{1_t} \\ u_{2_{yy}} + u_{2_{yy}} &= a_2^2 u_{2_t} \end{aligned} \tag{1}$$

das den folgenden Randbedingungen genügt:

$$\left.\begin{aligned} u_{1_x}(x = 0) \\ u_{2_x}(x = 0) \end{aligned}\right\} = 0 \qquad \begin{aligned} &(0 \leqq y \leqq a) \\ &(-b \leqq y \leqq 0) \end{aligned} \tag{2a}$$

$$\left.\begin{matrix} u_{1_y}(y=a) \\ u_{2_y}(y=-b) \end{matrix}\right\} = 0 \quad (0 \leqq x < d) \tag{2b}$$

$$\left.\begin{matrix} u_1(d, y, t) \\ u_2(d, y, t) \end{matrix}\right\} = Ae^{i\omega t} \quad \begin{matrix} (0 \leqq x < d) \\ (-b \leqq y \leqq 0) \end{matrix} \tag{2c}$$

und an der gemeinsamen Grenze

$$u_1(x, 0, t) = u_2(x, 0, t) \tag{2d}$$

$$\lambda_1 u_{1_y}(x, 0, t) = \lambda_2 u_{2_y}(x, 0, t) \qquad (0 < x < d). \tag{2e}$$

In der gegenwärtigen Arbeit geben wir eine Näherungslösung für diese Aufgabe, d.h. Approximationen für die Temperaturfunktionen u_1 und u_2. Die Aufgabe wird — ähnlich zu Arbeit [1] — mit Anwendung von aus vertauschbaren Blöcken bestehenden Hypermatrizen gelöst, jetzt ist jedoch die Annäherung an die Differentialgleichung an der Grenze der beiden Medien besser als zuvor.

Sucht man die Temperaturfunktionen in der Form

$$u_1(x, y, t) = f_1(x, y)e^{i\omega t}$$

und

$$u_2(x, y, t) = f_2(x, y)e^{i\omega t}$$

so kann man die Lösung der Differentialgleichungen dreier Veränderlichen (1) auf die Lösung der Differentialgleichung zweier Veränderlichen

$$\begin{aligned} \Delta f_1(x, y) &= a_1^2 i\omega f_1(x, y) \\ \Delta f_2(x, y) &= a_2^2 i\omega f_2(x, y) \end{aligned} \tag{3}$$

und

zurückführen. Im weiteren befassen wir uns mit der Lösung des letzteren Gleichungssystems, mit den Randbedingungen (2a), (2b), (2d), (2e) und

$$\left.\begin{matrix} f_1(d, y, t) \\ f_2(d, y, t) \end{matrix}\right\} = A \quad \begin{matrix} (0 \leqq y \leqq a) \\ (-b \leqq y \leqq 0). \end{matrix} \tag{4}$$

Teilen wir nun die Vereinigung der Bereiche D_1 und D_2 mit Hilfe von Parallelen zur x-Achse, sowie zur y-Achse in quadratische Teilbereiche auf. Nehmen wir an, dass parallel zur x-Achse im Inneren von D_1 insgesamt m solche Geraden und im Inneren von D_2 insgesamt n solche Geraden liegen, ferner, daß die x-Achse ebenfalls eine Teilgerade ist. Versehen wir nun die im Inneren des Vereinigungsbereiches verlaufenden, zur y-Achse parallelen Teilgeraden — von $y=0$ ausgehend — mit den laufenden Nummern $1, 2, \ldots, r$, die zur x-Achse parallelen Teilgeraden — vom größten y-Wert ausgehend — mit den laufenden Nummer $1, 2, \ldots, \ldots, (m+n)$, wobei die x-Achse ausgelassen wird (siehe Abbildung 1).

Die Näherungswerte von f_1 und f_2 werden nun in den auf obige Weise sich ergebenden Gitterpunkten bestimmt. Bezeichne (i, k) den Schnittpunkt der i-ten vertikalen Geraden mit der h-ten horizontalen Geraden. Benützen wir zur Approximierung des Laplaceschen Operators den Zusammenhang

$$\Delta f_{i,k} \approx \frac{f_{i,k-1} + f_{i-1,k} - 4f_{i,k} + f_{i+1,k} + f_{i,k+1}}{h^2} \tag{5}$$

(siehe KANTOROWITCH [2]). Die Randbedingungen werden folgendermaßen berücksichtigt. Da

$$f_{1_x}(x=0)=0,$$

gebrauchen wir die Form

$$f_{1_x}(x=0)=\frac{f(1,k)-f(0,k)}{h}$$

des Differenzquotienten und nehmen an, daß

$$f(1,k)=f(0,k)$$

gilt.

Ähnlicherweise gehen wir bezüglich aller solcher Randbedingungen vor, wo der Gradient gleich Null ist. In der Nähe der gemeinsamen Grenze wird die Annäherung genauer, als in der Nähe der übrigen Grenzen, wenn hier von der Approximation

$$(6)\qquad f'(x)\approx\frac{1}{h}\left[f(x+h)-f(x)-\frac{h^2}{2!}f''(x)\right]\approx$$

$$\approx\frac{1}{4!}[-f(x+2h)+6f(x+h)-5f(x)]$$

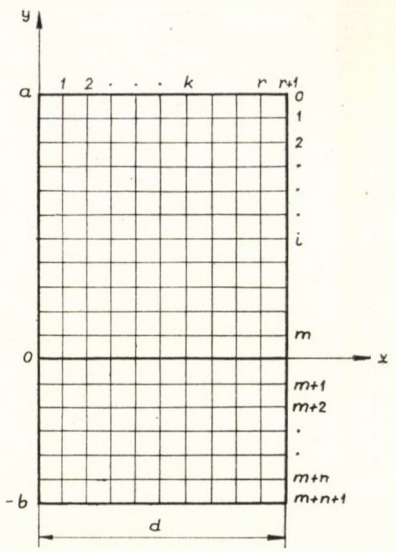


Fig. 1

Gebrauch gemacht wird. Bezeichnen wir nun den Schnittpunkt der i-ten zur y-Asche parallelen Geraden mit der gemeinsamen Grenze der beiden Bereiche (die an der x-Achse liegt) mit x_i. Da

$$f_1(x_i)=f_2(x_i)$$

(7) und

$$\lambda_1 f_{1_y}(x_i)=\lambda_2 f_{2_y}(x_i)$$

gelten, erhält man mit Benützung von (6) und unter Einführung der Funktion

$$(8)\qquad u(i,k)=\begin{cases}f_1(i,k) & (i,k\in D_1)\\ f_2(i,k) & (i,k\in D_2)\end{cases}$$

wobei

$$(9)\qquad u(x_i)\approx\frac{-\lambda_1}{5(\lambda_1+\lambda_2)}u(i,m-1)+\frac{6\lambda_1}{5(\lambda_1+\lambda_2)}u(i,m)+$$

$$+\frac{6\lambda_2}{5(\lambda_1+\lambda_2)}u(i,m+1)-\frac{\lambda_2}{5(\lambda_1+\lambda_2)}u(i,m+2)$$

für die Näherungslösung der Gleichungen (3) im Punkte (i,k) das Gleichungssystem

$$(10)\qquad u(i,k-1)+u(i-1,k)-4u(i,k)+u(i+1,k)+u(i,k+1)=$$

$$=\begin{cases}i\omega a_1^2h^2u(i,k) & (i=1,2,\ldots,r)\\ i\omega a_2^2h^2u(i,k) & (k=1,2,\ldots,m+n).\end{cases}$$

Gleichung (10) erfährt in Punkten, die nahe zum Rand liegen, die folgende Modifikation: Da laut unserer Annahme

$$u(0, 1) - u(1, 1) = u(1, 0) - u(1, 1) = 0$$

ist, so gilt

$$\Delta u(1, 1) \approx -2u(1, 1) + u(1, 2) + u(2, 1). \tag{11}$$

Falls $i > 1$ und $k = 1$ sind, so nimmt Gleichung (10) die folgende Gestalt an:

$$\left.\begin{array}{l} u(i-1, 1) - 3u(i, 1) + u(i+1, 1) + u(i, 2) \\ u(i-1, 1) - 3u(i, 1) + u(i, 2) + A \end{array}\right\} = i\omega a_1^2 h^2 u(i, 1) \quad (i \leqq r). \tag{12}$$

Man erhält ähnliche Gleichungen in der Nähe der Grenze an allen Stellen, wo der Gradient gleich Null ist.

Führt man in den Punkten (i, m), bzw. $(i, m+1)$ die Bezeichnungen

$$\alpha = \frac{\lambda_1}{5(\lambda_1 + \lambda_2)} \quad \text{und} \quad \beta = \frac{\lambda_2}{5(\lambda_1 + \lambda_2)}$$

ein, so nimmt Gleichung (10) die Form

$$\text{a)} \quad (1-\alpha)u(i, m-1) + (6\alpha - 3)u(i, m) + u(i+1, m) + 6\beta u(i, m+1) - \\ - \beta u(i, m+2) = a_1^2 h^2 u(i, m) \qquad (i = 1)$$

$$\text{b)} \quad (1-\alpha)u(i, m-1) + u(i-1, m) + (6\alpha - 4)u(i, m) + u(i+1, m) + \\ + 6\beta u(i, m+1) - \beta u(i, m+2) = i\omega a_1^2 h^2 u(i, m) \qquad (1 < i < r)$$

$$\text{c)} \quad (1-\alpha)u(i, m-1) + u(i-1, m) + (6\alpha - 4)u(i, m) + 6\beta u(i, m+1) - \\ - \beta u(i, m+2) + A = i\omega a_1^2 h^2 u(i, m) \qquad (i = r)$$

$$\text{d)} \quad -\alpha(i, m-1) + 6\alpha u(i, m) + (6\beta - 3)u(i, m+1) + u(i+1, m+1) + \\ + (1-\beta)u(i, m+2) = i\omega a_2^2 h^2 u(i, m+1) \qquad (i = 1)$$

$$\text{e)} \quad -\alpha u(i, m-1) + 6\alpha u(i, m) + u(i-1, m+1) + (6\beta - 4)u(i, m+1) + \\ + u(i+1, m+1) + (1-\beta)u(i, m+2) = i\omega a_2^2 h^2 u(i, m+1) \qquad (1 < i < r)$$

$$\text{f)} \quad -\alpha u(i, m-1) + 6\alpha u(i, m) + u(i-1, m+1) + u(i-1, m+1) + \\ + (6\beta - 4)u(i, m+1) + (1-\beta)u(i, m+2) + A = \\ = i\omega a_2^2 h^2 u(i, m+1) \qquad (i = r) \tag{13}$$

an. Die Gleichungen (13) a), b) und c) beziehen sich auf $u(i, m)$, die Gleichungen (13) d), e) und f) auf $u(i, m+1)$.

Schreibt man nun Gleichung (10) für die Werte $u(i, k)$ auf, so erhält man ein aus $r(m+n)$ Gleichungen bestehendes und ebenso viele Unbekannte enthaltendes inhomogenes lineares Gleichungssystem. So wurde die Lösung der Aufgabe auf die Lösung dieses Gleichungssystems zurückgeführt.

Das Gleichungssystem lautet in Matrixgestalt folgendermaßen

$$\mathbf{\Delta u} + \mathbf{f} = \mathbf{A u}. \tag{14}$$

Die Elemente der Matrix $\mathbf{\Delta}$ sind die Koeffizienten der in der Näherungsform von $\Delta u(i, k)$ auftretenden Funktionswerte. Jedes kr-te Element des Vektors $\mathbf{f}$ ist gleich A $(k = 1, 2, \ldots, m+n)$, alle anderen Elemente sind gleich Null, und $\mathbf{A}$ ist eine Diagonalmatrix, deren ersten $r \cdot m$ Elemente gleich $i\omega A_1^2 h^2$, die übrigen Elemente $i\omega A_2^2 h^2$ sind betragen. Mit der Bezeichnung $\mathbf{Q} = \mathbf{A} - \mathbf{\Delta}$ läßt sich (14) in der Form

$$\mathbf{Q u} = \mathbf{f} \tag{15}$$

schreiben, und die Lösung ist

$$\mathbf{u} = \mathbf{Q}^{-1} \mathbf{f}. \tag{16}$$

Im weiteren wollen wir die Matrix $\mathbf{Q}^{-1}$ bestimmen. $\mathbf{Q}$ ist eine quadratische Matrix von der Ordnung $r(m+n)$. Zerlegen wir $\mathbf{Q}$ in vier Blöcke: $\mathbf{Q} = \begin{bmatrix} \mathbf{P} & \mathbf{R} \\ \mathbf{S} & \mathbf{T} \end{bmatrix}$ derart, dass $\mathbf{P}$ eine quadratische Matrix von der Ordnung $(r \cdot m)$ und $\mathbf{T}$ eine quadratische Matrix von der Ordnung $(r \cdot m)$ sei. Nun wollen wir die Blöcke $\mathbf{P}, \mathbf{R}, \mathbf{S}, \mathbf{T}$ der Matrix $\mathbf{Q}$ angeben. Im nachfolgenden erweist es sich als zweckmäßig, diese Matrizen in Blöcken von der Ordnung r zu zerlegen. Bezeichne $\mathbf{K}$ die folgende quadratische Matrix von der Ordnung r:

$$\mathbf{K} = \begin{bmatrix} 1 & -1 & 0 & 0 & \ldots\ldots\ldots\ldots & \\ -1 & 2 & -1 & 0 & \ldots\ldots\ldots\ldots & \\ 0 & -1 & 2 & -1 & 0 \ldots\ldots\ldots\ldots & \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & & & \\ & & & & -1 \quad 2 \quad -1 & \\ & & & & 0 \quad -1 \quad 2 & \end{bmatrix} \begin{matrix} (1 \\ (2 \\ . \\ . \\ . \\ (r \end{matrix} \tag{17}$$

Mit Hilfe von $\mathbf{K}$ lässt sich die Matrix $\mathbf{P}$ in der Form (18) schreiben.

Die Matrix $\mathbf{R}$ hat m Blockzeilen und n Blockspalten. Die ersten beiden Elemente der m-ten Blockzeile betragen $-6\beta\mathbf{E}$ und $\beta\mathbf{E}$, die anderen Blöcke der Matrix sind gleich Null.

Die Matrix $\mathbf{S}$ besteht aus n Blockzeilen und m Blockspalten. Die beiden letzten Blöcke der letzten Blockzeile sind $\alpha\mathbf{E}$ und $-6\alpha\mathbf{E}$, die übrigen Blöcke sind gleich Null.

Die Matrix $\mathbf{T}$ lässt sich mit Hilfe ihrer Blöcke r-ter Ordnung in der Form (19) schreiben.

Mit $\mathbf{E}$ bezeichnen wir die Einheitsmatrix r-ter Ordnung und mit $\mathbf{O}$ die Nullmatrix r-ter Ordnung.

Ein Vergleich mit der Blöcken (26), (27) und (28) aus Arbeit [1] zeigt, daß die Abweichungen in den entsprechenden Ecken und den dazu benachbarten Blöcken der Matrizen zu finden sind.

$$(19)\qquad \mathbf{T}=\begin{bmatrix} \mathbf{K}+(2+i\omega a_2^2h^2-6\beta)\mathbf{E} & (\beta-1)\mathbf{E} & \mathbf{O} & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -\mathbf{E} & \mathbf{K}+(2+i\omega a_2^2h_2^2)\mathbf{E} & -\mathbf{E} & \mathbf{O} & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & \ddots & \ddots & \ddots & & & & & & & & & & \\ & & -\mathbf{E} & \mathbf{K}+(2+i\omega a_2^2h^2)\mathbf{E} & & -\mathbf{E} & & & & & & & & \\ & & & -\mathbf{E} & & \mathbf{K}+(1+i\omega a_1^2h^2)\mathbf{E} & & & & & & & & \end{bmatrix}\begin{matrix}(1\\(2\\\vdots\\\vdots\\(r\end{matrix}$$

$$(18)\qquad \mathbf{P}=\begin{bmatrix} \mathbf{K}+(1+i\omega a_1^2h^2)\mathbf{E} & -\mathbf{E} & \mathbf{O} & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -\mathbf{E} & \mathbf{K}+(2+i\omega a_2^2)\mathbf{E} & -\mathbf{E} & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & \ddots & \ddots & & & & & & & & & & & \\ & & -\mathbf{E} & \mathbf{K}+(2+i\omega a_1^2h^2)\mathbf{E} & & -\mathbf{E} & & & & & & & & \\ & & \mathbf{O} & -\mathbf{E} & & \mathbf{K}+(2+i\omega a_1^2h^2-6\alpha)\mathbf{E} & & & & & & & & \end{bmatrix}$$

Wird nun die Matrix $\mathbf{Q}^{-1}$ ähnlich wie $\mathbf{Q}$ in die Blöcke $\mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{Z}$ zerlegt, so können diese vier Blöcke mit Hilfe der folgenden Formeln berechnet werden [3],

$$\mathbf{U} = \mathbf{P}^{-1} + \mathbf{P}^{-1}\mathbf{R}(\mathbf{T} - \mathbf{S}\mathbf{P}^{-1}\mathbf{R})^{-1}\mathbf{S}\mathbf{P}^{-1} \tag{20}$$

$$\mathbf{V} = -\mathbf{P}^{-1}\mathbf{R}(\mathbf{T} - \mathbf{S}\mathbf{P}^{-1}\mathbf{R})^{-1}\mathbf{S}\mathbf{P}^{-1} \tag{21}$$

$$\mathbf{W} = -(\mathbf{T} - \mathbf{S}\mathbf{P}^{-1}\mathbf{R})^{-1}\mathbf{S}\mathbf{P}^{-1} \tag{22}$$

$$\mathbf{Z} = (\mathbf{T} - \mathbf{P}\mathbf{S}^{-1}\mathbf{R})^{-1}. \tag{23}$$

Die Lösung der Aufgabe wird erleichtert durch die Erkenntnis, daß die Blöcke r-ter Ordnung der Matrizen $\mathbf{P}, \mathbf{R}, \mathbf{S}, \mathbf{T}$ Polynome der Matrix $\mathbf{K}$, also (bezüglich der Multiplikation) vertauschbar sind. Wir behandeln die Blöcke wie skalare Elemente (Egerváry [4]), die Blöcke r-ter Ordnung der Matrix $\mathbf{Q}^{-1}$ ergeben sich so als Funktionen von $\mathbf{K}$.

Der Übersichtlichkeit halber wollen wir die einzelnen Matrizen als Hyperdiaden aufschreiben. Wir betrachten als Elemente Blöcke r-ter Ordnung. Die Spaltenvektoren wurden mit unteren Indizes, die Reihenvektoren mit oberen Indizes gekennzeichnet.

Führen wir im weiteren die Bezeichnung $\mathbf{T} - \mathbf{S}\mathbf{P}^{-1}\mathbf{R} = \mathbf{D}$ ein. Damit läßt sich (20) in der Gestalt

$$\mathbf{U} = \mathbf{P}^{-1} + \alpha\beta\mathbf{P}_m^{-1}(-6\mathbf{D}_{11}^{-1} + \mathbf{D}_{21}^{-1})(\mathbf{P}^{-1^{m-1}} - 6\mathbf{P}^{-1^{m}}) \tag{25}$$

schreiben. Die Matrix $\mathbf{V}$ ist in Hyperdiadenform

$$\mathbf{V} = -\mathbf{P}_m^{-1}(-6\beta\mathbf{D}^{-1^1} + \beta\mathbf{D}^{-1^2}). \tag{26}$$

und ähnlicherweise ist

$$\mathbf{W} = -\mathbf{D}_1^{-1}(\alpha\mathbf{P}^{-1^{m-1}} - 6\alpha\mathbf{P}^{-1^{m}}). \tag{27}$$

Die Matrix $\mathbf{Z}$ ergibt sich als die Inverse von $\mathbf{D}$, also gilt

$$\mathbf{Z} = \mathbf{D}^{-1}. \tag{28}$$

Nun schreiben wir die Matrizen $\mathbf{P}^{-1}$ und $\mathbf{D}^{-1}$ auf. Mit Einführung der Bezeichnungen

$$\begin{aligned} K + (2 + i\omega a_1^2 h^2)E &= 2\operatorname{ch}\Theta \\ K + (2 + i\omega a_2^2 h^2)E &= 2\operatorname{ch}\Phi \end{aligned} \tag{29}$$

läßt sich $\mathbf{P}$ folgendermaßen schreiben:

$$\mathbf{P} = \begin{bmatrix} 2\operatorname{ch}\Theta - 1 & -1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -1 & 2\operatorname{ch}\Theta & -1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & \cdot & \cdot & \cdot & & & & & & & \\ & & & \cdot & \cdot & & & & & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & -1 & 2\operatorname{ch}\Theta & -1 & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & \alpha - 1 & 2\operatorname{ch}\Theta - 6\alpha \end{bmatrix}. \tag{30}$$

Zur Bestimmung der Kehrmatrizen machen wir hier — ähnlich wie in [1] — von dem folgenden bekannten Zusammenhang Gebrauch. Ist $\mathbf{C}_n$ eine quadratische Matrix n-ter Ordnung, und zwar

$$\mathbf{C}_n = \begin{bmatrix} 2\operatorname{ch}\Theta & -1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ -1 & 2\operatorname{ch}\Theta & -1 & 0 & \dots & \dots & \dots & \dots & \dots \\ & \ddots & \ddots & \ddots & & & & & \\ \dots & \dots & \dots & \dots & \dots & \dots & -1 & 2\operatorname{ch}\Theta & -1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & -1 & 2\operatorname{ch}\Theta \end{bmatrix} \tag{31}$$

so gilt

$$\det \mathbf{C}_n = \frac{\operatorname{sh}(n+1)\Theta}{\operatorname{sh}\Theta}. \tag{32}$$

Mit Benützung dieses Zusammenhanges ergeben sich

$$\det \mathbf{P} = \frac{\operatorname{ch}\left(m+\frac{1}{2}\right)\Theta - 6\alpha \operatorname{ch}\left(m-\frac{1}{2}\right)\Theta + \alpha \operatorname{ch}\left(m-\frac{3}{2}\right)\Theta}{\operatorname{ch}\frac{\Theta}{2}} \tag{33}$$

und

$$(\mathbf{P}^{-1})_{IJ} = \frac{\operatorname{adj}\mathbf{P}_{IJ}}{\det \mathbf{P}} \tag{34}$$

$$(\mathbf{P}^{-1})_{IJ} = \tag{35}$$

$$= \begin{cases} \dfrac{\operatorname{ch}\left(I-\frac{1}{2}\right)\Theta}{\operatorname{ch}\frac{\Theta}{2}\det \mathbf{P}}\left[\operatorname{sh}(m+1-J)\Theta\right] - 6\alpha \operatorname{sh}(m-J)\Theta + \operatorname{sh}(m-1-J)\Theta & (I<J<m) \\ \dfrac{\operatorname{sh}(m+1-I)\Theta - 6\alpha \operatorname{sh}(m-I)\Theta + \alpha \operatorname{sh}(m-1-I)\Theta}{\operatorname{sh}\Theta \operatorname{ch}\frac{\Theta}{2}\det \mathbf{P}} \operatorname{ch}\left(J-\frac{1}{2}\right)\Theta & (I \geqq J) \\ \dfrac{\operatorname{ch}\left(I-\frac{1}{2}\right)\Theta}{\operatorname{ch}\frac{\Theta}{2}\det \mathbf{P}} & (I \leqq J = m). \end{cases}$$

Nun befassen wir uns mit der Bestimmung von $\mathbf{D}^{-1}$. Mit den obigen Bezeichnungen ist

$$\mathbf{D} = \begin{bmatrix} 2\operatorname{ch}\Phi - 6\Psi & (\Psi-1)\mathbf{E} & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -\mathbf{E} & 2\operatorname{ch}\Phi & -\mathbf{E} & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & \ddots & \ddots & \ddots & & & & & & \\ & & & & & & & & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & -\mathbf{E} & 2\operatorname{ch}\Phi & -\mathbf{E} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & -\mathbf{E} & 2\operatorname{ch}\Phi - \mathbf{E} \end{bmatrix} \tag{36}$$

wobei

$$\Psi = \frac{-\alpha\beta}{\operatorname{ch}\left(m+\frac{1}{2}\right)\Theta - 6\alpha\operatorname{ch}\left(m-\frac{1}{2}\right)\Theta + \alpha\operatorname{ch}\left(m-\frac{3}{2}\right)\Theta} \cdot \tag{37}$$

$$\cdot\left[\operatorname{ch}\left(m-\frac{3}{2}\right)\Theta - 6\operatorname{ch}\left(m-\frac{1}{2}\right)\Theta\right].$$

Ebenfalls mit Gebrauch von (31) erhält man für die Kehrmatrix von $\mathbf{D}$

$$\frac{\operatorname{sh} I\Phi - 6\Psi\operatorname{sh}(I-1)\Phi + \Psi\operatorname{sh}(I-2)\Phi}{\operatorname{sh}\Phi\operatorname{ch}\frac{\Phi}{2}\det\mathbf{D}}\operatorname{ch}\left(n-J+\frac{1}{2}\right)\Phi \quad (I<J)$$

$$\frac{\operatorname{ch}\left(n-J+\frac{1}{2}\right)\Phi}{\operatorname{ch}\frac{\Phi}{2}\operatorname{sh}\Phi\det\mathbf{D}}[\operatorname{sh} I\Phi - 6\Psi\operatorname{sh}(I-1)\Phi + \Psi\operatorname{sh}(I-2)\Phi] \quad (I \geqq J > 1) \tag{38}$$

$$\frac{\operatorname{ch}\left(n-I+\frac{1}{2}\right)\Phi}{\operatorname{ch}\frac{\Phi}{2}\det\mathbf{D}} \quad (I \geqq J = 1)$$

wobei (IJ) den in der I-ten Blockzeile und J-ten Blockspalte stehenden Block r-ter Ordnung bezeichnet. Diese Blöcke sind Funktionen von $\mathbf{K}$. Nach Einsetzen erhält

man die Matrix $\mathbf{Q}^{-1}$. In Kenntnis von $\mathbf{Q}^{-1}$ ist

$$u_{Ii} = A \sum_{J=1}^{m+n} (\mathbf{Q}^{-1})_{IJ;ir} \tag{39}$$

wobei u_{Ii} das $r(I-1)+i$-te Element des in (15) vorkommenden Vektors bezeichnet.

Die Summierung der Blöcke $(\mathbf{Q}^{-1})_{IJ}$ und die Einsetzung von $\mathbf{K}$ in diese Summe erfolgt auf folgende Weise. Aus den Formeln (25)—(28) sieht man, daß die Summierung an den Blöcken $(\mathbf{P}^{-1})_{IJ}$ bzw. $(\mathbf{Q}^{-1})_{IJ}$ durchgeführt werden muß. Da

$$\sum_{k=1}^{n} \operatorname{ch}\left(k-\frac{1}{2}\right)\Theta = \frac{\operatorname{sh} n\Theta}{2\operatorname{sh}\frac{\Theta}{2}}, \tag{40}$$

so gilt

$$\sum_{J=1}^{I} (\mathbf{P}^{-1})_{IJ} = \tag{41}$$

$$= \frac{\operatorname{sh}(m+1-I)\Theta - 6\alpha \operatorname{sh}(m-I)\Theta + \alpha \operatorname{sh}(m-1-I)\Theta}{\operatorname{sh}\Theta \operatorname{ch}\frac{\Theta}{2} \det \mathbf{P}} \frac{\operatorname{sh} I\Theta}{2\operatorname{sh}\frac{\Theta}{2}}.$$

Aus dem Zusammenhang

$$\sum_{k=1}^{n} \operatorname{sh} k\Theta = \frac{\operatorname{sh} n\frac{\Theta}{2} \operatorname{sh}(n+1)\frac{\Theta}{2}}{\operatorname{sh}\frac{\Theta}{2}} \tag{42}$$

folgt

$$\sum_{J=I+1}^{m} (\mathbf{P}^{-1})_{IJ} = \frac{\operatorname{ch}\left(I-\frac{1}{2}\right)\Theta}{\operatorname{ch}\frac{\Theta}{2} \operatorname{sh}\Theta \det \mathbf{P}} \cdot \tag{43}$$

$$\cdot \left\{ \left[\operatorname{sh}(m+1-I)\frac{\Theta}{2} \operatorname{sh}(m-I)\frac{\Theta}{2} - 6\alpha \operatorname{sh}(m-I)\frac{\Theta}{2} \cdot \operatorname{sh}(m-1-I)\frac{\Theta}{2} + \right.\right.$$

$$\left.\left. + \alpha \operatorname{sh}(m-1-I)\frac{\Theta}{2} \operatorname{sh}(m-2-I)\frac{\Theta}{2} \right] \left(\operatorname{sh}\frac{\Theta}{2}\right)^{-1} \right\}.$$

Ähnlicherweise kann man die Summierung über die Blöcke erhalten. Also ist

$$\sum_{j=1}^{I} (\mathbf{D}^{-1})_{IJ} = \frac{\operatorname{ch}\left(n-I+\frac{1}{2}\right)\Phi}{\operatorname{ch}\frac{\Phi}{2}\operatorname{sh}\Phi\det\mathbf{D}} \cdot \tag{44}$$

$$\cdot \frac{\operatorname{sh} I\frac{\Phi}{2}\operatorname{sh}(I+1)\frac{\Phi}{2} - 6\Psi\operatorname{sh}(I-1)\frac{\Phi}{2}\operatorname{sh} I\frac{\Phi}{2} + \Psi\operatorname{sh}(I-2)\frac{\Phi}{2}\operatorname{sh}(I-1)\frac{\Phi}{2}}{\operatorname{sh}\frac{\Phi}{2}} .$$

Die Summenformel der Blöcke für Summierung von $(I+1)$ bis n ist die folgende:

$$\sum_{J=I+1}^{n} (\mathbf{D}^{-1})_{IJ} = \frac{\operatorname{sh} I\Phi - 6\Psi\operatorname{sh}(I-1)\Phi + \Psi\operatorname{sh}(I-2)\Phi}{\operatorname{sh}\Phi\operatorname{ch}\frac{\Phi}{2}\det\mathbf{D}} \frac{\operatorname{sh}(n-I)\Phi}{2\operatorname{sh}\frac{\Phi}{2}} . \tag{45}$$

Wie bekannt, gilt für die Funktion $f(\mathbf{A})$ irgendeiner Matrix $\mathbf{A}$

$$f(\mathbf{A}) = \Sigma f(\lambda_k) l_k(\mathbf{A})$$

wobei λ_k den Eigenwert der Matrix $\mathbf{A}$ bezeichnet und $l_k(z)$ das Lagrangesche Interpolationspolynom an der Stelle λ_k ist. Da die Eigenwerte von $\mathbf{K}$ die folgenden sind:

$$\lambda_k = 4\sin^2\left[\frac{2k-1}{2r+1}\cdot\frac{\pi}{2}\right]$$

(siehe [1]), erhält man mit Benützung der Zusammenhänge (40), (41), (42), (43) und unter Einführung der Bezeichnungen

$$\xi(\Theta) = \operatorname{ch}\left(m+\frac{1}{2}\right)\Theta - 6\alpha\operatorname{ch}\left(m-\frac{1}{2}\right)\Theta + \alpha\operatorname{ch}\left(m-\frac{3}{2}\right)\Theta$$

und

$$\eta(\Theta,\Phi) = \operatorname{ch}\left(n+\frac{1}{2}\right)\Phi - 6\Psi(\Theta)\operatorname{ch}\left(n-\frac{1}{2}\right)\Phi + \Psi(\Theta)\operatorname{ch}\left(n-\frac{3}{2}\right)\Phi$$

für die Lösung der Gleichungen (1)

$$u_{1,2}(x_i, y_I, t) = \frac{4}{2r+1} a e^{i\omega t} \sum_{k=1}^{r} (-1)^{k-1} F_{1,2}(\Theta_k, \Phi_k) \cdot$$

$$\cdot \cos\left[(2I-1)\frac{2k-1}{2r+1}\frac{\pi}{2}\right] \sin\left[\frac{2k-1}{2r+1}\pi\right], \tag{46}$$

wobei

$$F_1(\Theta_k, \Phi_k) = \frac{1}{2\xi(\Theta)\,\mathrm{sh}\,\Theta\,\mathrm{sh}\,\dfrac{\Theta}{2}} \left\{ \mathrm{sh}\,(I-1)\,\Theta\,[\mathrm{sh}\,(m+1-I)\,\Theta - 6\alpha\,\mathrm{sh}\,(m-I)\,\Theta + \alpha\,\mathrm{sh}\,(m-1-I)\,\Theta] + 2\,\mathrm{ch}\left(I-\frac{1}{2}\right)\Theta \cdot \right.$$

$$\cdot \left[\mathrm{sh}\,(m+1-I)\,\frac{\Theta}{2}\,\mathrm{sh}\,(m+2-I)\,\frac{\Theta}{2} - 6\alpha\,\mathrm{sh}\,(m-I)\,\frac{\Theta}{2}\,\mathrm{sh}\,(m+1-I)\,\frac{\Theta}{2} + \alpha\,\mathrm{sh}\,(m-1-I)\,\frac{\Theta}{2}\,\mathrm{sh}\,(m-I)\,\frac{\Theta}{2} \right] \Big\} +$$

$$+ \alpha\beta\,\mathrm{ch}\left(I-\frac{1}{2}\right)\Theta \left[-6\,\mathrm{ch}\left(n-\frac{1}{2}\right)\Phi + \mathrm{ch}\left(n-\frac{3}{2}\right)\Phi \right] \left\{ \mathrm{sh}\,(m-2)\,\Theta\,[\mathrm{sh}\,2\Theta - 6\alpha\,\mathrm{sh}\,\Theta] + \right.$$

$$+ 2\,\mathrm{ch}\left(m-\frac{3}{2}\right)\Theta \left[\mathrm{sh}\,\Theta\,\mathrm{sh}\,\frac{3}{2}\,\Theta - 6\alpha\,\mathrm{sh}\,\frac{\Theta}{2}\,\mathrm{sh}\,\Theta \right] - 6\left[(1-\alpha)\,\mathrm{sh}\,(m-1)\,\Theta\,\mathrm{sh}\,\Theta + 2\,\mathrm{ch}\left(m-\frac{1}{2}\right)\Theta\,\mathrm{sh}\,\frac{\Theta}{2}\,\mathrm{sh}\,\Theta \right] \Big\} \cdot$$

$$(47) \quad \cdot \left(2\xi(\Theta)\,\mathrm{sh}\,\Theta\,\mathrm{sh}\,\frac{\Theta}{2} \left\{ \xi(\Theta)\,\mathrm{ch}\left(n-\frac{1}{2}\right)\Phi + \alpha\beta \left[\mathrm{ch}\left(m-\frac{3}{2}\right)\Theta - 6\,\mathrm{ch}\left(m-\frac{1}{2}\right)\Theta \right] \left[6\,\mathrm{ch}\left(n-\frac{1}{2}\right)\Phi - \mathrm{ch}\left(n-\frac{3}{2}\right)\Phi \right] \right\} \right)^{-1} -$$

$$- \beta\,\mathrm{ch}\left(I-\frac{1}{2}\right)\Theta \left\{ -12\,\xi(\Theta)\,\mathrm{ch}\left(n-\frac{1}{2}\right)\Phi\,\mathrm{sh}\,\frac{\Phi}{2}\,\mathrm{sh}\,\Phi - 6\xi)(\Theta)\,\mathrm{sh}\,(n-1)\,\Phi\,\mathrm{sh}\,\Phi - 6\alpha\beta \left[\mathrm{ch}\left(m-\frac{3}{2}\right)\Theta + 36\,\mathrm{ch}\left(m-\frac{1}{2}\right)\Theta \right] \cdot \right.$$

$$\cdot \mathrm{sh}\,(n-1)\,\Phi\,\mathrm{sh}\,\Phi + 2\xi(\Theta)\,\mathrm{ch}\left(n-\frac{3}{2}\right)\Phi\,\mathrm{sh}\,\Phi\,\mathrm{sh}\,\frac{3}{2}\,\Phi + \xi(\Theta)\,\mathrm{sh}\,(n-2)\,\Phi\,\mathrm{sh}\,2\Phi +$$

$$+ \alpha\beta \left[\mathrm{ch}\left(m-\frac{3}{2}\right)\Theta - 6\,\mathrm{ch}\left(m-\frac{1}{2}\right)\Theta \right] \left[12\,\mathrm{ch}\left(n-\frac{3}{2}\right)\Phi\,\mathrm{sh}\,\frac{\Phi}{2}\,\mathrm{sh}\,\Phi + 6\,\mathrm{sh}\,(n-2)\,\Phi\,\mathrm{sh}\,\Phi \right] \Big\} \cdot$$

$$\cdot \left(2\xi(\Theta)\,\mathrm{sh}\,\Phi\,\mathrm{sh}\,\frac{\Phi}{2} \left\{ \xi(\Theta)\,\mathrm{ch}\left(n+\frac{1}{2}\right)\Phi + \alpha\beta \left[\mathrm{ch}\left(m-\frac{3}{2}\right)\Theta - 6\,\mathrm{ch}\left(m-\frac{1}{2}\right)\Theta \right] \left[6\,\mathrm{ch}\left(n-\frac{1}{2}\right)\Phi - \mathrm{ch}\left(n-\frac{3}{2}\right)\Phi \right] \right\} \right)^{-1}.$$

$$F_2(\Theta_k, \Phi_k) = -\alpha \operatorname{ch}\left(n-I+\frac{1}{2}\right)\Phi\left\{\operatorname{sh}(m-2)\Theta[\operatorname{sh}2\Theta-6\alpha\operatorname{sh}\Theta]+2\operatorname{ch}\left(m-\frac{3}{2}\right)\Theta\left[\operatorname{sh}\Theta\operatorname{sh}\frac{3}{2}\Theta-6\alpha\operatorname{sh}\frac{\Theta}{2}\operatorname{sh}\Theta\right]-\right.$$

$$\left.-6[(1-\alpha)\operatorname{sh}(m-1)\Theta\operatorname{sh}\Theta+2\operatorname{ch}\left(m-\frac{1}{2}\right)\Theta\operatorname{sh}\frac{\Theta}{2}\operatorname{sh}\Theta]\right\}\cdot$$

$$\cdot\left(2\operatorname{ch}\Phi\operatorname{sh}\frac{\Phi}{2}\left\{\xi(\Theta)\operatorname{ch}\left(n+\frac{1}{2}\right)\Phi+\alpha\beta\left[\operatorname{ch}\left(m-\frac{3}{2}\right)\Theta-6\operatorname{ch}\left(m-\frac{1}{2}\right)\Theta\right]\left[6\operatorname{ch}\left(n-\frac{1}{2}\right)\Phi-\operatorname{ch}\left(n-\frac{3}{2}\right)\Phi\right]\right\}\right)^{-1}+$$

$$+2\xi(\Theta)\operatorname{ch}\left(n-I+\frac{1}{2}\right)\Phi\operatorname{sh}I\frac{\Phi}{2}\operatorname{sh}(I+1)\frac{\Phi}{2}+\xi(\Theta)\operatorname{sh}(n-I)\Phi\operatorname{sh}I\Phi+\alpha\beta\left[\operatorname{ch}\left(m-\frac{3}{2}\right)\Theta-6\operatorname{ch}\left(m-\frac{1}{2}\right)\Theta\right]\cdot$$

$$\cdot\left[12\operatorname{sh}(I-1)\frac{\Phi}{2}\operatorname{sh}I\frac{\Phi}{2}\operatorname{ch}\left(n-I+\frac{1}{2}\right)\Phi-2\operatorname{sh}(I-2)\frac{\Phi}{2}\operatorname{sh}(I-1)\frac{\Phi}{2}\operatorname{ch}\left(n-I+\frac{1}{2}\right)\Phi+\right.$$

$$\left.+6\operatorname{sh}(n-I)\Phi\operatorname{sh}(I-1)\Phi-\operatorname{sh}(n-I)\Phi\operatorname{sh}(I-2)\Phi\right]\cdot$$

$$\cdot\left(2\operatorname{sh}\Theta\operatorname{sh}\frac{\Theta}{2}\xi(\Theta)\operatorname{ch}\left(n+\frac{1}{2}\right)\Phi+\alpha\beta\left[\operatorname{ch}\left(m-\frac{3}{2}\right)\Theta-6\operatorname{ch}\left(m-\frac{1}{2}\right)\Theta\right]\left[6\operatorname{ch}\left(n-\frac{1}{2}\right)\Phi-\operatorname{ch}\left(n-\frac{3}{2}\right)\Phi\right]\right\}\right)^{-1}. \tag{48}$$

gelten.

Die Werte Θ_k und Φ_k ergeben sich aus den Zusammenhängen

$$\lambda_k + 2 + i\omega a_1^2 h^2 = 2 \operatorname{ch} \Theta_k$$

bzw.

$$\lambda_k + 2 + i\omega a_2^2 h = 2 \operatorname{ch} \Phi_k .$$

Die in der Arbeit angegebene Lösungsmethode eignet sich auch für Programmierung an Rechenautomaten.

LITERATURVERZEICHNIS

[1] KOSIK, P.: Über die Näherunglösung eines Wärmeleitungsproblems durch Anwendung der Theorie der Hypermatrizen, *MTA Mat. Kut. Int. Közl.* **9** A. 3. (1964).
[2] KANTOROWITCH, L. W.—KRYLOW, W. I.: *Näherungsmethoden der höheren Analysis*, Deutscher Verlag d. Wissenschaften, Berlin 1956.
[3] Булгаков, Б. В.: Колебания. Гос. изд. техн.-теор. лит. Москва, 1954.
[4] EGERVÁRY, J.: Mátrix-függvények kanonikus előállításáról és annak néhány alkalmazásáról, *MTA III. (Mat. és Fiz.) Oszt. Közl.* **3** (1953) 417—458.

Mathematisches Institut der Ungarischen Akademie der Wissenschaften, Budapest

(Eingegangen: 29. Juni 1969.)

PÁL RÉVÉSZ: THE LAWS OF LARGE NUMBERS

AKADÉMIAI KIADÓ, BUDAPEST, 1967. 176 p.

This book is the first one, completely elaborating all known laws of this field.

In the Introduction a definition of the branch of the title is given, by this definition any law of large numbers states the convergence (in some sense) of a sequence

$$\zeta_n = \frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n}$$

where $\xi_1, \xi_2, \ldots$ are random variables, or states something about the rate of the convergence of ζ_n. The theorems are classified by the properties of the sequence $\xi_1, \xi_2, \ldots$ and by the kind of the convergence.

In Chapter 0 the preliminary material is collected: the most important definitions and theorems applied in the book.

Chapter 1 contains the basic concepts, the definitions and some general properties of the weak, strong and mean law of large numbers.

Chapter 2 represents the amplest and most important chapter of the laws of large numbers, the laws for independent random variables. It also treats the problem of the rate of convergence, the law of iterated logarithm and the weighted averages of independent variables.

Chapters 3, 4, 6, 7 discuss the laws for other classes of (discrete) stochastic processes as orthogonal, stationary and symmetrically dependent sequences, Markov chains etc.

Chapter 5 deals with subsequences of sequences of random variables and Chapter 8 with weakly dependent variables as centered or mixing ones.

Chapters 9 and 10 treat some special questions.

Chapter 11 gives examples for the application of this area (in number theory, in statistics and in information theory).

The book gives a survey of the results and the most important methods of proving in this field. Therefore, the author enters into the details of the proofs mostly in cases the methods of the proofs are general ones or the results are due to the author and have not been published before.

The deeper study of the details and of related topics is facilitated by an up-to-date bibliography (with 123 head words.)

The book completes the knowledge of the specialist of this field and at the same time, it is suitable as an introduction to the subject for mathematicians working on other fields.

The book has a German edition too (Pál Révész: Die Gesetzte der großen Zahlen, Akadémiai Kiadó, Budapest 1968.)

J. Komlós

THEORY OF GRAPHS

PROCEEDING OF THE COLLOQUIUM HELD AT TIHANY, HUNGARY
Edited by P. ERDŐS and G. KATONA
AKADÉMIAI KIADÓ, BUDAPEST, 1968, 370 p.

The János Bolyai Mathematical Society organized a colloquium on the Theory of Graphs at Tihany (Lake Balaton, Hungary) 5—9 September 1966. The volume contains the papers and problems presented at this colloquium.

Graph theory has undergone a rapid development during the past two decades, and is nowadays widely applied in the theory of electrical networks, operations research, sociology, etc. The first international meeting on graph theory was held on Dobogókő (Hungary) in 1959. The first one was followed by many others: in 1959 at Halle (GDR), 1963 at Princeton and at Smolenice (Czehoslovakia), in 1966 at Waterloo (Canada) and in Rome.

The volume contains 35 papers of the following authors: W. G. BROWN, R. K. GUY, N. S. MENDELSOHN (Canada); J. BOSÁK, A. KOTZIG, A. ROSA, B. ZELINKA, Š. ZNÁM (Czehoslovakia); H. J. FINCK, M. HASSE, H. J. VOSS, H. WALTHER, W. WESSEL (German Federal Republic); E. C. MILNER, C. St. J. A. NASH-WILLIAMS, J. SHEEHAN (Great Britain); J. Ch. BOLAND (Holland); M. BÁNKFALVI, ZS. BÁNKFALVI, B. BOLLOBÁS, J. DÉNES, P. ERDŐS, T. GALLAI, A. HAJNAL, G. KATONA, G. KORVIN, L. LOVÁSZ, J. PELIKÁN, R. PÉTER, M. SIMONOVITS (Hungary); L. W. BEINEKE, G. CHARTRAND, F. HARARY, D. KLEITMAN (USA); HOANG TUY (Democratic Republic of Viet-nam); G. RINGEL (West-Berlin). The above list of authors ensures the high level of the volume.

The papers deal among others with the following topics: Euler and Hamiltonian lines of special finite and infinite graphs, properties of the graphs with prescribed valencies, problems connected with the diameter of a graph, extremal problems, limit value problems, embedding of graphs into different surfaces, colouring problems, enumeration problems, generalized graphs as subsets of finite sets and algebraic methods in graph theory. We must emphasize the great number of unsolved problems presented at a special half-day session of the colloquium, which are contained in the volume. We are sure that these interesting problems will have a stimulating effect on research in graph theory.

The volume is the result of careful work of the editors. This valuable volume has been published by the Publishing House of the Hungarian Academy of Sciences in a worthy form.

B. Andrásfai

PROCEEDINGS OF THE COLLOQUIUM ON INFORMATION THEORY

Edited by A. RÉNYI

PUBLISHED BY THE J. BOLYAI MATHEMATICAL SOCIETY BUDAPEST, HUNGARY
VOL. I—II. (520 PAGES)

(Distributor: Kultúra Book Export Dept. Budapest 62 *P. O. B.* 149, *Hungary, Price* $ 18.*)*

Because of the rapid development of information theory and its applications in many different fields (communication engineering, biology, linguistics etc.), in the last years several meetings have been organized on this very important subject.

The Colloquium on Information Theory, organized by the J. Bolyai Mathematical Society (sponsored by the Federation of Technical and Scientific Societies of Hungary and the Hungarian National Committee of the URSI) and held at the University L. Kossuth in Debrecen (Hungary) from 19 to 24 September 1967, was very successfull as it can be seen from the present volume.

The two volumed Proceedings contains the text of 34 lectures presented at the Colloquium, making these papers available also to those who were not present at the meeting.

The names of the authors arranged according to countries, are as follows: L. L. CAMPBELL (Canada), A. PEREZ, I. VAJDA (Czechoslovakia), E. A. GRAHAM, C. PICARD, J. L. RIGAL, N. AGGARWAL, C. CANONGE (France), P. ZIESCHE (German Democratic Republic), R. AHLSWEDE (German Federal Republic), A. ÁDÁM, S. BENDE, S. CSIBI, I. CSISZÁR, Z. DARÓCZY, J. DÉNES. J. FRITZ, O. GULYÁS, B. GYIRES, P. JUHÁSZ-NAGY, G. KATONA, J. KOMLÓS, T. NEMETZ, A. RÉNYI, G. SCHERMANN, G. TUSNÁDY, I. VINCZE (Hungary), B. FORTE (Italy), S. GUIAŞU (Rumania), F. J. BEUTLER, R. C. BOSE, J. G. CALDWELL, I. T. FRISCH, W. D. GREGG, W. ROTHFARB, G. J. SIMMONS, M. SOBEL, S. WATANABE (USA), I. A. OVSEEVICH, M. S. PINSKER (USSR).

The subjects of the papers included are as follows: Foundations on information theory, coding theorems, algebraic methods in the theory of effective code-construction (particularly of error-correcting codes), i.e. coding theorems and design techniques, search theory, statistical communication and control theory, application of information theory in probability and statistics, in biology, physics, linguistics and also in other branches of mathematics (e.g. in ergodic theory and approximation theory).

K. Bognár

INDEX

Printed in Hungary

A kiadásért felel az Akadémiai Kiadó igazgatója — Műszaki szerkesztő: Farkas Sándor
A kézirat nyomdába érkezett: 1970. I. 29. — Terjedelem: 18,75 (A/5) ív, 8 ábra

70-532 — 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. Jährlich erscheint ein Band.

Adresse der Redaktion: Budapest V., Reáltanoda u. 13—15, Ungarn.
Technischer Redaktor: Gy. Petruska

Abonnementspreis pro Band (pro Jahr): $ 16.00. Bestellbar bei dem 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 oiginaux, en francais, anglais, allemand ou russe, du domaine des mathématiques.

Rédaction: Budapest V., Reáltanoda u. 13—15, Hongrie.
Rédacteur téchnique: Gy. Petruska

Le prix de l'abonnement: $ 16.00 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 Bibliotheque 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 и немецком, английском, французском и русском языках. Отдельные выпуски составляют ежегодно один том.

Адрес редакции: Budapest V., Reáltanoda u. 13—15, Венгрия.
Технический редактор: Gy. Petruska.

Подпусная цена на год (за один том): $ 16.00. 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.

All the reviews of the Hungarian Academy of Sciences may be obtained among others from the folloving 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 Librairie
48, rue Gay Lussac
Paris 5

GERMAN DEMOCRATIC REPUBLIC

Deutscher Buchexport 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-ehome
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 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

IX. 30. 1970. **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 V.
FASC. 3—4.
1970

AKADÉMIAI KIADÓ, BUDAPEST

Studia Scientiarum Mathematicarum Hungarica

A Magyar Tudományos Akadémia matematikai folyóirata

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

Technikai szerkesztő: Petruska 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. Petruska

Subscription rate: $ 16.00 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 MATHEMATICA

TOMUS V

INDEX

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 215—220.

AN ANALOGUE OF A PROBLEM OF J. BALÁZS

by

A. K. VARMA and S. K. GUPTA

Recently Prof. J. Balázs [3] has considered an interesting problem on weighted (0, 2) interpolation. By weighted (0, 2) interpolation he seeks to find the polynomial $f(x)$ when the values of $f(x)$ and $[\varrho(x)f(x)]''$ are prescribed at the given abscissas, $\varrho(x)$ being given weight function. The main results of his paper are as follows.

THEOREM 1. 1. (J. Balázs): *Let $x_k's$ be the zeros of ultraspherical polynomial $P_n^{(\lambda)}(x)$. For n even and y_{k0}, y_{k2} are arbitrary prescribed numbers in advance then there exists a unique polynomial $R_n(x)$ of degree $\leqq 2n$ such that*

$$R_n(x_k) = y_{k0} \qquad k = 1, 2, \ldots, n \tag{1. 1}$$

$$\left[(1-x^2)^{\frac{1+\lambda}{2}} R_n(x)\right]''_{x=x_k} = y_{k2} \qquad k = 1, 2, \ldots, n \tag{1. 2}$$

$$R_n(0) = \sum_{k=1}^{n} y_{k0} l_k^2(0) \tag{1. 3}$$

For n odd there need not exist a unique polynomial $R_n(x)$ of degree $\leqq 2n$ satisfying the above conditions. If the last condition (1. 3) *is dropped, there does not exist a unique polynomial $R_n(x)$ of degree $\leqq 2n-1$ satisfying the above conditions (for both n even and odd).*

THEOREM 1. 2. (J. Balázs): *Let $f(x)$ be a continuous function in $[-1, +1]$ and let $f'(x) \in \operatorname{Lip} \mu$, $\mu > \frac{1}{2}$. Further*

$$f(x_{kn}) = y_{k0}, \quad y_{k2} = o\left(\sqrt{n}\right)(1-x_{kn}^2)^{\frac{\alpha-3}{2}}, \qquad k = 1, 2, \ldots, n \tag{1. 4}$$

then the sequence of polynomials $\{R_n(x, f)\}$ converges uniformly to $f(x)$ in $-1+\varepsilon \leqq x \leqq 1-\varepsilon$, $0<\varepsilon<1$ (ε being arbitrary fixed positive number).

The object of this note is to modify the problem of Balázs in such a manner so that the corresponding sequence of interpolatory polynomials may converge uniformly to $f(x)$ in $-1+\varepsilon \leqq x \leqq 1-\varepsilon$ for a wider class of functions and yet the degree of polynomials be $\leqq 2n-1$. We limit ourselves here only to Tchebycheff abscissas of the second kind.

Let us denote the zeros of $u_n(x) = \dfrac{\sin(n+1)\theta}{\sin\theta}$, $\cos\theta = x$ by

$$-1 < x_n < x_{n-1} < \cdots < x_2 < x_1 < +1 \tag{1. 5}$$

THEOREM 1. 3. *If n is even, then to prescribed values y_{k0}, y_{k2} $(k=1, 2, \ldots, n)$, there is a uniquely determined polynomial $R_n(x)$ of degree $\leqq 2n-1$ such that*

$$R_n(x_k) = y_{k0}, \quad [(1-x^2)^{\alpha} R_n(x)]''_{x=x_k} = y_{k2} \tag{1. 6}$$

for $k = 1, 2, \ldots, n$ *and* $\alpha \neq \frac{3}{4}$, $\alpha \neq \frac{9}{4}$ $\alpha > 0$.

If n is odd there is in general no unique polynomial $R_n(x)$ of degree $\leqq 2n-1$ which satisfies (1. 6). *For* $\alpha = \frac{9}{4}$ *or* $\alpha = \frac{3}{4}$. *the interpolatory polynomial $R_n(x)$ does not exist uniquely either for n even or for n odd.*

Obviously $R_n(x, f)$ has the following representation corresponding to a given continuous function $f(x)$ in $[-1, +1]$

$$R_n(x, f) = \sum_{k=1}^{n} f(x_{kn})\, r_{kn}(x) + \sum_{k=1}^{n} y_{k2}\, \varrho_{kn}(x) \tag{1. 7}$$

where $r_{kn}(x)$ and $\varrho_{kn}(x)$ are stated in (2. 3) and (2. 1). Now we prove

THEOREM 1. 4. *Let $f(x)$ be a continuous function in the closed interval* $[-1, +1]$ *and let it satisfy the Zygmund condition*

$$|f(x+h) - 2f(x) + f(x-h)| = o(h) \tag{1. 8}$$

in $(-1, +1)$. For $\alpha = \frac{7}{4}$, the sequence $\{R_n(x, f)\}$ *converges uniformly to $f(x)$ in every closed interval* $-1+\varepsilon \leqq x \leqq 1-\varepsilon$, ε *being fixed* $(0<\varepsilon<1)$, *provided*

$$|y_{k2}| = \frac{o(n)}{(1-x_{kn}^2)^{3/4}}, \qquad k = 1, 2, \ldots, n. \tag{1. 9}$$

2. Here we shall give the explicit representation of the fundamental polynomials for $\alpha > \frac{3}{4}$ and $\alpha \neq \frac{9}{4}$. Let us denote $\frac{7}{4} - \alpha = p$, then for n even we have

$$\varrho_{kn}(x) = \frac{(1-x^2)^{p-1} u_n(x)}{2(1-x_{kn}^2)^{\alpha-1} u_n'(x_{kn})} \left[A_k \int_{-1}^{x} \frac{u_n(t)}{(1-t^2)^p}\, dt + \int_{-1}^{x} \frac{l_{kn}(t)}{(1-t^2)^p}\, dt \right], \tag{2. 1}$$

where

$$A_k \int_{-1}^{+1} \frac{u_n(t)}{(1-t^2)^p}\, dt + \int_{-1}^{+1} \frac{l_{kn}(t)}{(1-t^2)^p}\, dt = 0, \tag{2. 2}$$

$$l_{kn}(t) = \frac{u_n(t)}{(t-x_k)\, u_n'(x_{kn})},$$

$$r_{kn}(x) = l_{kn}^2(x) + \frac{u_n(x)\, l_{kn}'(x)}{2u_n'(x_{kn})} + c_k \varrho_{kn}(x) + u_n(x)\, q_{n-1}(x), \tag{2. 3}$$

where

$$(2.4)\quad (1-x^2)^{1-p}q_{n-1}(x) = -\frac{\left(p+\frac{1}{2}\right)}{u_n'(x_{kn})}\left[B_k\int_{-1}^{x}\frac{u_n(t)\,dt}{(1-t^2)^p}+\int_{-1}^{x}\frac{tl_{kn}'(t)}{(1-t^2)^p}\,dt\right],$$

$$(2.5)\quad c_k = \big(n(n+2)+2\alpha-3\big)(1-x_{kn}^2)^{\alpha-1}-\frac{(2\alpha-3)(2\alpha-5)\,x_{kn}^2}{(1-x_{kn}^2)^{2-\alpha}},$$

$$(2.6)\quad B_k\int_{-1}^{+1}\frac{u_n(t)}{(1-t^2)^p}\,dt+\int_{-1}^{+1}\frac{tl_{kn}'(t)}{(1-t^2)^p}\,dt = 0.$$

3. In order to prove the convergence theorem stated above we need the following lemmas.

LEMMA 3. 1. *For all* x *in* $-1 \leqq x \leqq +1$, *we have*

$$(3.1)\quad \left|\int_{-1}^{x}u_n(t)\,dt\right| \leqq \frac{2}{n+1},$$

$$(3.2)\quad \left|\int_{-1}^{x}l_{kn}(t)\,dt\right| \leqq \frac{24(1-x_{kn}^2)^{1/2}}{(n+1)}, \qquad k = 1, 2, \ldots, n,$$

$$(3.3)\quad \left|\int_{-1}^{x}tl_{kn}'(t)\,dt\right| \leqq 13, \quad n \geqq 4, \qquad k = 1, 2, \ldots, n.$$

PROOF. Proof of (3. 1) is clear. From a result of L. FEJÉR we have

$$(3.4)\quad l_{kn}(t) = \frac{2(1-x_{kn}^2)}{n+1}\sum_{r=0}^{n-1}u_r(x_{kn})u_r(t).$$

Since

$$(3.5)\quad \int_{-1}^{x}u_r(t)\,dt = \frac{\cos(r+1)\theta+(-1)^r}{r+1}, \qquad x = \cos\theta$$

and

$$(3.6)\quad \left|\sum_{j=1}^{n}\frac{\sin j\theta}{j}\right| \leqq 2\pi^{1/2}.$$

Now using (3. 4) → (3. 6) gives us (3. 2). For (3. 3) we observe that

$$(3.7)\quad \int_{-1}^{x}tl_{kn}'(t)\,dt = xl_{kn}(x)+l_{kn}(-1)-\int_{-1}^{x}l_{kn}(t)\,dt$$

Now we use

$$(3.8)\quad |l_{kn}(x)| \leqq 4 \quad \text{for} \quad -1 \leqq x \leqq +1$$

and (3. 2) we get (3. 3).

LEMMA 3. 2. *For n even and all x such that*

$$-1+\varepsilon \leqq x \leqq 1-\varepsilon$$

we have

$$|\varrho_{kn}(x)| \leqq \frac{24(1-x_{kn}^2)^{3/4}}{\varepsilon^{3/2} n^2}, \qquad k = 1, 2 \ldots, n \tag{3. 9}$$

$$\sum_{k=1}^{n} (1-x_{kn}^2)^{-3/4} |\varrho_{kn}(x)| \leqq \frac{24}{\varepsilon^{3/2} n}, \qquad n = 4, 6, \ldots \tag{3. 10}$$

PROOF. Since

$$\int_{-1}^{+1} u_n(t)\,dt = \frac{2}{n+1} \qquad n \text{ even.} \tag{3. 11}$$

We get from (3. 2) and (2. 2)

$$|A_k| \leqq 12\,(1-x_{kn}^2)^{1/2}. \tag{3. 12}$$

Now it is clear that (3. 9) follows from (2. 1), (3. 12), (3. 1) and (3. 2).

LEMMA 3. 3. *For n even, and* $-1+\varepsilon \leqq x \leqq 1-\varepsilon$ *we have*

$$|r_{kn}(x) \leqq \frac{1}{2} l_{kn}^2(x) + \frac{44}{\varepsilon^{3/2}}, \qquad n = 4, 6, \ldots \tag{3. 13}$$

$$\sum_{k=1}^{n} |r_{kn}(x)| \leqq \frac{50n}{\varepsilon^{3/2}} \tag{3. 14}$$

PROOF. By using the relation

$$u_n(x)\, l'_{kn}(x) = u'_n(x)\, l_{kn}(x) - u'_n(x_{kn})\, l_{kn}^2(x), \tag{3. 15}$$

we have

$$r_{kn}(x) = \frac{1}{2} l_{kn}^2(x) + \frac{u'_n(x)\, l_{kn}(x)}{2u'_n(x_{kn})} + c_{kn}\varrho_{kn}(x) + u_n(x)\, q_{n-1}(x) \tag{3. 16}$$

where $q_{n-1}(x)$ is defined by (2. 4). But from (2. 6), (3. 11) and (3. 3) we get

$$|B_{kn}| \leqq \frac{13}{2}(n+1), \qquad k = 1, 2, \ldots, n \tag{3. 17}$$

Similarly on using (3. 9), (2. 5), $\alpha = \frac{7}{4}$ we have

$$|c_{kn}\varrho_{kn}(x)| \leqq \frac{28}{\varepsilon^{3/2}}, \qquad n = 4, 6, \ldots . \tag{3. 18}$$

Since

$$|u'_n(x)| \leqq \frac{(n+2)}{\varepsilon^{3/2}}, \qquad -1+\varepsilon \leqq x \leqq 1-\varepsilon$$

and

$$|l_{kn}(x)| \leqq 4$$

we have

$$\left| \frac{u_n'(x) l_{kn}(x)}{2u_n'(x_{kn})} \right| \leqq \frac{3}{\varepsilon^{3/2}} \qquad n \geqq 4. \tag{3.19}$$

Further from (2. 4), (3. 1), (3. 3) and (3. 17) we get

$$|u_n(x) q_{n-1}(x)| \leqq \frac{13(1-x_{kn}^2)}{(n+1)\varepsilon^{3/2}} \tag{3.20}$$

for $-1+\varepsilon \leqq x \leqq 1-\varepsilon$. Now using (3. 16)—(3. 20) we get (3. 14).

LEMMA 3. 4. *Let $f(x)$ be a continuous function in $[-1, +1]$ satisfying the Zygmund condition (1. 8) in $(-1, +1)$ then there exists a sequence of polynomials $\varphi_n(x)$ with the following properties:*

$$|f(x)-\varphi_n(x)| = o\left(\frac{1}{n}\right)\left[(1-x^2)^{1/2}+\frac{1}{n}\right] \tag{3.21}$$

$$|\varphi_n'(x)| = o(\log n) \tag{3.22}$$

$$|\varphi_n''(x)| = o(n) \min\left[(1-x^2)^{-1/2}, n\right] \tag{3.23}$$

which hold uniformly in $[-1, +1]$.

For the proof of this lemma see FREUD, G. [4].

4. PROOF *of the convergence theorem.* Owing to the uniqueness theorem we have

$$\varphi_n(x) = \sum_{k=1}^{n} \varphi_n(x_{kn}) r_{kn}(x) + \sum_{k=1}^{n} [(1-x^2)^{7/4} \varphi_n(x)]''_{x=x_{kn}} \varrho_{kn}(x) \tag{4.1}$$

where $\varphi_n(x)$ is defined by Lemma 3. 4. Since

$$|R_n(f, x)-f(x)| \leqq |R_n(f, x)-\varphi_n(x)| + |\varphi_n(x)-f(x)| \tag{4.2}$$

Therefore on using (1. 7) and the above representation of $\varphi_n(x)$ we get

$$|R_n(x, f)-\varphi_n(x)| \leqq |S_1| + |S_2| + |S_3|$$

where on using (3. 21) and (3. 14) we have

$$|S_1| = \sum_{k=1}^{n} |f(x_{kn})-\varphi_n(x_{kn})|\,|r_{kn}(x)| = o\left(\frac{1}{n}\right)\frac{50n}{\varepsilon^{3/2}} = o(1)$$

Applying (1. 9) and (3. 10) we have

$$|S_2| = \sum_{k=1}^{n} |y_{k2}|\,|\varrho_{kn}(x)| = o(1)$$

Lastly, on using (3. 22), (3. 23) we have

$$[(1-x^2)^{7/4}\varphi_n(x)]''_{x=x_{kn}} \leqq \frac{9}{(1-x_{k)}^2)^{1/4}} + o(\log n)\, o(n) \min [n, (1-x_k^2)^{-1/2}]$$

Therefore on using the estimation of $\varrho_{kn}(x)$ as stated (3. 9) we get

$$|S_3| = o(1)$$

Therefore $|R_n(f, x) - \varphi_n(x)| = o(1)$ and so using (3. 21) we get $|R_n(x, f) - f(x)| = o(1)$. This completes the proof of the theorem.

REFERENCES

[1] BALÁZS, J. and TURÁN, P.: Notes on interpolation II, *Acta. Math. Acad. Sci. Hung.* **8** (1957), 201—215.

[2] BALÁZS, J. and TURÁN, P.: Notes on interpolation III, Convergence, *ibid* **9** (1958), 195—214.

[3] BALÁZS, J.: On weighted (0, 2) interpolation on ultraspherical abscissas (Hungarian), *A Magyar Tud. Akad. Oszt. Közl.* **11** (3) (1961), 305—338.

[4] FREUD, G.: Bemerkung über die Konvergenz eines interpolations verfahrens von P. Turán, *Acta. Math. Acad. Sci. Hung.* **9** (1958), 337—341.

[5] SURÁNYI, J. and TURÁN, P.: Notes on interpolation I, ibid **6** (1955), 66—79.

University of Alberta Edmonton, Canada, University of Rajasthan Jaipur, India

(Received December 6, 1966)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 221—228.

ÜBER DIE ANZAHL DER KNOTENPUNKTE EINES LÄNGSTEN WEGES IN PLANAREN, KUBISCHEN, DREIFACH ZUSAMMENHÄNGENDEN GRAPHEN

von

R. LANG und H. WALTHER

Dieser Artikel stellt eine Fortsetzung der im Literaturverzeichnis angegebenen Arbeit [4] dar.

Ein *elementarer Weg* eines Graphen G ist ein doppelpunktfreier Kantenzug. Ein *längster Weg* W von G ist ein elementarer Weg von G mit maximaler Knotenpunktanzahl. Mit $|G|$ bezeichnen wir die Anzahl der Knotenpunkte von G, mit $|W|$ die Anzahl der in einem längsten Weg W von G enthaltenen Knotenpunkte. $M(G)$ sei die Anzahl der in allen längsten Wegen von G enthaltenen Knotenpunkte, $P(G)$ sei die Anzahl der in keinem längsten Weg von G enthaltenen Knotenpunkte.

Das Anliegen dieser Arbeit ist der Beweis des folgenden Satzes:

SATZ: *Es gibt eine Folge $\{G_n\}$ von planaren, kubischen, dreifach knotenzusammenhängenden Graphen mit folgenden Eigenschaften:*

(a) *Die Anzahl $|W_n|$ der Knotenpunkte, die in einem längsten Weg W_n von G_n liegen, ist gleich der Anzahl $M(G_n)$ der in allen längsten Wegen liegenden Knotenpunkte,*

(b) $$\lim_{n\to\infty} \frac{|W_n|}{|G_n|} = 0,$$

(c) $$\lim_{n\to\infty} \frac{P(G_n)}{|G_n|} = 1.$$

Zum Beweis benötigen wir einige Hilfssätze.

HILFSSATZ 1: *Der Graph T von Abb. 1 besitzt keinen Hamiltonkreis.*

(Beweis: siehe [3].)

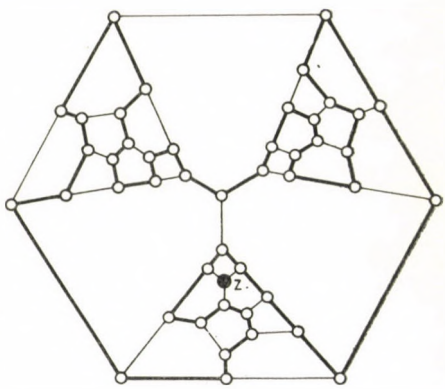


Abb. 1

Aus Abb. 1 ist ersichtlich, daß ein längster Kreis von T genau 45 der 46 Knotenpunkte von T enthält. Abb. 1 zeigt einen längsten Kreis, der den fest vorgegebenen Knotenpunkt z nicht enthält. T ist offensichtlich planar, kubisch und dreifach knotenzusammenhängend. Aus T bauen wir gemäß Abb. 2 ein Gebilde T'. Das Gebilde T' besitzt 55 Knotenpunkte. Der Knotenpunkt z werde besonders markiert (siehe Abb. 2). Das Gebilde H_1 (Abb. 3) entstehe aus dem Gebilde T', indem jeder Knotenpunkt von T' durch ein Dreieck ersetzt wird. Das Gebilde H_2 entstehe aus T' (Abb. 4), indem mit Ausnahme des Knotenpunktes z alle Knotenpunkte von T' durch ein Dreieck ersetzt werden. H_1 besitzt 165 Knotenpunkte,

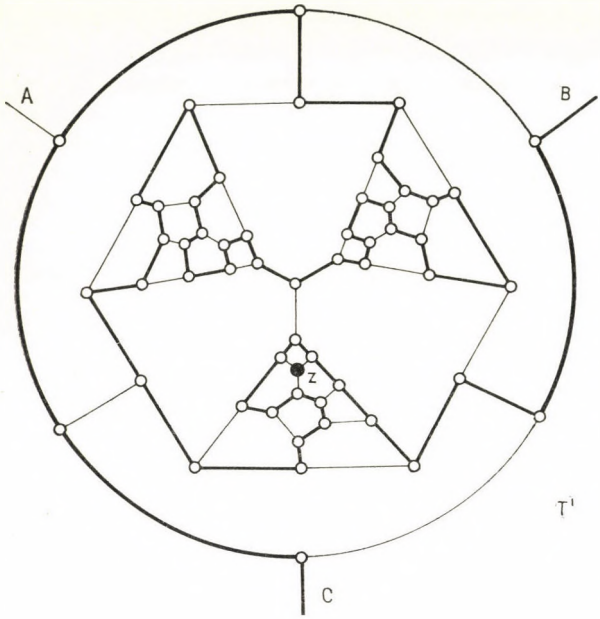


Abb. 2

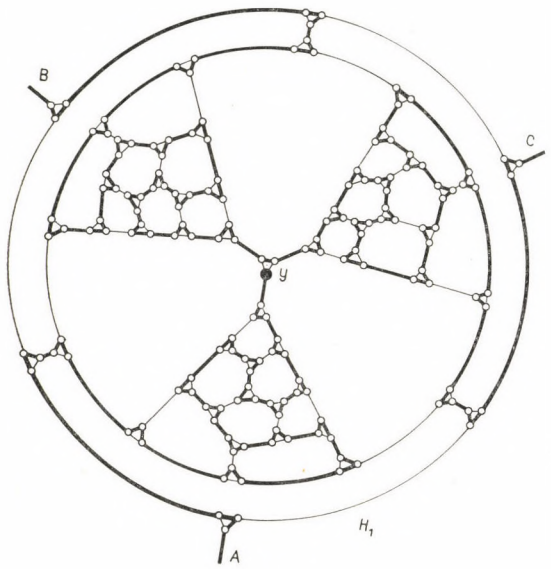


Abb. 3

H_2 besitzt 163 Knotenpunkte. In H_1 zeichnen wir den Knotenpunkt y besonders aus (Abb. 3).

Definition: Sei H_1 oder H_2 Teil eines Graphen G und sei W ein längster Weg von G. Wir sagen, H_1 bzw. H_2 wird von W *durchlaufen*, wenn Knotenpunkte von H_1 bzw. H_2 in W liegen, jedoch kein Endpunkt von W in H_1 bzw. H_2 liegt; H_1 bzw. H_2 wird von W *erfaßt*, wenn alle drei Kantenansätze A, B, C von H_1 bzw. H_2 in W liegen.

HILFSSATZ 2: *Ist H_1 (Abb 3) Teil eines Graphen G mit einem längsten Weg W und wird H_1 von W erfaßt, dann liegen alle* 165 *Knotenpunkte von H_1 in W.*

BEWEIS: Abb. 3 zeigt eine Möglichkeit.

HILFSSATZ 3: *Ist H_2 Teil eines Graphen G mit einem längsten Weg W und wird H_2 von W durchlaufen, dann liegen* 162 *der* 163 *Knotenpunkte von H_2 in W, und z liegt nicht in W.*

BEWEIS: Siehe [4], Hilfssatz 4. *(Abb. 4)*

Wir konstruieren nun eine Graphenfolge $\{G_n\}$, welche die im Satz genannten Eigenschaften besitzt.

Der Graph G_1 entstehe durch Zusammenfügen zweier Gebilde H_1 (Abb. 5). Der Graph G_1 besitzt 330 Knotenpunkte. Wir können einen längsten Weg W_1 von G_1 angeben, der sämtliche Knoten-

punkte von G_1 enthält und dessen Endpunkte y_1 und y_1' sind.

Der Graph G_2 entstehe aus G_1, indem wir die Knotenpunkte y_1 und y_1' durch je ein Gebilde H_1 ersetzen, während wir alle übrigen Knotenpunkte von G_1 durch je ein Gebilde H_2 ersetzen. Die y-Knotenpunkte (Abb. 3) der beiden eingesetzten Gebilde H_1 bezeichnen wir mit y_2 und y_2'.

Der Graph G_{n+1} entstehe aus G_n, indem wir die Knotenpunkte y_n und y_n' durch je ein Gebilde H_1 ersetzen, alle anderen Knotenpunkte von G_n ersetzen wir jedoch durch je ein Gebilde H_2. Die y-Knotenpunkte der beiden eingesetzten Gebilde H_1 bezeichnen wir mit y_{n+1} und y_{n+1}'.

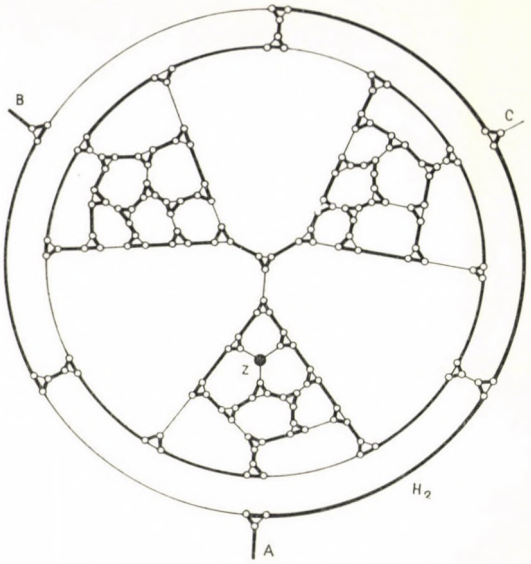


Abb. 4

Wir werden zeigen, daß diese Konstruktion so beschaffen ist, daß die Endpunkte jedes längsten Weges von G_n in den zwei verschiedenen Gebilden H_1 liegen.

Definition: Wir nennen einen Knotenpunkt x_n von G_n einen *Vorgänger* von x_m aus G_m ($n \leqq m$), wenn beim Zusammenziehen von G_m auf G_n (das Einsetzen der Gebilde H_1 und H_2, welches von G_n zu G_m geführt hat, wird rückgängig gemacht)

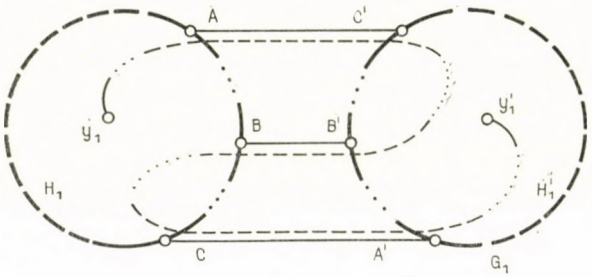


Abb. 5

der Knotenpunkt x_m in den Knotenpunkt x_n übergeht. Der Knotenpunkt x_m von G_m heißt auch *Nachfolger* von x_n aus G_n. Insbesondere ist jeder Knotenpunkt sein eigener Nachfolger und Vorgänger.

HILFSSATZ 4: *Es gibt in G_n einen längsten Weg W_n, dessen Endpunkte y_n und y_n' sind und der von den beiden Gebilden H_1 alle drei Kantenansätze enthält.*

BEWEIS: Für $n=1$ ist die Behauptung richtig, wie Abb. 5 zeigt. Es sei W_{n-1} ein längster Weg von G_{n-1}, der in y_{n-1} und y'_{n-1} endet. Beim Übergang von G_{n-1} zu G_n wird jeder innere Knotenpunkt von W_{n-1} durch ein Gebilde H_2 ersetzt,

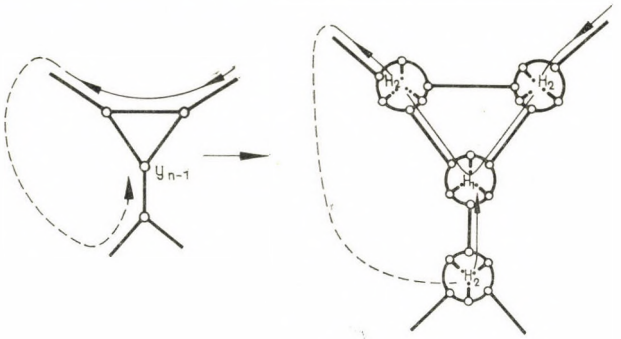


Abb. 6

während die Endpunkte dieses Weges durch ein H_1 ersetzt werden. Wir konstruieren einen Weg V_n von G_n aus W_{n-1}, der aus jedem der in die Knotenpunkte von W_{n-1} eingesetzten H_2 genau 162 der 163 Knotenpunkte enthält (Hilfssatz 3) und gemäß Abb. 6 von den beiden in y_{n-1} und y'_{n-1} eingesetzten Gebilden H_1 alle drei Kantenansätze enthält, in y_n und y'_n endet und alle Kotenpunkte der beiden H_1 enthält (Hilfssatz 2). Es bleibt zu zeigen, daß V_n ein längster Weg von G_n ist. Die Knotenpunktanzahl von V_n ist

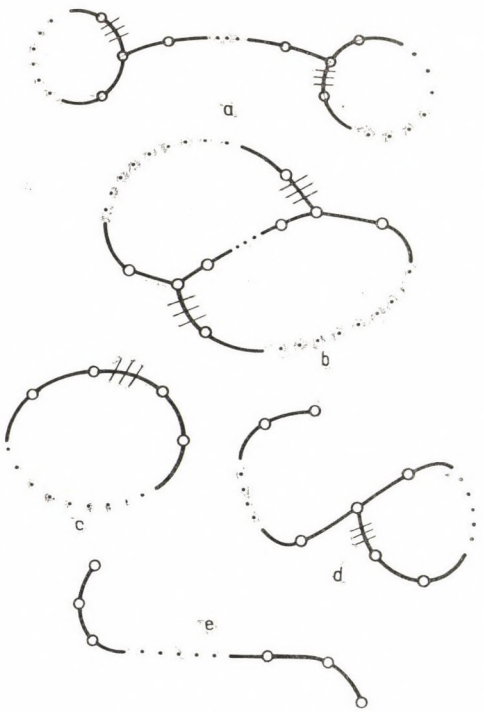


Abb. 7

$$(1) \qquad |V_n| = (|W_{n-1}|-2)\cdot 162 + 2\cdot 165.$$

Angenommen, es gäbe in G_n einen längsten Weg W_n mit

$$|W_n| > |V_n|.$$

Beim Zusammenziehen von G_n auf G_{n-1} geht W_n in eines der 5 Gebilde der Abb. 7 über.

Wie Abb. 7 zeigt, kann man aus den Gebilden a und b durch Löschen zweier Kanten und aus den Gebilden c und d durch Löschen einer Kante stets einen Weg V_{n-1} konstruieren mit gleicher Anzahl von Knotenpunkten. Da W_{n-1} ein längster Weg von G_{n-1} ist, gilt

$$(2) \qquad |W_{n-1}| \geqq |V_{n-1}|.$$

In den Fällen *a*, *b*, *d*, *e* werden von W_n genau $|V_{n-1}|-2$ Gebilde H (H_1 oder H_2) durchlaufen und höchstens von 2 Gebilden H werden alle Knotenpunkte erfaßt, im Falle *c* werden $|V_{n-1}|-1$ Gebilde durchlaufen und höchstens von einem Gebilde H werden alle Knotenpunkte erfaßt. Für die Anzahl der Knotenpunkte von W_n gilt also

$$|W_n| \leqq 162|V_{n-1}|+6. \tag{3}$$

Aus (1), (2), (3) ergibt sich

$$|W_n| \leqq 162|V_{n-1}|+6 \leqq 162|W_{n-1}|+6 = |V_n|.$$

Das ist aber ein Widerspruch zu der Annahme $|W_n| > |V_n|$. Also ist V_n längster Weg von G_n. Damit ist der Hilfssatz bewiesen.

HILFSSATZ 5: *Sei W_n ein längster Weg von G_n. Dann liegt in jedem der beiden in y_{n-1} und y'_{n-1} eingesetzten Gebilde H_1 ein Endpunkt von W_n, und alle Knotenpunkte dieser beiden H_1 liegen in W_n.*

BEWEIS: Angenommen, der Hilfssatz wäre falsch. Es sei W_n ein längster Weg von G_n, der nicht die Eigenschaften des Hilfssatzes hat. Ziehen wir G_n auf G_{n-1} zusammen, so geht W_n in eines der 5 Gebilde von Abb. 7 über. Gemäß den Überlegungen beim Beweis von Hilfssatz 4 kann man aus diesem Gebilde durch Weglassen von zwei bzw. einer bzw. keiner Kante einen Weg V_{n-1} bilden. V_{n-1} ist ein längster Weg von G_{n-1}, da andernfalls ($|W_{n-1}| > |V_{n-1}|$) W_n nur Knotenpunkte aus $|V_{n-1}|$ Gebilden H enthält, von denen mindestens $|V_{n-1}|-2$ durchlaufen werden und von höchstens zweien alle Knotenpunkte in W_n liegen können, das heißt aber, da W_n von jedem durchlaufenen Gebilde H nur 162 Knotenpunkte enthalten kann,

$$|W_n| \leqq 162(|V_{n-1}|-2)+2\cdot 165 < 162|W_{n-1}|+6.$$

Aus dem Beweis von Hilfssatz 4 ergibt sich aber

$$|W_n| = 162|W_{n-1}|+6.$$

Ist aber V_{n-1} ein längster Weg von G_{n-1}, dann kann man sich entsprechend überlegen, daß W_n von der in der Behauptung des Hilfssatzes angegebenen Form ist. Damit ist der Hilfssatz bewiesen.

HILFSSATZ 6: *Für die Anzahl der Knotenpunkte eines längsten Weges W_n von G_n gilt*

$$|W_n| = 162^{n-1}|W_1|+6\frac{162^{n-1}-1}{162-1} \qquad (n \geqq 1).$$

BEWEIS: Aus dem Beweis von Hilfssatz 4 ergibt sich

$$|W_n| = 162|W_{n-1}|+6.$$

Die Richtigkeit der obigen Formel ergibt sich sofort durch vollständige Induktion.

HILFSSATZ 7: *Für die Anzahl $|G_n|$ der Knotenpunkte von G_n gilt*

$$|G_n| = 163^{n-1}|G_1|+4\frac{163^{n-1}-1}{163-1} \qquad (n \geqq 1).$$

Der Beweis folgt unmittelbar aus der Konstruktion der Graphenfolge, denn es gilt

$$|G_n| = 163|G_{n-1}|+4.$$

HILFSSATZ 8: *Sei x_n aus G_n Nachfolger eines Knotenpunktes z aus G_i $(i \leqq n)$. Dann gibt es keinen längsten Weg W_n von G_n, der x_n enthält.*

BEWEIS: Angenommen, der Hilfssatz wäre falsch. Es sei n die kleinste natürliche Zahl, für die es einen längsten Weg W_n von G_n gibt und einen Knotenpunkt x_n, der Nachfolger eines Knotenpunktes z aus G_i $(i \leqq n)$ ist, der auf W_n liegt. Wegen Hilfssatz 3 ist x_n kein z, ist also ein „echter" Nachfolger eines Knotenpunktes z. Wir ziehen G_n auf G_{n-1} zusammen, dabei gehe der Knotenpunkt x_n in x_{n-1} über. Dann geht W_n in eines der Gebilde der Abb. 7 über. Durch Entfernen von 2 (Abb. 7a, b) bzw. 1 (Abb. 7c, d) bzw. 0 (Abb. 7e) Kanten entsteht aus diesem Gebilde ein Weg V_{n-1} mit gleicher Knotenpunktanzahl. Da x_{n-1} ein Nachfolger eines z ist, kann V_{n-1} wegen der Minimalität von n kein längster Weg von G_{n-1} sein. Es gibt also einen längsten Weg W_{n-1} mit $|W_{n-1}| > |V_{n-1}|$. Wie im Beweis von Hilfssatz 4 ausgeführt wurde, gilt

$$|W_n| = 162|W_{n-1}|+6.$$

Mit $|W_{n-1}| > |V_{n-1}|$ erhält man

$$|W_n| > 162|V_{n-1}|+6.$$

Andererseits enthält W_n Knotenpunkte aus $|V_{n-1}|$ Gebilden H_1 bzw. H_2, von denen mindestens $|V_{n-1}|-2$ durchlaufen werden, also gilt

$$|W_n| \leqq 162(|V_{n-1}|-2)+2\cdot 165 = 162|V_{n-1}|+6.$$

Das ist aber ein Widerspruch. Damit ist der Hilfssatz bewiesen.

HILFSSATZ 9: *Ist x_n aus G_n nicht Nachfolger irgendeines Knotenpunktes z, dann liegt x_n in jedem längsten Weg von G_n.*

BEWEIS: Angenommen, der Hilfssatz wäre falsch. Es sei n die kleinste natürliche Zahl, für die es einen Knotenpunkt x_n und einen längsten Weg W_n in G_n derart gibt, daß x_n weder Nachfolger eines Knotenpunktes z ist noch in W_n liegt. Wie man sich überzeugt, liegt kein Knotenpunkt desjenigen Gebildes H in W_n, in dem x_n liegt. Wir ziehen wieder G_n auf G_{n-1} zusammen. Dabei geht W_n in eines der Gebilde von Abb. 7 über. Durch Entfernen geeigneter Kanten entsteht ein Weg V_{n-1} (siehe auch Hilfssatz 4). Der Vorgänger x_{n-1} aus G_{n-1} von x_n liegt nicht in V_{n-1}. Da x_{n-1} nicht Nachfolger eines Knotenpunktes z ist (denn x_n ist nicht Nachfolger eines z) und da n die kleinste natürliche Zahl obiger Eigenschaft ist, kann V_{n-1} nicht längster Weg von G_{n-1} sein. Mit den gleichen Argumenten wie beim Beweis des vorigen Hilfssatzes kann man zeigen, daß dann W_n nicht längster Weg von G_n ist. Das ist aber ein Widerspruch.

FOLGERUNG: *Jeder längste Weg W_n von G_n enthält genau die Knotenpunkte, die nicht Nachfolger irgendeines Knotenpunktes z sind.*

Damit ist die Aussage (a) des Satzes bewiesen.

Zum Beweis von Aussage (b) des Satzes bilden wir mit den Ergebnissen der Hilfssätze 6 und 7

$$\lim_{n\to\infty}\frac{|W_n|}{|G_n|}=\lim_{n\to\infty}\frac{162^{n-1}|W_1|+6\dfrac{162^{n-1}-1}{162-1}}{163^{n-1}|G_1|+4\dfrac{163^{n-1}-1}{163-1}}\leqq\lim_{n\to\infty}\left(\frac{162}{163}\right)^{n-1}\cdot\frac{|W_1|+1}{|G_1|}=0.$$

Die Aussage (c) des Satzes folgt unmittelbar aus (a) und (b).

MOTZKIN und GRÜNBAUM [2] konstruierten Graphen $G(n)$ mit n Knotenpunkten und der Länge $p(n)$ eines längsten Weges, für die

$$p(n)<2\cdot n^{1-\alpha}\quad\text{mit}\quad\alpha=2^{-19}\quad\text{gilt}\quad(n=2,4,\ldots).$$

Die von uns in dieser Arbeit konstruierte Graphenfolge ist nicht für alle geradzahligen Knotenpunktanzahlen erklärt. Die von MOTZKIN und GRÜNBAUM konstruierte Graphenfolge besitzt offenbar auch die Eigenschaft (b) unseres Satzes, jedoch nicht die Eigenschaften (a) und (c), wie man sich an Hand der Konstruktion leicht überzeugt.

Aus der Graphenfolge $\{G_m\}$ kann man eine andere Folge $\{G_{m,j,i}\}$ von Graphen konstruieren, die für jedes geradzahlige n einen Graphen mit n Knotenpunkten besitzt und die Eigenschaft hat, daß ein Knotenpunkt genau dann in einem längsten Weg liegt, wenn er in allen längsten Wegen liegt (solche Graphen nennen wir *zulässig*).

Sei $P_1^m, P_2^m, \ldots, P_i^m, \ldots, P_k^m$ eine Anordnung der Knotenpunkte von G_m, wobei $W_m=(P_2^m, P_3^m, \ldots, P_i^m, P_1^m)$ mit $i=|W_m|$ ein beliebiger, aber fester längster Weg von G_m ist, und es gelte $k=|G_m|$.

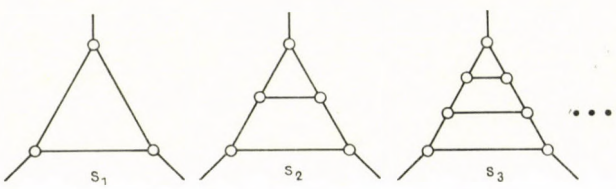


Abb. 8

$G_{m,0,i}$ entstehe aus G_m, indem man in P_1^m ein Gebilde S_i (Abb. 8) einsetzt, $i=1, 2, \ldots, 81$.

$G_{m,1,i}$ entstehe aus G_m, indem man in P_1^m ein Gebilde H_1 und in P_2^m ein Gebilde S_i einsetzt, $i=1, 2, \ldots, 81$.

$G_{m,2,i}$ entstehe aus G_m, indem man in P_1^m und P_2^m je ein Gebilde H_1 und in P_3^m ein Gebilde S_i einsetzt, $i=1, 2, \ldots, 80$.

$G_{m,j,i}$ entstehe aus G_m, indem man in P_1^m und P_2^m je ein H_1, in $P_3^m, P_4^m, \ldots, P_j^m$ je ein H_2 und in P_{j+1}^m ein S_i einsetzt, $i=1, 2, \ldots, 80$; $j=3, 4, \ldots, k$.

$G_{m,k,81}=G_{m+1}$.

Man kann sich davon überzeugen, daß alle Graphen $G_{m,j,i}$ zulässig sind und daß es zu jedem geradzahligen n einen Graphen $G_{m,j,i}$ mit $|G_{m,j,i}|=n$ gibt.

Für die Anzahl $|W_{m,j,i}|$ der Knotenpunkte eines längsten Weges $W_{m,j,i}$ von $G_{m,j,i}$ gilt offenbar

$$|W_{m,j,i}| \leqq |W_{m+1}|,$$

und es gilt

$$\max_{\substack{1\leqq i\leqq 81(80)\\ 0\leqq j\leqq k=|G_m|}} \frac{|W_{m,j,i}|}{|G_{m,j,i}|} \leqq \frac{|W_{m+1}|}{|G_m|}.$$

Sei nun n derart, daß

$$|G_m| < n \leqq |G_{m+1}|$$

gilt.

Dann gilt für die Länge $q(n)$ eines längsten Weges $W_{m,j,i}$ von $G_{m,j,i}$ mit $|G_{m,j,i}|=n$:

$$\frac{|W_{m,j,i}|}{|G_{m,j,i}|} = \frac{q(n)}{n} \leqq \frac{|W_{m+1}|}{|G_m|} = \frac{162^m|W_1|+6\dfrac{162^m-1}{162-1}}{163^{m-1}|G_1|+4\dfrac{163^{m-1}-1}{163-1}} <$$

$$< 164\left(\frac{162}{163}\right)^m, \quad \text{da} \quad |W_1| = |G_1| = 330 \text{ ist.}$$

In Anlehnung an die obere Schranke $M(n)=2n^{1-\alpha}$ für die Länge $p(n)$ eines längsten Weges bei MOTZKIN und GRÜNBAUM kann man nachweisen, daß für alle $\alpha \leqq 2^{-10}$

$$\frac{\dfrac{q(n)}{n}}{\dfrac{M(n)}{n}} = \frac{q(n)}{M(n)} \leqq \frac{164\left(\dfrac{162}{163}\right)^m}{2n^{1-\alpha}} < 83\left(\frac{162}{163}\cdot 163^{\alpha}\right)^m \xrightarrow[m\to\infty]{} 0$$

gilt, wie man unschwer zeigen kann. Diese Schranke ist also ein wenig besser, als die von GRÜNBAUM und MOTZKIN angegebene.

LITERATUR

[1] BERGE, C.: *Théorie des Graphes et ses Applications,* Dunod, Paris, 1958.
[2] GRÜNBAUM, B. und MOTZKIN, T. S.: Longest simple paths in polyhedral graphs, *J. London Math. Soc.* **37** (1962), 152—160.
[3] TUTTE, W.T.: On Hamiltonian circuits, *J. London Math. Soc.* **21** (1946), 98—101.
[4] WALTHER, H.: Über die Anzahl der Knotenpunkte eines längsten Kreises in planaren, kubischen, dreifach knotenzusammenhängenden Graphen, *Studia Sci. Math. Hung.* **2** (1967), 391—398.

Institut für Mathematik der TH Ilmenau, DDR

(Eingegangen: 17. Dezember 1967.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 229—247.

GROUP-TESTING TO CLASSIFY ALL UNITS IN A TRINOMIAL SAMPLE

by

S. KUMAR

1. Introduction

A finite number N of units are to be classified into one of the three disjoint categories by means of group testing. The three categories are labeled as *good*, *mediocre* and *defective*. A group test is a simultaneous test on x units ($1 \leqq x \leqq N$) with one of the following three possible outcomes: (i) all the x units are good, (ii) among the x units at least one is mediocre and none are defective, (iii) at least one of the x units is defective. For the sample outcome (ii) and $x \geqq 2$ we do not know which ones or how many units are mediocre and similarly for the sample outcome (iii) and $x \geqq 2$. The problem is to define a simple and efficient procedure (or an optimal procedure) for classifying all the N units into one of the three disjoint categories. Each unit is assumed to represent an independent observation from a trinomial population with known a priori probabilities q_1, q_2 and q_3 (with $q_i \geqq 0$ for $i=1, 2, 3$ and $q_1+q_2+q_3 = 1$) of being good, mediocre and defective respectively.

In section 2 a procedure R_1 which describes a mode of action for any given value of $\vec{q}=(q_1, q_2, q_3)$ is defined. This procedure is similar to the procedure R_1 defined for the binomial problem in [6]. In section 3 some properties of the procedure R_1 are studied. Under R_1, the identification of the units in the same test group is not required. Another procedure R_2, where the identification of the units in the same group test is sometimes required is proposed in section 4. Section 5 deals with some properties of the optimal procedure R_0. These properties are similar to the properties of the optimal procedure for the binomial problem in [11]. In section 6 it is shown that, for $q_1 < \frac{1}{2}[q_2 - 1 + (5q_2^2 - 6q_2 + 5)^{1/2}]$, expected number of tests for the classification of N units under the procedure R_1 is the same as under the procedure R_0. In section 7, a theorem concerning the lower bound for the expected number of tests under any group-testing procedure using information theory is proved. Section 8 deals with finding the lower bound for the expected number of tests under any grouptesting procedure. HUFFMAN's method of construction [4] of compact codes is used for finding the lower bound. Table I gives the size of the next test group for various situations arising in the classification of N ($\leqq 8$) units under R_1, with $q_1 = .90$, $q_2 = q_3 = .05$. Table II gives the comparison of the expected number of tests for the procedure R_1 and information theory lower bounds for any procedure for the classification of N ($\leqq 8$) units with $q_1 = .90$, $q_2 = q_3 = .05$.

2. The Procedure R_1

The procedure R_1 is defined by a number of recursion formulae and boundary conditions. Before writing the formula for R_1, we shall need some definitions and results. A set of units will be called a defective set if it is known to contain at least one defective unit.

For a set of size m, the conditional probability that Z_1, the number of defective units present, equals z given that the set is defective is

$$\text{(2. 1)} \qquad \mathsf{P}\{Z_1 = z | Z_1 \geqq 1\} = \frac{\binom{m}{z} q_3^z q_{[2]}^{m-z}}{1 - q_{[2]}^m} \qquad (z = 1, 2, \ldots, m)$$

where $q_{[2]} = q_1 + q_2$.

Let x denote the size of a proper (i.e., non-trivial and non-empty) subset randomly chosen from the defective set of size m and let Z_2 denote the number of defective units present in this subset. Then the probability that it contains at least one defective unit is

$$\text{(2. 2)} \qquad \mathsf{P}\{Z_2 \geqq 1 | Z_1 \geqq 1\} = \sum_{z=1}^{m} \sum_{y=1}^{z} \frac{\binom{z}{y}\binom{m-z}{x-y}}{\binom{m}{x}} \frac{\binom{m}{z} q_3^z q_{[2]}^{m-z}}{1 - q_{[2]}^m} = \frac{1 - q_{[2]}^x}{1 - q_{[2]}^m}$$

where we use the hypergeometric identity and we define $\binom{z}{y} = 0$ if $y > z$ or $z < 0$.

Let Y_1 be the chance variable representing the number of mediocre units present in the defective set and Y_2 be the chance variable representing the number of mediocre units present in the proper subset of size x randomly chosen from the defective set of size m. Then, for a defective set of size m, the conditional probability that $Y_1 + Z_1$, the number of mediocre plus the number of defective units present, equals a is

$$\text{(2. 3)} \qquad \mathsf{P}\{Y_1 + Z_1 = a | Z_1 \geqq 1\} = \frac{\sum_{z=1}^{a} \frac{m!}{z!(a-z)!(m-a)!} q_1^{m-a} q_2^{a-z} q_3^z}{1 - q_{[2]}^m} =$$

$$= \frac{\binom{m}{a} [(q_2 + q_3)^a - q_1^a] q_1^{m-a}}{1 - q_{[2]}^m}, \qquad (a = 1, 2, \ldots, m)$$

Hence the probability that a randomly chosen proper subset of size x from the defective set of size m, contains all good units is

(2. 4)

$$\mathsf{P}\{Y_2 + Z_2 = 0 | Z_1 \geqq 1\} = \sum_{r=1}^{m-x} \frac{\binom{m-r}{x}}{\binom{m}{x}} \frac{\binom{m}{r} [(q_2 + q_3)^r - q_2^r] q_1^{m-r}}{1 - q_{[2]}^m} = \frac{q_1^x (1 - q_{[2]}^{m-x})}{1 - q_{[2]}^m}.$$

Thus the probability that a proper subset of size x randomly chosen from a defective set of size m contains at least one mediocre unit and no defective unit is

(2. 5)

$$\mathsf{P}\{Y_2 \geqq 1, Z_2 = 0 | Z_1 \geqq 1\} = 1 - \frac{1-q_{[2]}^x}{1-q_{[2]}^m} - \frac{q_1^x(1-q_{[2]}^{m-x})}{1-q_{[2]}^m} = \frac{(1-q_{[2]}^{m-x})(q_{[2]}^x - q_1^x)}{1-q_{[2]}^m}.$$

A set of units will be called a mediocre set if it is known that it contains at least one mediocre unit and no defective unit; let W_1 denote the (random) number of mediocre units it contains. Let x denote the size of a proper subset randomly chosen from a mediocre set of size m and let W_2 denote the number of mediocre units in the subset. Then the probability that the subset has only good units is given by

$$\mathsf{P}\{W_2 = 0 | W_1 \geqq 1\} = \frac{q^x(1-q^{m-x})}{1-q^m} \tag{2. 6}$$

where $q = q_1/q_{[2]}$ and the probability that this proper subset of size x is also a mediocre set is given by

$$\mathsf{P}\{W_2 \geqq 1 | W_1 \geqq 1\} = \frac{1-q^x}{1-q^m}. \tag{2. 7}$$

Now we shall introduce two lemmas which are of importance in writing the recursion formulae for the procedure R_1.

LEMMA 1. Given a mediocre set of size m and given that a randomly chosen subset of size x contains at least one mediocre unit, then the a posteriori distribution of the remaining $(m-x)$ units is that of a binomial sample with probabilities $q_1/q_{[2]}$ and $q_2/q_{[2]}$ of being good and mediocre, respectively.

We shall omit its proof because its proof is similar to that of lemma 2.

LEMMA 2. Given a defective set of size d and given that a randomly chosen proper subset of size x contains at least one defective unit, then the a posteriori distribution associated with the remaining $d-x$ units is the same as the original trinomial distribution.

PROOF. Let A be the set of size x randomly chosen from the defective set and B be the set of remaining $(d-x)$ units. Let A_1, B_1 denote the random number of good units, A_2, B_2 the random number of mediocre units and A_3, B_3 the random number of defective units present in A and B respectively. Let $\vec{A} = (A_1, A_2, A_3)$ and $\vec{B} = (B_1, B_2, B_3)$. Then for (b_1, b_2, b_3) such that b_i are non-negative integers and $\sum_{i=1}^{3} b_i = d-x$, we have

$$P_j = \mathsf{P}\{B_1 = b_1, B_2 = b_2, B_3 = b_3 \mid A_3 + B_3 \geqq 1,\ A_3 \geqq 1\} \tag{2. 8}$$

where P_j is defined by (2. 8). Since $A_3 \geqq 1$ implies $A_3 + B_3 \geqq 1$, therefore

$$P_j = \frac{\mathsf{P}\{B_1 = b_1, B_2 = b_2, B_3 = b_3, A_3 \geqq 1\}}{\mathsf{P}\{A_3 \geqq 1\}}. \tag{2. 9}$$

At the outset (and in the unconditional probability above) all the units are inde-

pendently and trinomially distributed with common probabilities q_1, q_2 and q_3 of being good, mediocre or defective respectively. Since the sets A and B are disjoint, it follows that $\vec{A}$ and $\vec{B}$ are independent vector chance variables. Hence the numerator in (2. 9) factors and after cancellation we get

$$P_j = \mathrm{P}\{B_1 = b_1, B_2 = b_2, B_3 = b_3\} = \frac{(m-x)!}{b_1!\, b_2!\, b_3!}\, q_1^{b_1} q_2^{b_2} q_3^{b_3},$$

which proves the lemma.

A set of units will be called a *trinomial set* if, given the past history of testing, the a posteriori distribution of this set of units is that of independent trinomial chance variables with common probabilities q_1 of being good, q_2 of being mediocre and q_3 of being defective $(q_1+q_2+q_3 = 1)$.

A set of units will be called a *conditional binomial* set if, given the past history of testing, the a posteriori distribution if this set of units is that of independent binomial chance variables with common probabilities $q = q_1/q_{[2]}$ of being good and $q_2/q_{[2]}$ of being mediocre.

The procedure R_1 requires that at every stage the unclassified units be separated into at most four sets, namely, the trinomial set, the defective set, the mediocre set and the conditional binomial set.

Let $G(c;\ m, n;\ d, e;\ \vec{q}) = G(m, n;\ d, e)$ denote the expected number of group-tests remaining to be performed if the procedure R_1 is used and if presently, the number of classified units is c, the mediocre set is of size m, the conditional binomial set is of size $n-m$, the defective set is of size d and the trinomial set is of size $e-d = N-c-n-d$; the a priori probability of a unit being good, mediocre and defective are known constants q_1, q_2 and $q_3 = 1-q_1-q_2$, respectively and we are using $\vec{q} = (q_1, q_2, q_3)$. For the special case when $m = d = 0$ we use the notation $H(n; e)$ instead of $G(0, n;\ 0, e)$. The values of c, d, e, m, n vary as the procedure R_1 of group testing is carried out. The situation of unclassified units will be referred to as a G-situation or $G(m, n;\ d, e)$-situation if $\max(m, d) \geqq 2$ and as a H-situation or $H(n; e)$-situation if $m = d = 0$. The case when $\max(m, d) = 1$ is excluded in the above definition because then the G-situation can be changed into H-situation without any group test (see the boundary conditions below) by classifying the unit in the mediocre set, if any, as mediocre and the unit in the defective set, if any, as defective.

Recursion formulae defining procedure R_1. For any H-situation with $e \geqq 1$, $n \geqq 0$ (and $m = d = 0$) we take a sample of size x from the trinomial set and we then have

$$H(n; e) = 1 + \min_{1 \leqq x \leqq e} \{q_1^x H(n; e-x) + (q_{[2]}^x - q_1^x) G(x, n+x; 0, e-x) + \\ + (1 - q_{[2]}^x) G(0, n; x, e)\}. \tag{2.10}$$

For any H-situation with $e = 0$, $n \geqq 1$ we take a sample of size x from the conditional binomial set and we then have

$$H(n; 0) = 1 + \min_{1 \leqq x \leqq n} \{q^x H(n-x; 0) + (1-q^x) G(x, n; 0, 0)\}. \tag{2.11}$$

With the help of lemma 1, (2. 6) and (2. 7), for any G-situation with $m \geqq 2$ (and

for any values of $n \geqq m$, $e \geqq d \geqq 0$) we take a sample of size x from the mediocre set and we then have

(2. 12)

$$G(m,n;\,d,e) = 1 + \min_{1 \leqq x \leqq m-1} \left\{ \frac{q^x(1-q^{m-x})}{1-q^m} G(m-x, n-x;\, d, e) + \frac{1-q^x}{1-q^m} G(x, n;\, d, e) \right\}.$$

With the help of Lemma 2, (2. 2), (2. 4) and (2. 5), for any G-situation with $m=0$, $d \geqq 2$ we take a sample of size x from the defective set and we then have

(2. 13)

$$G(0,n;\,d,e) = 1 + \min_{1 \leqq x \leqq d-1} \left\{ \frac{q_1^x(1-q_{[2]}^{d-x})}{1-q_{[2]}^d} G(0, n;\, d-x, e-x) + \right.$$
$$\left. + \frac{(q_{[2]}^x - q_1^x)(1-q_{[2]}^{d-x})}{1-q_{[2]}^d} G(x, n+x;\, d-x, e-x) + \frac{1-q_{[2]}^x}{1-q_{[2]}^d} G(0, n;\, x, e) \right\}.$$

The boundary conditions state that for all $\vec{q} = (q_1, q_2, q_3)$

(2. 14) $$H(0;\,0) = 0.$$

(2. 15) $$G(1, n;\, d, e) = G(0, n-1;\, d, e) \quad \text{for} \quad n \geqq 1,\ e \geqq d \geqq 0.$$

(2. 16) $$G(0, n;\, 1, e) = H(n;\, e-1) \quad \text{for} \quad n \geqq 0,\ e \geqq 1.$$

In (2. 10) to (2. 13) the expression in the braces is the conditional expected number of additional group-tests required to classify all units under procedure R_1 given the size x of the next group-test. It follows from (2. 10), (2. 14), (2. 15) and (2. 16) that $H(0;\,1) = 1$ for all $\vec{q}$.

REMARK 1. To justify writing $G(x, n;\, d, e)$ on the right side of (2. 12) we make use of Lemma 1. If the proper subset of size x randomly chosen from the mediocre set of size m is known to contain at least one mediocre unit, then the a posteriori distribution associated with the remaining $(m-x)$ units is exactly the same as the distribution associated with $(m-x)$ independent units in the conditional binomial set. These $m-x$ units are then recombined with $n-m$ units in the conditional binomial set giving a total of $n-x$ conditional binomial units, and this justifies the expression $G(x, n;\, d, e)$ in (2. 12). Similar use of Lemma 2 is made in writing $G(0, n;\, x, e)$ in (2. 13).

REMARK 2. These four recursion formulae, together with boundary conditions allow one to compute successively for any $\vec{q} = (q_1, q_2, q_3)$ the function $H(0;\,1)$, $G(2, 2;\, 0, 0)$, $G(0, 0;\, 2, 2)$, $H(0;\, 2)$, ... to any desired values of m, n, d, and e.

REMARK 3. The positive integer x which accomplishes the minimization in (2. 10), (2. 11), (2. 12) and (2. 13) for each situation characterized by the integers m, n, d and e is particularly important, since this is the size of the next group to be tested according to the procedure R_1. These integers $x = x_H(n;\, e;\, \vec{q})$ and $x = x_G(m, n;\, d, e;\, \vec{q})$ implicitly define the procedure R_1.

REMARK 4. If $m = d = 0$, $e \geqq 1$ then it follows from (2. 10) that under procedure R_1 a subset of size x with $1 \leqq x \leqq e$ is taken from the trinomial set without mixing

it with units from the conditional binomial set. If $m>1$, then it follows from (2. 12) that under procedure R_1 a subset of size x with $1 \leqq x < m$ is taken from the mediocre set without mixing it from units from the other sets. If $m=0$, $d>1$, then it follows from (2. 13) that under procedure R_1 a subset of size x with $1 \leqq x < d$ is taken from the defective set without mixing it with units from the other sets. Any lack of optimality for procedure R_1 can only arise from this "no mixing" assumption. Another procedure, which partially drops the "no mixing" assumption at the expense of more complication is introduced in a later section.

3. Properties of the Procedure R_1

In this section we consider properties of the procedure R_1 which are concerned with the size of the next test group; in particular we are interested in determining what this size depends on.

PROPERTY 1. We state this property as a

THEOREM. *For any $G(m, n; d, e)$-situation with $m \geqq 2$, the size of the next group test under the procedure R_1, defined by* (2. 12) *does not depend on n, d or e.*

PROOF. For this situation the procedure under R_1 is to break down the mediocre set until a mediocre unit is found and removed. (Instead of randomizing the units in this set each time before a test group is selected, it is assumed, without any loss of generality, the order is randomized only at the outset; units removed for testing will be taken in that order.) If the ith unit is the first mediocre unit, then the breaking down of the mediocre set leads to a situation in which the mediocre set is empty and the size of the conditional binomial set is increased by $m-i$ whereas the sizes of the other sets remain the same, i.e., it leads to $G(0, n-i; d, e)$-situation.

Let $F(m, \bar{q}) = F(m)$ be defined as the expected number of group tests required to breakdown the mediocre set of size m and reach (for the first time) a situation in which the mediocre set is empty, where $\bar{q}$ is given and the procedure R_1 is used. It follows from this definition that $F(m)$ does not depend on n, d or e. Using q for $q_1/q_{[2]}$, we have

$$G(m, n; d, e) = F(m) + \sum_{i=1}^{m} \frac{(1-q)q^{i-1}}{1-q^m} G(0, n-i; d, e). \tag{3.1}$$

Let $\left(\frac{1-q^m}{1-q}\right) F(m) = F^*(m)$ and $\left(\frac{1-q^m}{1-q}\right) G(m, n; d, e) = G^*(m, n; d, e)$.

Replacing F by F^* and G by G^* in (3. 1) and (2. 12) we obtain

$$G^*(m, n; d, e) = F^*(m) + \sum_{i=1}^{m} q^{i-1} G(0, n-i; d, e) \tag{3.2}$$

(3.3)

$$G^*(m, n; d, e) = \sum_{i=1}^{m} q^{i-1} + \min_{1 \leqq x < m} \{q^x G^*(m-x, n-x; d, e) + G^*(x, n; d, e)\}.$$

Substituting (3. 2) in (3. 3) and observing that the summation terms cancel, we obtain

$$F^*(m) = \frac{1-q^m}{1-q} + \min_{1 \leqq x < m} \{q^x F^*(m-x) + F^*(x)\} \tag{3.4}$$

which does not depend on n or d or e. The boundary condition is $F^*(1) = 0$ for all $\bar{q}$, which also does not depend on n, d or e. It is clear from the derivation that (3. 4) which does not depend on n or d or e, must define the same integer value x as defined by (2. 12).

It follows from (3. 4) that for any $G(m, n; d, e)$-situation with $m \geqq 2$ we can use the G-tables for the binomial problem in [6].

PROPERTY 2. It will now be shown that in the $G(0, n; d, e)$-situation the size x of the next group test taken from the defective set of size d may also depend on the size n of the conditional binomial set (as well as on the size $e-d$ of the trinomial set). We shall show only the first part, for this we consider the following example with $e-d = 0$. Let $q_1 = .78$, $q_2 = .12$ and $q_3 = .10$. Consider the situations $G(0, 0; 4, 4)$ and $G(0, 1; 4, 4)$. It is easy to show numerically that $x = 3$ cannot be the size of the next group test. Let $G^{(x)}(0, i; 4. 4)$, $(i = 0, 1; x = 1, 2)$ be the expected number of tests when x is the size of the next group test (from the defective group of size 4). From (2. 10)—(2. 16) we obtain for $i = 0$

$$G^{(1)}(0, 0; 4, 4) - G^{(2)}(0, 0; 4, 4) =$$

$$= \frac{1}{1-q_{[2]}^4}[(q_{[2]} - q_{[2]}^4) - (1-q_{[2]}^2)(q_1 q_2 + q_2 q_{[2]} + 1) = -.014,$$

which means that $x = 1$ is for this situation the size of the next group test under procedure R_1. Similarly for $i = 1$

$$G^{(1)}(0, 1; 4, 4) - G^{(1)}(0, 1; 4, 4) =$$

$$= \frac{1}{1-q_{[2]}^4}[q_{[2]} - q_{[2]}^4 - \{q_1 q_2 q_{[2]} q_3 + q_2 q_{[2]}(1-q_{[2]}^2) H(2; 0) + 1 - q_{[2]}^2 + q_1 q_2 q_3\}]$$

$$= .023,$$

which means that $x = 2$ is for this situation the size of the next group test under procedure R_1. Hence for the $G(0, n; d, e)$-situation it follows that the size x of the next group test may depend on the size of at least one of other sets.

PROPERTY 3. Using the results of [6], the following results for the situation $G(m, n; d, e)$ with $m \geqq 2$ are evident.

(i) Under the procedure R_1, with $2 \leqq m \leqq n$ and $0 < \frac{q_1}{q_{[2]}} < .618$,

$$F(m) = \frac{q_1}{q_2} + \frac{q_{[2]}^m - q_1^{m-1} q_{[2]} - m q_1^m}{q_{[2]}^m - q_1^m}$$

where $F(m)$ is the expected number of group-tests required to "break down" a mediocre set of size m.

(ii) Under the procedure R_1, with $0 \leqq \frac{q_1}{q_{[2]}} < .618$, the units from the mediocre set are tested one at a time until a mediocre unit is found.

(iii) Under the procedure R_1, for the $H(n;\ 0)$-situation with $0 < \frac{q_1}{q_{[2]}} < .618$, all the units are tested one at a time.

4. Procedure Allowing Mixing of the Trinomial and Conditional Binomial Sets

In this situation a new procedure R_2 is introduced where the mixing of the units from the trinomial and conditional binomial sets is allowed whenever we are in an H-situation. For this procedure R_2 three of the four recursion formulae are the same and will not be repeated here. For the two types of G-situations with $m \geqq 2$ or $m=0$, $d \geqq 2$ the procedure R_2 is based on the same recursion formulae, namely, (2. 12) and (2. 13), repsectively. Under R_2 the recursion formula for the H-situation with $e \geqq 1$ is replaced by

(4. 1)

$$H(n;\ e) = 1 + \min \begin{cases} \min\limits_{1 \leqq x \leqq e} \{q_1^x H(n; e-x) + (q_{[2]}^x - q_1^x) G(x, n+x; 0, e-x) + \\ \qquad + (1 - q_{[2]}^x) G(0, n; x, e)\} \\ \min\limits_{1 \leqq y \leqq n} \{q_1^e q^y H(n-y; 0) + (q_{[2]}^e - q_1^e q^y) G(e+y, n+e; 0, 0) + \\ \qquad + (1 - q_{[2]}^e) G(0, n; x, e)\} \end{cases}$$

The recursion formula for the H-situation with $e=0$, $n \geqq 1$ is again the same as (2. 11). The boundary conditions are also the same as in (2. 14)—(2. 16).

It follows from the structure of (4. 1) that under the procedure R_2 for any H-situation with $e \geqq 1$ the next test group will be either

(1) A set which is a subset of the trinomial set with no units from the conditional binomial set, or

(2) A set which contains the trinomial set and a non-empty subset of the conditional binomial set.

It should be pointed out that in order to apply the procedure R_2 without loss of "information" it is necessary for the H-situation with $e \geqq 1$ that the units of the trinomial set must be distinguishable from the units of the conditional binomial set, through the units within each set need not be distinguishable. Therefore, for any problem in which such an identification of the units is impossible economically or impractical, the procedure R_2 should not be used.

It will now be shown that, in some cases, the expected number of tests under the procedure R_2 is less than the expected number of tests under the procedure R_1. We need consider only a single numerical example. Let $N=3$, $q_1 = .65$, $q_2 = .25$ and $q_3 = .10$. Let $H_i(0;\ 3)$ be the expected number of tests under the procedure R_i $(i=1, 2)$. It is easy to show that

$$H_1(0;\ 3) - H_2(0;\ 3) = .046. \tag{4. 2}$$

Thus, in this example, the expected number of tests under the procedure R_2 is less than the expected number of tests under the procedure R_1.

5. Some Properties of the Optimal Procedure R_0

In this section we discuss some properties of the optimal procedure R_0. These properties are concerned with the question of when we should test one-at-a-time and are similar to the corresponding properties shown for the optimal procedure for the binomial problem in [11].

A group-testing procedure is called optimal among all procedures for given N and $\vec{q}=(q_1, q_2, q_3)$ if it minimizes the expected number of tests. Let R_0 denote the optimal group-testing procedure and let $\mathsf{E}(T|R_0, N, \vec{q})=\mathsf{E}(T)$ denote the expected number of groups-tests to be performed if the procedure R_0 is used for N units and for given $\vec{q}=(q_1, q_2, q_3)$. Let q_2^* be ddfined by

$$q_2^* = \frac{1}{2}[q_2 - 1 + (5q_2^2 - 6q_2 + 5)^{1/2}] \tag{5.1}$$

The optimal procedure R_0 for $N=1$ is, of course trivial and we now consider R_0 for $N \geqq 2$.

THEOREM. *The optimal procedure R_0 has the following properties for $N \geqq 2$:*

(i) *If $q_1 < q_2^*$, then $\mathsf{E}T = N$ and the units are tested one-at-a-time.*
(ii) *If $q_1 > q_2^*$, then $\mathsf{E}T < N$.*
(iii) *If $q_1 = q_2^*$, then $\mathsf{E}T = N$ and the optimal procedure is not unique.*

PROOF. We shall prove (i) by showing that, for $q_1 < q_2^*$, a group-testing procedure cannot be optimal if groups of more than one unit occur at any state of testing.

A group-testing procedure can be represented by a tree in the following way. The group G_1 on the top represents the group to be tested first. Let G_2 represent the group to be tested next if G_1 turns out to be good, G_3 represent the group to be tested next if G_1 turns out to be mediocre, G_4 represent the group to be tested

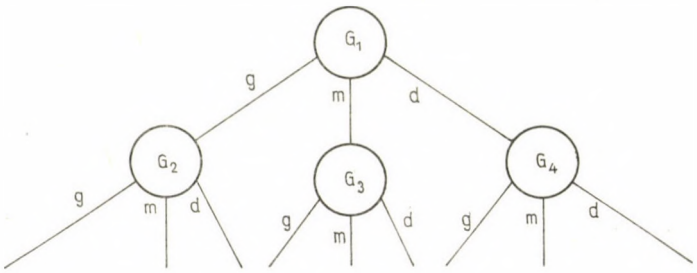


Figure 1

next if G_1 turns out to be defective. We write G_2, G_3 and G_4 below G_1 and connect these to G_1 by a "path". We can proceed with the representation of the procedure in a similar way. Two groups will be called "occurring on the same branch of the tree" if one of them can be reached from the other by descending all the way along one of the connecting paths.

We define a group-testing procedure to be "reasonable" if it has the following properties:

1. A group will not occur more than once on the same branch of the tree. This implies that if a group has already been tested we shall not test the same group again.

2. Let G be the group at any branch point B of the tree. No unit of G has been "previously" classified (i.e., classified by any branch point which has a direct path descending to G).

3. Let G be the group at any branch point B of the tree. There does not exist a group G' containing G on any of the branches below B.

4. A test will be skipped if the available information at that time enables us to infer the result of the test.

Any group-testing procedure which does not satisfy the above properties can be modified by removing elements from groups and skipping unnecessary tests so as to satisfy these properties. The number of tests needed to classify any sample is definitely not increased by these modifications.

The result (i) of the theorem is proved by considering an arbitrary procedure which satisfies the properties 1 to 4 and modifying it so that the expected number of tests under the new procedure is less than under the old procedure whenever $q_1 < q_2^*$. We start with a procedure which tests more than one unit at some point and we use the word "*plan*" to indicate a portion of this procedure. Let B be a branch point on our tree such that the group G to be tested at B has x units where $x \geqq 2$ and all the tests below B (if there are any below B) require the units to be tested individually. The branch of the tree is denoted by I, II or III in Fig. 2, according as the group test indicates that G is good, mediocre or defective. Let this plan be called the old plan; now we introduce another plan which will be called new plan. Let u denote a unit of G; u can be any unit of G except that if G is defective, then we assume that u is different from the last one to be tested among the units

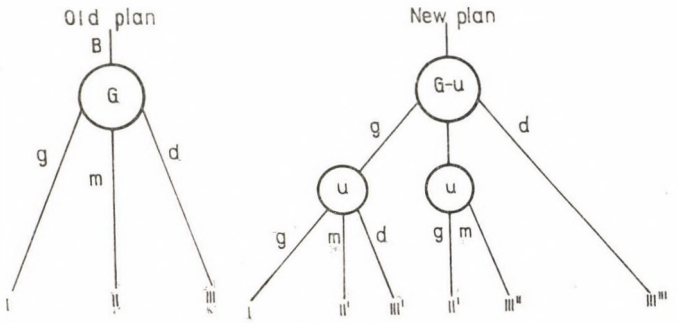


Figure 2

of G under the old plan. Instead of testing G at the branch point B, under the new plan we test $G-u$. If $G-u$ is found defective we continue as under the old plan where G is found defective. It might be possible to infer the results of some tests and thereby reduce the number of tests by using the additional information available under the new plan. Hence this branch may be different under the new plan and we denote it by III‴ instead of III.

If $G-u$ is mediocre, we test the unit u on the next test. If u is found good or mediocre we continue as under the old plan where G is found mediocre. We denote the branch to be followed by II′ instead of II because the availability of new information again might enable us to infer the results of some tests. If it is found defective we continue as under the old plan where G is found defective. We denote the branch to be followed by III″ because the availability of new information might enable us to infer the result of some test.

If $G-u$ is good, we test u on the next test. If u is good we continue as under the old plan where G is found good. If u is mediocre (or defective) we continue as under the old plan when G is found mediocre (or defective). We denote the branch to be followed by II′ (or III′) because the availability of new information might enable us to infer the result of some tests. The remainder of the procedure (i.e., everything which is not below B) is left unchanged. The procedure corresponding to the old plan (or new plan) is referred to as old procedure (or new procedure).

It is evident from the above construction of the new test plan that, for any sample, the number of tests under the new procedure can at most exceed by one the number of tests for the same sample under the old procedure.

Now we shall show that the only samples for which more tests are needed under the new procedure are those for which B is reached and G is found good under the old plan.

If the branch point B is not reached, then the number of tests for any sample under the old and the new procedures are equal since the two procedures are identical up to B. Hence in the following discussion we can assume that the branch point B is reached.

If $G-u$ is defective, we follow the same procedure as under the old plan except possibly for skipping a test which may have been necessary under the old plan. Hence, in this case, the number of tests under the new plan is less than or equal to the number of tests under the old plan.

If $G-u$ is mediocre, and u is mediocre or good, then we are following the old plan as in the case when G is mediocre except possibly skipping a test under the new plan which may have been necessary under the old plan. If $G-u$ is mediocre and u is defective, we need $(x+1)$ tests for classifying all the units of G under the old plan (one test for G and afterwards one test for each of the x units of G) whereas under the new plan we shall need either x tests (or $x+1$ tests) according as it is possible (or not) to infer the result of one test.

If $G-u$ is good and u is mediocre or defective, we need x (or $x+1$) tests for classifying all the units in G under the old plan according as it is possible (or not) to infer the result of one test whereas we need two tests under the new plan.

If all the units in G are good, we need one test to classify all the units in G under the old plan whereas two tests will be needed under the new plan.

Hence it is established that the only samples for which more tests are needed under the new procedure are those for which B is reached and G is found good under the old procedure.

We discuss the two cases $x=2$ and $x\geqq 3$ separately:

Case I. $x=2$

Let $G=(u, v)$. Under the new plan of testing we need two tests to classify the units u and v. Under the old plan of testing the following cases will require three tests to classify the units u and v, i.e., we have a saving of one test for the new plan when

(i) G is defective and the first unit to be tested after G under the old plan along the branch III is defective, and also when

(ii) G is mediocre and the first unit to be tested after G under the old plan along the branch II is mediocre.

In addition we would be using two tests under the new plan when

(iii) G is good whereas we would have used only one test under the old plan (i.e., we have a loss of one test). Combining the results from (i), (ii) and (iii) the expected number of tests saved under the new plan is $(1-q_1-q_2)+q_2(q_1+q_2)-q_1^2$. Hence we have a positive saving when

$$(1-q_2+q_2^2)-q_1(1-q_2)-q_1^2 > 0. \tag{5.2}$$

Case II. $x \geqq 3$

We now show that there is a saving of $x-2$ (or $x-1$) tests when all the $x-1$ units in G except u are good. Under the old plan we would perform $(x-1)$ individual tests to find that all the units in $G-u$ are good and none of these are required under the new plan. Under the old plan a test on the unit u may or may not be necessary but it is necessary under the new plan. Hence in this situation there is a saving of at least $(x-2)$ tests.

Moreover we shall save one test in the following situations:

(1) *Let u be a defective unit.*

(a) Let b denote the last unit of $G-u$ to be tested under the old plan when G is defective. It will also be the last unit of $G-u$ to be tested under the new plan if $G-u$ is defective. If all the units of $G-u$, except, b are good we will save one test by using inference in the new plan; this inference is not available to us under the old plan.

(b) Let c be the last element of $G-u$ to be tested when $G-u$ is mediocre. If c is mediocre and all the units in $G-u$ except c are good, then to classify all the units of G under the new plan we need x tests whereas under the old plan of testing we would have needed $(x+1)$ tests to classify the units of G. (The assumption that u is not the last unit tested under the old plan is used here.)

(2) *Let u be a mediocre unit.*

Let d be the last unit of $G-u$ to be tested under the old plan if G is mediocre. It will also be the last unit of $G-u$ to be tested under the new plan if $G-u$ is mediocre. If all the units in $G-u$, except d, are good we will save one test by using inference in the new plan; this inference is not available to us under the old plan.

Finally if G is good we use two tests under the new plan whereas we would have used only one test under the old plan; hence there is a loss of exactly one test in this case.

Combining all the above results we find that the expected number of tests saved (denoted by S) satisfies the inequality

$$S \geqq q_1^x \left(\frac{q_2^2+q_2q_3+q_3^2}{q_1^2} + (x-2)\frac{q_2+q_3}{q_1} - 1 \right) \tag{5.3}$$

Replacing q_3 by $1-q_1-q_2$ we find the right hand side of (5.3) is positive for all $x \geqq 3$ if

$$(1-q_2+q_2^2)-q_1(1-q_2)-q_2 > 0. \tag{5.4}$$

Therefore combining these two cases and noting that the inequalities in (5. 2) and (5. 4) are the same, it follows that for all x ($x \geqq 2$) there is a positive saving in the expected number of tests under the new procedure when (5. 4) holds. The inequality (5. 4) will be true whenever $q_1 < q_2^*$ where q_2^* is given by (5. 1).

Furthermore it is evident that the samples with the above mentioned cases will reach B. This proves statement (i) of the theorem.

When there are only two units, there are only two different procedures disregarding unreasonable procedures. Under the first procedure we test each unit individually and therefore we need two tests. Under the second procedure, to begin with, we test both units. If both units are good, we do not need any further test. If this set is mediocre (or defective) we test a single unit. We infer the nature of the second unit, if it is possible to do so, from the result of the test on the first unit; otherwise we test the second unit. The expected number of tests under the second procedure is easily computed to be

$$3 - q_1 - q_2 + q_2(q_1 + q_2) - q_1^2$$

and thus

$$\mathsf{E}(T) = \min \{2, 3 - q_1 - q_2 + q_2(q_1 + q_2) - q_1^2\}.$$

The second procedure is optimal if

$$1 - q_1 - q_2 + q_2(q_1 + q_2) - q_1^2 < 0 \tag{5. 5}$$

or equivalently $q_1 > q_2^*$ where q_2^* is given by (5. 1). Suppose $N = 2M$ is an even number and (5. 5) holds. We divide N units into M groups each of size 2 and use the optimal procedure mentioned above for each group of size 2. Under this scheme of testing, the expected number of tests to classify N units is

$$M[3 - q_1 - q_2 + q_2(q_1 + q_2) - q_1^2]. \tag{5. 6}$$

Now the quantity in square brackets in (5. 6) is less than 2 and the expression (5. 6) is less than $2M$ ($= N$). Since the expected number of tests under this procedure is less than N, so it must also be so under the optimal procedure. Likewise we can deal with the case $N = 2M + 1$ by dividing these units into $M + 1$ groups, M of which are of size 2 and a group containing a single unit. This proves statement (ii) of the theorem.

It is shown in the proof of statement (i) of the theorem that, for $x = 2$ the expected number of tests saved under the new plan is

$$(1 - q_1 - q_2) + q_2(q_1 + q_2) - q_1^2 \tag{5. 7}$$

and for $x \geqq 3$, the expected number of tests saved under the new plan is either

$$q_1^{x-2}\{(1 - q_1 - q_2) + q_2(q_1 + q_2) - q_1^2 + (x - 2)q_1(q_2 + q_3)\} \tag{5. 8}$$

or

$$q_1^{x-2}\{(1 - q_1 - q_2) + q_2(q_1 + q_2) - q_1^2 + (x - 3)q_1(q_2 + q_3)\}, \tag{5. 9}$$

depending on whether or not we can infer the result of one test. For $q_1 = q_2^*$, the value of (5. 7) which equals the first three terms in the braces is zero and thus the saving in the expected number of tests is non-negative. The expected number of tests under the new plan is not greater than that under the old plan. Hence introduc-

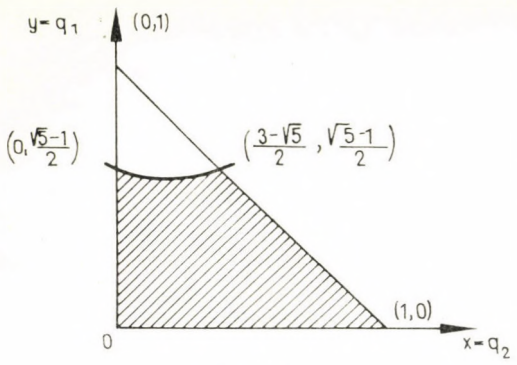


Figure 3

ing a sequence of new plans, each of which tests more units one at a time than the predecessor, it follows that the expected number of tests under the "one-at-a-time" procedure is not greater than the expected number of tests under the procedure arbitrarily chosen at the outset. This proves the statement (iii) of the theorem.

COROLLARY. For $q_1 < .6$, the optimal procedure tests one unit at a time.

PROOF. — Let $q_2 = x$, $q_1 = y$ so that $x + y \leqq 1$. The curve corresponding to the equality in (5. 4), given by

$$1 - x + x^2 + xy - y - y^2 = 0 \tag{5. 10}$$

represents a hyperbola. Here we are interested in the upper branch of the hyperbola as shown in the figure. Whenever the point $(q_2, q_1) = (x, y)$ lies in a shaded area, the inequality (5. 4) is satisfied and the optimal procedure is to test one unit at a time. By looking at the first and second derivatives of (5. 10), it is easy to see that this branch of hyperbols has minimum at $\left(\frac{1}{5}, \frac{3}{5}\right)$. Since the minimum is attained at $\left(\frac{1}{5}, \frac{3}{5}\right)$ we are led to the conclusion that when $q_1 < .6$, no matter what the value of q_2 (subject only to $q_1 + q_2 \leqq 1$), the optimal procedure for classifying all the units is to test each unit individually.

6. Comparison of the Procedure R_0 with R_1

In this section we compare the expected number of group tests under the procedure R_0 and R_1 for the classification of N units with known $\vec{q} = (q_1, q_2, q_3)$. Let $\mathsf{E}(T|R, N, \vec{q})$ denote the expected number of group tests to be performed if the procedure R is used for N units and for given $\vec{q} = (q_1, q_2, q_3)$. The procedure R is said to be equivalent to the procedure R' for $\vec{q} = (q_1, q_2, q_3)$ if $\mathsf{E}(T|R, N, \vec{q}) = \mathsf{E}(T|R', N, \vec{q})$ for every N.

We shall use $H(0; N)$ for $\mathsf{E}(T|R_1, N, \vec{q})$ where $H(0; N)$ is defined in (2. 10); also q_2^* is the same as defined in (5. 1).

THEOREM. *For $\vec{q} = (q_1, q_2, q_3)$ with $q_1 \leqq q_2^*$ the procedure R_0 is equivalent to the procedure R_1.*

PROOF. For $q_1 \leqq q_2^*$, $\mathsf{E}(T|R_0, N, \vec{q}) = N$ by the theorem in section 5 and, since R_0 is the optimal procedure among all procedures, we have

$$\mathsf{E}(T|R_0, N, \vec{q}) = N \leqq H(0; N). \tag{6. 1}$$

For any $\vec{q}=(q_1, q_2, q_3)$ and, in particular, for any $\vec{q}$ with $q_1 \leqq q_2^*$, we have

$$H(0; N) = 1 + \min_{1 \leqq x \leqq N} \{q_1^x H(0; N-x) + (q_{[2]}^x - q_1^x) G(x, x; 0, N-x) + (1-q_{[2]}^x), G(0, 0; x, N)\}$$

$$\leqq 1 + q_1 H(0;\ N-1) + (q_{[2]} - q_1) G(1, 1;\ 0, N-1) + (1-q_{[2]}) G(0, 0;\ 1, N)$$

$$= 1 + H(0;\ N-1).$$

By substituting $N=1, 2, \ldots$ we get

$$H(0;\ N) \leqq N. \tag{6.2}$$

Combining (6. 1) and (6. 2) we find that for any $\vec{q}$ with $q_1 \leqq q_2^*$

$$H(0;\ N) = \mathsf{E}(T|R_0, N, \vec{q})$$

This proves the theorem.

7. Lower Bound for any Group Testing Procedure from Information Theory

Let $H(N|R)$ be the expected number of group tests needed to classify N units under a procedure R for given $\vec{q}=(q_1, q_2, q_3)$.

THEOREM.

$$H(N|R) \geqq -N[q_1 \log_3 q_1 + q_2 \log_3 q_2 + q_3 \log_3 q_3]$$

PROOF. The total reduction in entropy associated with the classification of N unit where each unit is assumed to represent an idependent observation from a trinomial population with parameter $\vec{q}=(q_1, q_2, q_3)$ is given by

$$I = -N\left(\sum_{i=1}^{3} q_i \log_2 q_i\right). \tag{7.1}$$

The expected number of tests, in which the total reduction in entropy is carried out, is $H(N|R)$. The reduction in entropy associated with each test is at most $\log_2 3$, thus we have for any procedure R and any $\vec{q}=(q_1, q_2, q_3)$

$$H(N|R) \log_2 3 \geqq -N(q_1 \log_2 q_1 + q_2 \log_2 q_2 + q_3 \log_2 q_3)$$

or

$$H(N|R) \geqq -N(q_1 \log_3 q_1 + q_2 \log_3 q_2 + q_3 \log_3 q_3).$$

These lower bounds have been calculated for $\vec{q}=(.90, .05, .05)$ and $N \leqq 8$ and are given in Table II.

8. Lower Bound for any Group Testing Procedure from Coding Theory

HUFFMAN has given a procedure for the construction of compact codes. Using his results we can obtain a lower bound for any q and for any group-testing procedure.

Let the set of symbols comprising a given alphabet be called $S=\{s_1, s_2, \ldots, s_q\}$.

Then we define a code as a mapping of all possible sequences of symbols of S into sequences of symbols of some other alphabet $X=\{x_1, x_2, \ldots, x_r\}$. S is called the source alphabet and X the code alphabet. A compact code for a source S is a code which has the smallest average word length if we encode the symbols from S one at a time.

At the outset there are N units, each of which is good, mediocre or defective. Thus there are 3^N possible states of nature, one of which is true. If we represent each test that gives a good outcome, a mediocre outcome and a defective outcome by the digits zero, one and two respectively, then a procedure is identical with a 3-ary code. Thus a particular set of outcomes (i.e., a particular path in the tree of possible paths) corresponds in a one-to-one manner with a particular "word" of the code. The expected number of tests required is equal to the expected word length of the code.

For example, letting S, T and U denote good, mediocre and defective respectively, we consider two different codes corresponding to two group testing procedures for $N=2$ units.

State of Nature	*Probability*	*Code I*	*Code II*
SS	q_1^2	00	0
ST	q_1q_2	01	10
TS	q_1q_2	10	110
SU	q_1q_3	02	20
US	q_1q_3	20	220
TT	q_2^2	11	111
TU	q_2q_3	12	21
UT	q_2q_3	21	221
UU	q_3^2	22	222

Code I corresponds to the procedure in which each unit is tested individually; code II corresponds to the procedure in which the first test is on both units and subsequent tests are on each individual unit.

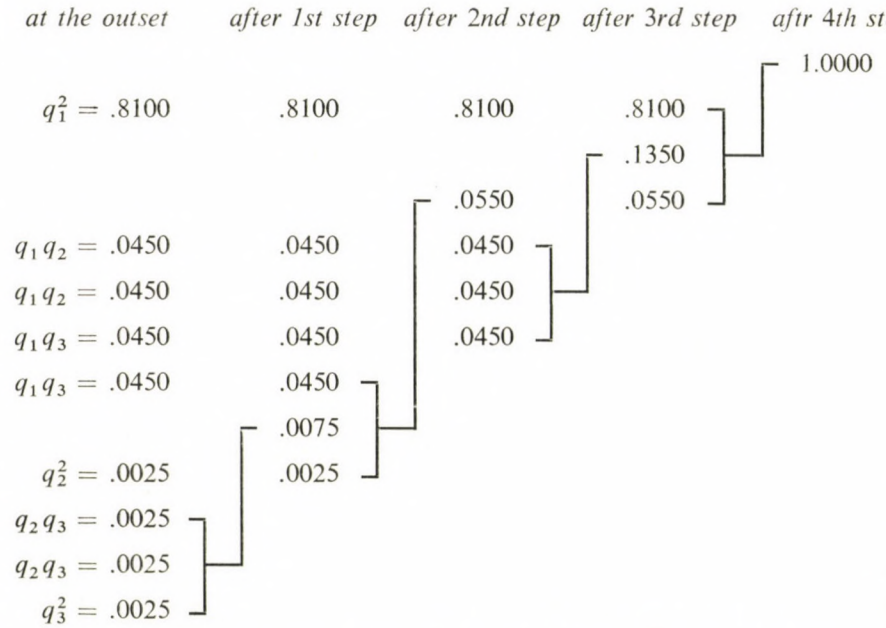
The expected word length of the code corresponding to any group testing procedure is clearly greater than or equal to the expected word length of the optimal Huffman code for encoding the 3^N words in the source with known probabilities. Thus the expected word length of the Huffman code is a lower bound on the expected number of tests for any group-testing procedure.

HUFFMAN has given a routine for finding the expected word length of Huffman codes. To describe the computation of the expected word length of Huffman code, let Q_i $(i=1, 2, \ldots, I)$ denote any set of a priori probabilities that sum to one; in our problem of group-testing these a priori probabilities are of a special trinomial structure $q_1^i q_2^j q_3^k$ where i, j, k are non-negative and sum to N. At the 1st step we order the Q_i's and add the three smallest and call the sum S_1; at the 2nd step we reorder the remaining set of 3^N-2 probabilities and again add the three smallest, calling the sum S_2, This is repeated until the odd number of probabilities remaining reduces to 1. Let S_j denote the sum of three smallest probabilities at the j^{th} step $(j=1, 2, \ldots, J)$. It is easy to verify that $J=\frac{1}{2}(3^N-1)$ and $S_J=1$. Then the Huff-

man lower bound (HLB) which depends on $\vec{q}$ and N is given by

$$\text{HLB} = \sum_{j=1}^{J} S_j$$

This method is illustrated below by an example with $\vec{q}_0 = (.90, .05, .05)$ and $N=2$:



The value of HLB in the above example is 1.1975 and ILB (Information theory lower bound) is .718 for $N=2$. It is easy to see that HLB does not correspond to a group test in this case. The optimal group-test in this case is to start with testing two units at the outset and the expected number of group-tests under the optimal procedure (as well as for procedure R_1) is then easily computed to be 1.2875.

The number of digits (shown below) in a Huffman code is the number of combinations indicated in the diagram above for the corresponding probability.

State of Nature	*Probability*	*Length of Huffman code word*
SS	.8100	1
ST	.0450	2
TS	.0450	2
SU	.0450	2
US	.0450	2
TT	.0025	2
TU	.0025	3
UT	.0025	3
UU	.0025	3

It is easily observed that the lengths of the Huffman code words (having the same probabilities for the alphabet in the source as the states of nature in our problem) are different from the corresponding word lengths of both code I and code II; the latter coresponds to the optimal group-testing procedure for this problem. Hence the HLB is not attained by a group-testing procedure in this case.

TABLE I

Values of $x_H(n; x; \vec{q}_0)$ and $x_G(m, n; d; e; \vec{q}_0)$ for $\vec{q}_0=(.90, .05, .05)$ for procedure R_1 for various H- and G-situations arising in the classification of $N(\leqq 8)$ units under R_1. *

$$x_H(n; \underset{\uparrow}{e}; \vec{q}_0) = e \quad \text{for } e \geqq 1 \text{ and } n \leqq N;$$

$$x_H(\underset{\uparrow}{n}; 0, \vec{q}_0) = n \quad \text{for } 1 \leqq n \leqq N;$$

$$x_G(\underset{\uparrow}{m}, n; d, e; \vec{q}_0) = \begin{cases} 1 & \text{for } m=2, 3 \\ 2 & \text{for } m=4, 5, 6 \\ 3 & \text{for } m=7 \\ 4 & \text{for } m=8; \end{cases}$$

$$x_G(0, n; \underset{\uparrow}{d}, e, \vec{q}_0) = \begin{cases} 1 & \text{for } d=2, 3 \\ 2 & \text{for } d=4, 5, 6 \\ 3 & \text{for } d=7, 8. \end{cases}$$

TABLE II

Comparison of the Expected Number of Tests for Procedure R_1 and Information Theory Lower Bounds for Any Procedure starting with a Trinomial set of size N with $\vec{q}_0=(.90, .05, .05)$

N	H(0; N)	Information Theory Lower Bound for Any Procedure
1	1.000	0.359
2	1.288	0.718
3	1.654	1.077
4	2.034	1.436
5	2.464	1.795
6	2.905	2.154
7	3.357	2.513
8	3.820	2.872

Acknowledgment. I wish to express my sincere gratitude to Professor M. SOBEL for proposing this investigation and constant encouragement and many valuable suggestions during its progress.

*The vertical arrow indicates what set the x's come from.

REFERENCES

[1] CARTER, F. L., JR.: Group testing in binomial and multinomial situations. *Technical Report* No. 3, Dept. of Statistics, V. P. I., Blacksburg, Va, 1960.

[2] DORFMAN, R.: The detection of defective members of large populations, *Ann. Math. Statist.* **14** (1943), 436—440.

[3] FELLER, W.: *An Introduction to Probability Theory and its Applications,* Second Edition. John Wiley and Sons, New York, 1957.

[4] HUFFMAN, D. A.: A method for the construction of minimum redundancy codes, *Proc. I. R. E.* **40** (1952), 1098—1101.

[5] KUMAR, S.: A Group testing problem, *Ann. Math. Statist.* **36** (1965), 727—728, 1318.

[6] SOBEL, M. and GROLL, P. A.: Group testing to eliminate efficiently all defectives in a binomial sample, *Bell System Tech. J.* **38** (1959), 1179—1252.

[7] SOBEL, M.: Group testing to classify all defectives in a binomial sample, A contribution in *Information and Decision Processes,* edited by Machol, R. E., McGraw Hill, 1960. p. 127—161.

[8] SOBEL, M.: Optimal group testing, *Technical Report* No. 72, Stanford Univ., Stanford, Calif., 1964.

[9] GROLL, P. A. and SOBEL, M.: Binomial group testing with an unknown number of defectives, *Technometrics* **8** (1966), 631—656.

[10] STERRETT, A.: On the detection of defective members of large populations, *Ann. Math. Statist.* **28** (1957), 1033—1036.

[11] UNGAR, P.: The cut off point for group testing, *Communications Pure Applied Math.* **13** (1960), 49—54.

University of Wisconsin — Milwaukee

(Received June 20, 1967.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 249—254.

PREDICTION OF VARIANCE IN TWO-STAGE SAMPLING DESIGNS

by

G. BAIKUNTH NATH[1] and V. P. GUPTA

Summary

An estimation procedure has been considered which gives a better estimate than the usual unbiased estimate (in the sense of smaller variance) for two-stage sampling. A test is required to find whether the effect due to units is negligible.

1. Introduction

The technique of two-stage sampling has been widely used in chemical, physical and biological analysis. The total population is divided into a number of units, each of these units is further divided into a number of subunits. A random sample of n' units and m' subunits from each of the selected unit is taken. The variance of the sample mean is often required.

If x_{ij} denote the observation in the j-th subunit of the i-th unit, and the effects of units and subunits are additive, then it may be represented by the model:

$$x_{ij} = \mu + v_i + \delta_{ij}, \tag{1.1}$$

where $i = 1, 2, 3, \ldots, n'$; $j = 1, 2, 3, \ldots, m'$. The term μ represents the population mean (a fixed constant), v_i and δ_{ij} represent effects due to units and subunits within units respectively. The v_i' s and δ_{ij}' s are stochastically independent random variables with zero means and variances σ_u^2 and σ_w^2 respectively. The variance of the sample mean $\bar{x}$ is, then, given by

$$V(\bar{x}) = \sigma_u^2/n' + \sigma_w^2/n'm'. \tag{1.2}$$

To estimate $V(\bar{x})$, we need estimates of σ_u^2 and σ_w^2 which may be obtained from the following analysis of variance given by Cochran [1953]:

TABLE 1

Analysis of variance of the experiment

Source	Degrees of Freedom	Mean Square	Expected Mean square
Between units	$m_1 = n-1$	s_b^2	$\sigma_b^2 = \sigma_w^2 + m\sigma_u^2$
Within units (between subunits)	$m_2 = n(m-1)$	s_w^2	$\sigma_w^2 = \sigma_w^2$

[1] Presently at Department of Mathematics, Univ. of Queensland, St. Lucia, 4067, Australia

Thus the unbiased estimate of $V(\bar{x})$ is

$$v(\bar{x}) = \{s_b^2 + (m-m')s_w^2/m'\}/n'm. \tag{1.3}$$

In many experimental situations we may have $\sigma_u^2 = 0$. Then the appropriate model for x_{ij} will be

$$x_{ij} = \mu + \delta_{ij}, \tag{1.4}$$

and the pooled estimate $(m_1 s_b^2 + m_2 s_w^2)/(m_1 + m_2)n'm'$ may be used for $V(\bar{x})$. Therefore, before taking (1. 1) or (1. 4) as a model for x_{ij}, we perform a preliminary test of significance as suggested by BANCROFT [1] to test the hypothesis that $\sigma_u^2 = 0$ against $\sigma_u^2 > 0$. These considerations lead us to the following procedure for estimating $V(\bar{x})$. If

$$s_b^2/s_w^2 < \beta, \quad \text{use} \quad V = (m_1 s_b^2 + m_2 s_w^2)/(m_1 + m_2)n'm'; \tag{1.5}$$

or

$$s_b^2/s_w^2 \geqq \beta, \quad \text{use} \quad V = s_b^2/n'm + (m-m')s_w^2/n'm'm; \tag{1.6}$$

where $\beta = F(m_1, m_2; \alpha)$ refers to the upper $100\alpha\%$ point of the F-distribution with m_1 and m_2 degrees of freedom.

The object of the present investigation is to study the bias and relative efficiency of the estimate thus obtained.

2. Expected value and mean square error of V

The mean squares s_b^2 and s_w^2 are distributed as $\chi_1^2\sigma_b^2/m_1$ and $\chi_2^2\sigma_w^2/m_2$ respectively, where χ_i^2 is central chi-square statistics with m_i $(i = 1, 2)$ degrees of freedom. Their joint distribution is given by

$$g(s_b^2, s_w^2) = K s_b^{m_1-2} s_w^{m_2-2} \exp\{-\tfrac{1}{2}(m_1 s_b^2/\sigma_b^2 + m_2 s_w^2/\sigma_w^2)\}\, d(s_b^2) d(s_w^2), \tag{2.1}$$

where K is constant.

Using the transformation $z_1 = s_b^2/s_w^2$, $z_2 = s_b^2$ and $z_3 = s_w^2$, we get

$$g_1(z_1, z_2) = K z_1^{-\frac{1}{2}m_2 - 1} z_2^{\frac{1}{2}m_{12} - 1} \exp\{-z_2(m_1 z_1/\sigma_b^2 + m_2/\sigma_w^2)/2z_1\}\, dz_1\, dz_2, \tag{2.2}$$

and

$$g_2(z_1, z_3) = K z_1^{\frac{1}{2}m_1 - 1} z_3^{\frac{1}{2}m_{12} - 1} \exp\{-z_3(m_1 z_1/\sigma_b^2 + m_2/\sigma_w^2)/2\}\, dz_1\, dz_3, \tag{2.3}$$

where

$$K = (m_1/\sigma_b^2)^{\frac{1}{2}m_1} (m_2/\sigma_w^2)^{\frac{1}{2}m_2} / 2^{\frac{1}{2}m_{12}} \Gamma(\tfrac{1}{2}m_1)\Gamma(\tfrac{1}{2}m_2),$$

and $m_{12} = m_1 + m_2$.

Let E_1 and E_2 denote the expected values of V under (1. 5) and (1. 6) respectively. Then we have

$$\mathsf{E}_1 = \frac{m_1}{m'n'm_{12}} \int_0^\infty \int_0^\beta z_2 g_1(z_1, z_2)\, dz_2\, dz_1 + \frac{m_2}{m'n'm_{12}} \int_0^\infty \int_0^\beta z_3 g_2(z_1, z_3)\, dz_3\, dz_1, \tag{2.4}$$

and

(2.5)

$$\mathsf{E}_2 = \frac{1}{mn'} \int_0^{\infty} \int_{\beta}^{\infty} z_2 g_1(z_1, z_2)\, dz_2\, dz_1 + \frac{1}{n'}\left(\frac{1}{m'} - \frac{1}{m}\right) \int_0^{\infty} \int_{\beta}^{\infty} z_3 g_2(z_1, z_3)\, dz_3\, dz_1 .$$

Solving (2.4) and (2.5) and adding, we get

(2.6)

$$\mathsf{E}(V) = V(\bar{x}) + (m'm_{12} - mm_1)\{\Phi I_{x_0}(\tfrac{1}{2}m_1, \tfrac{1}{2}m_2 + 1) - I_{x_0}(\tfrac{1}{2}m_1 + 1, \tfrac{1}{2}m_2)\}\sigma_b^2/n'm'mm_{12},$$

where $\Phi = \sigma_w^2/\sigma_b^2$, $x_0 = m_1\Phi\beta/(m_1\Phi\beta + m_2)$ and

$$I_{x_0}(p, q) = \frac{1}{B(p, q)} \int_0^{x_0} x^{p-1}(1-x)^{q-1}\, dx.$$

Hence the bias of V is given by

(2.7)

$$\text{Bias}\,(V) = (m'm_{12} - mm_1)\{\Phi I_{x_0}(\tfrac{1}{2}m_1, \tfrac{1}{2}m_2 + 1) - I_{x_0}(\tfrac{1}{2}m_1 + 1, \tfrac{1}{2}m_2)\}\sigma_b^2/n'm'mm_{12}.$$

It can easily be seen that

i) Bias (V) is positive, zero or negative according as

$$I_{x_0}(\tfrac{1}{2}m_1 + 1, \tfrac{1}{2}m_2)/I_{x_0}(\tfrac{1}{2}m_1, \tfrac{1}{2}m_2 + 1)$$

is less than, equal to, or greater than Φ,

ii) for a given set of values of n', m', n, m and Φ the Bias (V) is maximum at $\beta = 1$,

iii) for the sample $n' = kn$, the Bias (V) is $1/k$ times its value for $n' = n$,

iv) for $\beta = 0$, Bias $(V) = 0$ and hence V is an unbiased estimate of $V(\bar{x})$,

v) for $\beta = \infty$, i.e., assuming $\sigma_u^2 = 0$ always, we have

$$\mathsf{E}(V) = (m_1\,\sigma_b^2 + m_2\,\sigma_w^2)/m'n'm_{12}.$$

To calculate the mean square error of V, we need the expression for $\mathsf{E}(V^2)$ which can be derived along the same lines as for $\mathsf{E}(V)$, and then to use the relation $\text{MSE}\,(V) = \text{Var}\,(V) + \{\text{Bias}\,(V)\}^2$, where $\text{Var}\,(V) = \mathsf{E}(V^2) - \{\mathsf{E}(V)\}^2$. The final expression for MSE (V) is given by

$$\begin{aligned}\text{MSE}(V) = \sigma_b^4\big[&(2/m_1m_2)\{m_2m'^2 + m_1(m-m')^2\Phi^2\} + (1 + 2/m_1)(m^2m_1^2 - m'^2)\times \\ &I_{x_0}(\tfrac{1}{2}m_1 + 2, \tfrac{1}{2}m_2) + (1 + 2/m_2)\Phi^2\{m^2m_2^2 - (m-m')^2\}I_{x_0}(\tfrac{1}{2}m_1, \tfrac{1}{2}m_2 + 2) + \\ &(2\Phi/m_{12}^2)\{m^2m_1m_2 + m'm_{12}^2(m-m')^2\}I_{x_0}(\tfrac{1}{2}m_1 + 1, \tfrac{1}{2}m_2)I_{x_0}(\tfrac{1}{2}m_1, \tfrac{1}{2}m_2 + 1) - \\ &(2/m_{12})\{m'(mm_1 - m'm_{12}) + mm_1(m-m')\Phi\}I_{x_0}(\tfrac{1}{2}m_1 + 1, \tfrac{1}{2}m_2) - \\ &(2\Phi/m_{12})\{mm'm_2 + (m-m')(m'm_{12} - mm_1)\Phi\}I_{x_0}(\tfrac{1}{2}m_1, \tfrac{1}{2}m_2 + 1)\big]/(n'm'm)^2\end{aligned} \tag{2.8}$$

or

$$= \text{MSE}\,V(\bar{x}) + \psi(n', m', n, m, \Phi).$$

The quantity ψ thus represents reduction or increase in MSE (V) due to pooling or not pooling the mean squares s_b^2 and s_w^2 subsequent to the preliminary test of significance. For $\beta=0$, $\psi=0$ and hence MSE $(V)=$ MSE $V(\bar{x})$. For $\beta=\infty$, we have

$$\text{MSE}\,(V) = 2(m_1\sigma_b^4+m_2\sigma_w^4)/(n'm'm_{12})^2+\{\text{Bias}\,(V)\}^2.$$

3. Discussion of results

In this section we discuss the results of bias and mean square error of V. From table 1 we may note that $m_1<m_2$, except for the trivial case $m=n=1$. It may be of interest to examine how the variance of the sample mean is affected by increasing the number of units to the k-th multiple of n as given by the anova table, since an increase in the value of n diminishes the contributions from both the variances σ_u^2 and σ_w^2. We may observe from (2. 6) and (2. 8) that Bias (V) and MSE (V) will depend upon the level of significance α, the selection of which is at our disposal. For an empirical study of the behaviour we have made computations for the set $n=n'=10$, $m=m'=3$ corresponding to various values of α. The curves for Bias (V) at various levels of significance are shown in Fig. 1.

The bias decreases monotonically as Φ increases for $\alpha=0$. It is positive and maximum for a given value of Φ at $\beta=1$, and increases as Φ increases. When $\alpha=.25$ it fluctuates in the neighbourhood of zero for $.1\leqq\Phi\leqq.5$ and then

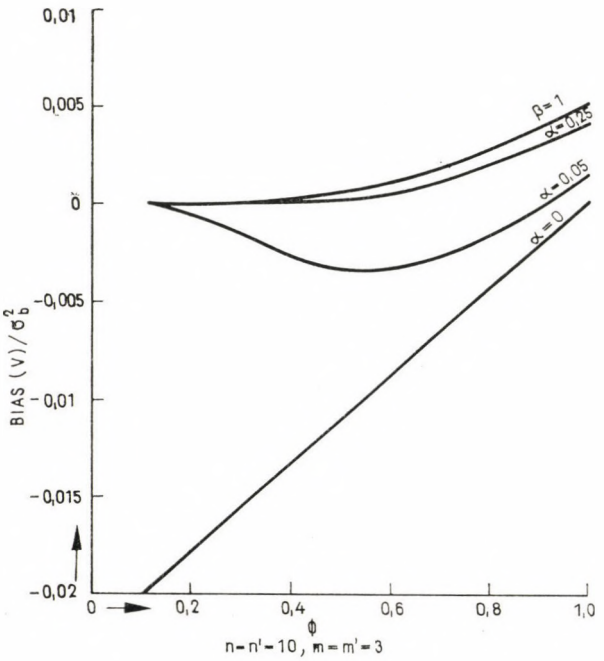


Fig. 1

increases in magnitude as Φ increases for $\Phi > .5$. However, the bias in magnitude is smaller almost everywhere for $\alpha = .25$ than its value at other levels of significance. The estimator V is unbiased for $\alpha = 1$.

In general, V is a biased estimator of $V(\bar{x})$. It is, therefore, desirable to examine the relative efficiency of V to $V(\bar{x})$ defined by

$$\text{R. E.} = \frac{2\{m_2 m'^2 \sigma_b^4 + m_1 (m - m')^2 \sigma_w^4\}/(n' m' m)^2 m_1 m_2}{\text{MSE}(V)} \cdot 100\%$$

The curves for relative efficiency have been shown in Fig. 2. It can easily be seen that the relative efficiency is invariant for any initial sample size n'. When $\alpha = 0$

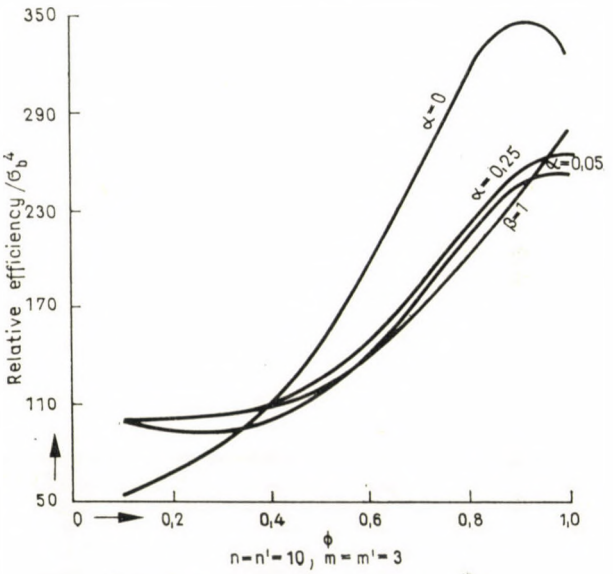


Fig. 2

the estimator V is most efficient for $.9 \leqq \Phi < 1.0$. In the case $\alpha = .01$ or $.05$ relative efficiency first decreases and then increases as Φ increases. For $\alpha \geqq .25$, the relative efficiency increases with Φ and the estimator V is more efficient than $V(\bar{x})$ almost everywhere.

To summarize, an indiscriminate use of $\alpha = 0$ or $\alpha = 1$ is not desirable; former is to be used when Φ is close to unity and the latter when Φ is very small. For intermediate values of Φ or when no a priori information on the magnitude of Φ is available, a value of $\alpha \geqq .25$ is most suitable.

4. Illustration

Ten thousand random normal deviates from Rand Corporation [1955] were punched on one thousand cards of eighty columns each. On each card rows numbered 0 through 9 were considered. A random sample of 10 cards was drawn and each of the selected card was subsampled for 3 rows. The punches in each selected row were counted. The analysis of variance of the sample is given in table 2.

TABLE 2

Analysis of variance of the sample

Source	Degrees of Freedom	Mean Square	Expected mean Square
Between cards	9	2.922	$\sigma_w^2 + 3\sigma_u^2$
Within cards (between rows)	20	2.915	σ_w^2

To test the hypothesis that $\sigma_u^2 = 0$ against $\sigma_u^2 > 0$, we compute the F-ratio $2.922/2.915 = 1.002$ with (9, 20) degrees of freedom, which is non-significant at $\alpha = .25$. Thus a better estimate of the variance of the number of punches in a row, i.e., $V(\bar{x})$ given by situation (1. 5) is .0972. This estimate, though slightly biased, is more efficient than the usual estimate .0974 given by (1. 3).

REFERENCES

[1] BANCROFT, T. A.; On biases in estimation due to the use of preliminary tests of significance, *Ann. Math. Statist.*, **15** (1944), 190.
[2] COCHRAN, W. G.: *Sampling Techniques*, John Wiley & Sons, New York, 1953.
[3] *A Million Random Digits with* 100,000 *Normal Deviates*, Rand Corporation, The Free Press, Illionis, 1955.

Centre for Advanced Study in Mathematics, Panjab University, Chandigarh, India

(Received August 29, 1968.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 255—265.

RANDOM MODELS OF LOGICAL SYSTEMS

by

I. RUZSA

Part II[1]

Models of Some Modal Logics

§ 4. Models of Łukasiewicz's modal systems

We shall apply the results of § 3 (especially *Th.* 3. 4) to the three- and the four-valued logical systems of ŁUKASIEWICZ. Let us denote here these systems by *Ł*3 and *Ł*4, respectively.

In *Ł*3, [2] the set O of the operators (see *Def.* 1 in Part I) is

$$O = O_1 \cup O_2 = \{\neg\} \cup \{\rightarrow\},$$

and the set T of the theorems is determined by a matrix

$$\mathbf{M3} = (C, D, \Psi)$$

which is *per definitionem* adequate to *Ł*3 (see *Def.* 8—10 in Part I); here $C = \{0, {}^1/_2, 1\}$, $D = \{1\}$; and the functions $\Psi(\neg) = \neg^*$, $\Psi(\rightarrow) = \overset{*}{\rightarrow}$ are defined by the following table:

$\overset{*}{\rightarrow}$	0	$\frac{1}{2}$	1	$\neg^*$
0	1	1	1	1
$\frac{1}{2}$	$\frac{1}{2}$	1	1	$\frac{1}{2}$
1	0	$\frac{1}{2}$	1	0

If we define the function operators $\dot{\neg}$ and $\dot{\rightarrow}$ as follows:

$$\dot{\neg} w = 1 - w,$$

$$w_1 \dot{\rightarrow} w_2 \equiv \begin{cases} \overset{*}{\rightarrow}(w_1, w_2), \text{ if } w_1 \text{ and } w_2 \text{ are discrete with respect to } C, \\ 1 \text{ otherwise} \end{cases}$$

then the random model structure $Z_3 = (0, 1, \theta)$ (where $\theta(\neg) = \dot{\neg}$, $\theta(\rightarrow) = \dot{\rightarrow}$) is *characteristic* to **M3** (see *Def.* 2, 3, 12 in Part I), and according to *Def.* 14—15 and *Th.* 3. 4 (in Part I), Z_3 is a perfect discrete representation of *Ł*3 in $\langle 0, {}^1/_2, 1\rangle$.

Let us consider now the more interesting system *Ł*4[3]. In this system

$$O = O_1 \cup O_2 = \{\neg, M\} \cup \{\rightarrow\},$$

---

[1] Part I ("Models of Valuing Logics") appeared in this *Studia*, **4** (1969), 301—312. The present paper is not self-contained. The notation is explained in Part I. Also definitions and theorems of Part I often will be quoted here.

[2] See e.g. [3]. On the axiomatization of *Ł*3 see [4] or [7, III. 2].

[3] See [5] or [6].

and T is determined by the matrix $\mathbf{M4}=(C, D, \Psi)$, where $C=\{{}^*0, 0, 1, 1^*\}$, $D=\{1^*\}$, and the functions $\Psi(\neg)=\neg^*$, $\Psi(M)=M^*$, $\Psi(\rightarrow)=\overset{*}{\rightarrow}$ are defined by the following table:

(1)

$\overset{*}{\rightarrow}$	^{*0}	0	1	1^*	$\neg^*$	M^*
^{*0}	1^*	1^*	1^*	1^*	1^*	0
0	1	1^*	1	1^*	1	0
1	0	0	1^*	1^*	0	1^*
1^*	^{*0}	0	1	1^*	^{*0}	1^*

REMARKS. (i) If p is a statement, then the meaning of Mp is "it is *possible* that p". — (ii) Let us agree that ${}^*0=-\varepsilon$, and $1^*=1+\varepsilon$, where ε is any fixed positive real number. Then the elements of C are numbers, and $\langle {}^*0, 0, 1, 1^*\rangle$ is a strictly increasing sequence. (In the original matrix of Łukasiewicz, the signs of the values are 4, 3, 2, and 1 instead of ^{*0}, 0, 1, and 1^*, respectively.) — (iii) In Ł4, $\wedge$, $\vee$, and $\leftrightarrow$ are defined operators with the usual definitions $[p\wedge q=\neg(p\rightarrow\neg q)$, $p\vee q=$ $=\neg p\rightarrow q$, $p\leftrightarrow q=(p\rightarrow q)\wedge(q\rightarrow p)]$. Furthermore, L is a one-placed operator defined by $Lp=\neg M\neg p$. The meaning of Lp is "it is *necessary* that p". M and L are called *modal operators*.

Let us define the function operators $\dot{\neg}$, $\dot{M}$ and $\dot{-}$ as follows:

$$(\dot{\neg}\, w)(E)=1-w(E); \tag{2}$$

$$\dot{M}w(E)=\begin{cases}0, & \text{if } w(E)\leqq 0,\\ 1^* & \text{otherwise;}\end{cases} \tag{3}$$

$$w_1\dot{-}w_2(E)=\begin{cases}1^*, & \text{if } w_1(E)=w_2(E)=0, \quad \text{or} \quad w_1(E)=w_2(E)=1,\\ \max\big(1-w_1(E), w_2(E)\big) & \text{otherwise.}\end{cases} \tag{4}$$

If w, w_1, w_2 are random variables on a probability space $\sum=(I, \mathscr{A}, P)$ with values in $[{}^*0, 1^*]$, then so are $\dot{\neg}w$, $\dot{M}w$ and $w_1\dot{-}w_2$. The truth of this statement is immediate for $\dot{\neg}$ and $\dot{M}$. As far as $\dot{-}$ is concerned, the measurability of $w_1\dot{-}w_2$ follows immediately from the identity

(5)

$$[w_1\dot{-}w_2<z]=[w_1>1-z]\cap[w_2<z]\cap\overline{[w_1=0]\cap[w_2=0]}\cap\overline{[w_1=1]\cap[w_2=1]}$$

(provided $z\leqq 1^*$), and (5) follows easily from (4).

Thus, if $\theta(\neg)=\dot{\neg}$, $\theta(M)=\dot{M}$, and $\theta(\rightarrow)=\dot{-}$, then

$$Z_Ł=({}^*0, 1^*, \theta)$$

is an *rms* of Ł4′.

Furthermore, if w, w_1, w_2 are discrete with respect to C, then $\dot{\neg}w=\neg^*(w)$, $\dot{M}w=M^*(w)$, and $w_1\dot{-}w_2=\overset{*}{\rightarrow}(w_1, w_2)$, as one can verify this fact by comparing table (1) and definitions (2), (3), (4). Thus $Z_Ł$ is *characteristic* of **M4**. By *Th.* 3. 4 (Part I) it then follows:

THEOREM 4. 1. *The rms $Z_Ł$ is a perfect discrete representation of the system Ł4 in $\langle {}^*0, 0, 1, 1^*\rangle$.*

System Ł4 contains system A_0, in the usual sense (any formula of A_0 is a formula of Ł4, and any theorem of A_0 is a theorem of Ł4). Although we cannot say that *rms* $Z_Ł$ contains *rms* Z_0 (in the sense that any random model belonging to Z_0, also belongs to $Z_Ł$), yet $Z_Ł$ is, in a certain sense, an extension of Z_0. Namely, the base interval of $Z_Ł$ contains the base interval of Z_0, the function operator $\dot{\neg}$ in both *rms*'s is the same, and for $z \leqq 1$ the identity

$$[w_1 \dot{-} w_2 < z] = [w_1 > 1 - z] \cap [w_2 < z] \qquad (z \leqq 1) \tag{6}$$

holds in both *rms*'s. (In $Z_Ł$, (6) follows easily from (5). In Z_0, $[w_1 \dot{-} w_2 < z] = [\dot{\neg} w_1 \dot{\vee} w_2 < z] = [\dot{\neg} w_1 < z] \cap [w_2 < z] = [w_1 > 1 - z] \cap [w_2 < z]$, using (12) and (7) of § 2, Part I). Thus $Z_Ł$ is an *rms* which is non-trivially characteristic of **M4**, and is a fairly natural extension of Z_0.

We have seen in Part I that the *rms* Z_0 is a perfect discrete representation of A_0 in $\langle 0, 1\rangle$, and an adequate total $^1/_2$-representation of A_0 (*Th.* 3. 5 and *Th.* 2. 2). Having got these results and *Th.* 4. 1, the question arises: Does there exist a real number d greater than ^{*0} for which $Z_Ł$ is an adequate total d-representation of Ł4? The following theorem gives a partial answer to this question.

THEOREM 4. 2. *For no positive d is the rms* $Z_Ł$ *an adequate total d-representation of the system* Ł4.

PROOF. Consider the formula

$$\alpha = Mp \to (M \neg p \to Mq)$$

(where p and q are variables) which is a theorem of Ł4. We give a random model Γ_0 belonging to $Z_Ł$ in which $W(\alpha) = 0$. By giving such a Γ_0, our theorem will be proved.

The construction of Γ_0 is as follows: Let $\sum$ be arbitrary. Put $W(p) = {}^1/_2$ and $W(q) = 0$; for other variables let W be arbitrary. Then we have (according to (2) and (3)) in $\Gamma_0 = (\Sigma, W)$ the following equalities:

$$W(\neg p) = 1/2, \quad W(Mp) = 1^*, \quad W(M \neg p) = 1^*, \quad W(Mq) = 0.$$

From this we get that $W\big(Mp \to (M \neg p \to Mq)\big) = 1^* \dot{-} (1^* \dot{-} 0)$. Since 1^* and 0 are discrete random variables with respect to the sequence $\langle ^*0, 0, 1, 1^*\rangle$, and $Z_Ł$ is characteristic of **M4**, we can apply the matrix **M4** to compute $1^* \dot{-} (1^* \dot{-} 0)$:

$$1^* \dot{-} 0 = 0, \quad 1^* \dot{-} (1^* \dot{-} 0) = 1^* \dot{-} 0 = 0.$$

Thus we have $W\big(Mp \to (M \neg p \to Mq)\big) = 0$.

REMARK. I can not answer the question whether there exists a nonpositive d for which $Z_Ł$ is an adequate total d-representation of Ł4.

The negative result just obtained yields us a starting point to investigate certain fragments of Ł4 having total d-representations in $Z_Ł$ for certain positive d. This is the aim of the next §.

§ 5. Models of some subsystems of Ł4

We introduce two subsystems of Ł4, denoted by Ł1 and Ł1*. The common language of both systems is Ł4′. We denote the set of the theorems of Ł1 and Ł1* by $T1$ and $T1^*$, respectively.

Let T° be the smallest set satisfying the following conditions (a) to (c): (a) If α is a *valid* formula of $A_0, p_1, \ldots, p_n$ are all the variables occurring in α, $\beta_1, \ldots, \beta_n$ are formulas of Ł4, and α_1 is the result of the substitutions $M\beta_i$ for p_i (for $i=1, \ldots, n$) in α, then $\alpha_1 \in T^\circ$. (b) If α and β are formulas of Ł4, then the following formulas (7) to (18) are in T°:

$$\neg M\alpha \to (\neg M \neg \alpha \to \neg M\beta) \tag{7}$$

$$M(\alpha \vee \beta) \leftrightarrow (M\alpha \vee M\beta) \tag{8}$$

$$M(\alpha \wedge \beta) \leftrightarrow (M\alpha \wedge M\beta) \tag{9}$$

$$M\alpha \leftrightarrow MM\alpha \tag{10}$$

$$M(M\alpha \to \alpha) \tag{11}$$

$$M \neg \neg \alpha \leftrightarrow M\alpha \tag{12}$$

$$M(\neg \alpha \to \alpha) \leftrightarrow M\alpha \tag{13}$$

$$M \neg (\alpha \to \neg \alpha) \leftrightarrow M\alpha \tag{14}$$

$$M \neg (\alpha \to \neg \beta) \leftrightarrow M \neg (\beta \to \neg \alpha) \tag{15}$$

$$M(\neg \alpha \to \beta) \leftrightarrow M(\neg \beta \to \alpha) \tag{16}$$

$$M \neg \neg (\alpha \to \neg \beta) \leftrightarrow M(\neg \neg \alpha \to \neg \beta) \tag{17}$$

$$M \neg (\neg \alpha \to \beta) \leftrightarrow M \neg (\neg \alpha \to \neg \neg \beta) \tag{18}$$

(c) If $\alpha \in T^\circ$, and $\alpha \to \beta \in T^\circ$, then $\beta \in T^\circ$.

DEFINITION of $T1$: $T1$ is the smallest set such that: (a) $T^\circ \subseteq T1$; (b) $\alpha \to M\alpha \in T1$, and $L\alpha \to \alpha \in T1$, for all $\alpha \in F$; (c) if $\alpha \leftrightarrow \beta \in T1$, then $M\alpha \leftrightarrow M\beta \in T1$.

DEFINITION of $T1^*$: $T1^*$ is the smallest set for which (a) $T^\circ \subseteq T1^*$, and (b) if $\alpha \leftrightarrow \beta \in T1$, then $M\alpha \leftrightarrow M\beta \in T1^*$.

We shall prove that $Z_Ł$ is a total 1*-representation of Ł1*, and is a total 1-representation of Ł1. To do this, we need some lemmas.

LEMMA 3. *The following identities hold in $Z_Ł$ for arbitrary random variables w, w_1, w_2:*

$$\dot{\neg}\dot{\neg} w = w$$

$$\dot{\neg} w \dot{-} w = w, \qquad w \dot{-} \dot{\neg} w = \dot{\neg} w,$$

$$\dot{\neg} w_1 \dot{-} w_2 = \dot{\neg} w_2 \to w_1, \qquad w_1 \dot{-} \dot{\neg} w_2 = w_2 \dot{-} \dot{\neg} w_1.$$

Our statement follows immediately from (2) and (3).

LEMMA 4. *If w is a random variable discrete with respect to $\langle {}^*0, 0, 1, 1^*\rangle$, then* $\dot{\neg}((w \dot{-} w) \dot{-} \dot{\neg}(w \dot{-} w)) = 1^*$.

PROOF. By the preceding lemma,

$$\dot{\neg}\left((w \dot{-} w) \dot{-} \dot{\neg}(w \dot{-} w)\right) = \dot{\neg}\left(\dot{\neg}(w \dot{-} w)\right) = w \dot{-} w.$$

Using that w is discrete with respect to $\langle {}^*0, 0, 1, 1^*\rangle$, we may use table (1) which shows that $w \dot{-} w = 1^*$ (constantly).

LEMMA 5. *If ω is one of the formulas* (12) *to* (18), *then in any random model belonging to Z_L, $W(\omega) = 1^*$.*

PROOF. *For* (12): By *Lemma* 3, $W(\neg\neg\alpha) = W(\alpha)$, thus $W(M\neg\neg\alpha) = W(M\alpha) = w$, and this w is discrete with respect to $\langle {}^*0, 0, 1, 1^*\rangle$. Now let us take into consideration that

$$\gamma_1 \leftrightarrow \gamma_2 = \neg\left((\gamma_1 \to \gamma_2) \to \neg(\gamma_2 \to \gamma_1)\right),$$

thus

$$W(\gamma_1 \leftrightarrow \gamma_2) = \dot{\neg}\left((W(\gamma_1) \dot{-} W(\gamma_2)) \dot{-} \dot{\neg}(W(\gamma_2) \dot{-} W(\gamma_1))\right). \tag{19}$$

Put $M\neg\neg\alpha$ for γ_1, $M\alpha$ for γ_2, w for $W(\gamma_1)$ and $W(\gamma_2)$, and use the preceding lemma; this yields the desired proof.

The proof for the other formulas is quite analogous, thus it will be omitted here. (The identities of *Lemma* 3 are to be used.)

LEMMA 6. *If ω is one of the formulas* (7) *to* (11), *then in any random model belonging to Z_L, $W(\omega) = 1^*$.*

PROOF. *For* (7): We use (2), (3), and (4). The dots ... occurring below indicate some omitted expressions which do not play any role in the computation.

$$[\neg M\alpha \to (\neg M\neg\alpha \to \neg M\beta) < 1^*] = [M\alpha < 1^*] \cap [M\neg\alpha < 1^*] \cap \ldots =$$

$$= [\alpha \leqq 0] \cap [\alpha \geqq 1] \cap \ldots = \emptyset.$$

For (8): Case (a): $W(M(\alpha \vee \beta))(E) = 0$. Then (by (3)) $W(\alpha \vee \beta)(E) \leqq 0$, i.e. $W(\neg\alpha \to \beta)(E) \leqq 0$. From this it follows by (4) that $W(\alpha)(E) \leqq 0$, and $W(\beta)(E) \leqq 0$. Hence $W(M\alpha)(E) = W(M\beta)(E) = 0$, and $W(M\alpha \vee M\beta)(E) = 0$. — Case (b): $W(M(\alpha \vee \beta))(E) = 1^*$. Then $W(\alpha \vee \beta)(E) = W(\neg\alpha \to \beta)(E) > 0$. Subcase (i): $W(\neg\alpha)(E) = W(\beta)(E) = 0$. Then $W(\alpha)(E) = 1$, $W(M\alpha)(E) = 1^*$, $W(M\beta)(E) = 0$, thus $W(M\alpha \vee M\beta)(E) = 1^*$. — Subcase (ii): $W(\neg\alpha)(E) = W(\beta)(E) = 1$. Then $W(M\alpha)(E) = 0$, $W(M\beta)(E) = 1^*$, thus $W(M\alpha \vee M\beta)(E) = 1^*$. — Subcase (iii): Neither (i), nor (ii) is satisfied. Then one of $W(\alpha)(E)$, $W(\beta)(E)$ is positive, consequently one of $W(M\alpha)(E)$, $W(M\beta)(E)$ is 1^*. Then $W(M\alpha \vee M\beta)(E) = 1^*$.
— Thus we proved that $W(M(\alpha \vee \beta)) = W(M\alpha \vee M\beta) = w$, and w is discrete with respect to $\langle {}^*0, 0, 1, 1^*\rangle$. The completion is the same as in the corresponding proof for (12) in *Lemma* 5.

For (9) the method of the proof is the same as for (8), thus it will be omitted here.

For (10): It is immediate that $W(M\alpha) = W(MM\alpha)$. The completion is the same as in *Lemma* 5.

For (11): We use (2), (3), and (6).

$$[M(M\alpha \to \alpha) < 1^*] = [M(M\alpha \to \alpha) = 0] = [M\alpha \to \alpha \leqq 0] =$$

$$= [M\alpha \geqq 1] \cap [\alpha \leqq 0] = [M\alpha = 1^*] \cap [\alpha \leqq 0] = [\alpha > 0] \cap [\alpha \leqq 0] = \emptyset.$$

LEMMA 7. *If* $\alpha_1 \in T^\circ$ *by* (a) *of the definition of* T°, *then in any random model belonging to* $Z_Ł$, $W(\alpha_1) = 1^*$.

PROOF. We use the notation of (a) just quoted. Let $\Gamma = (\Sigma, W)$ be an arbitrary random model belonging to $Z_Ł$, and let $\Gamma_1 = (\Sigma, W_1)$ be another random model which coincides with Γ, except that

$$W_1(p) = \begin{cases} W(M\beta_i) & \text{if} \quad p = p_i \quad \text{for} \quad i = 1, \ldots, n, \\ 1 & \text{otherwise.} \end{cases}$$

Since $W(M\beta_i)$ is always discrete with respect to $\langle {}^*0, 0, 1, 1^* \rangle$, it follows that Γ_1 is discrete with respect to the same sequence. Using that α is a theorem of A_0, and therefore α is a theorem of Ł4, we get that $W_1(\alpha) = 1^*$. On the other hand, $W(\alpha_1) = W_1(\alpha)$, thus $W(\alpha_1) = 1^*$.

LEMMA 8. *If (in a random model belonging to* $Z_Ł$*)* $W(\alpha \leftrightarrow \beta) \geqq 1$, *then* $W(M\alpha \leftrightarrow M\beta) = 1^*$.

PROOF. Since the possible values of $W(M\alpha)$ and $W(M\beta)$ are 0 and 1^*, we see from table (1) that $(W(M\alpha) \dot{-} W(M\beta))(E)$ has the value 1^* except when $W(M\alpha)(E) = 1^*$ and $W(M\beta)(E) = 0$. In the latter case

$$W(\alpha)(E) > 0, \quad \text{and} \quad W(\beta)(E) \leqq 0. \tag{20}$$

From the assumption and (19) it follows easily that $W(\alpha) \dot{-} W(\beta) \geqq 1$, and $W(\beta) \dot{-} W(\alpha) \geqq 1$. These mean (by (6)) that

$$[\alpha > 0] \cap [\beta < 1] = \emptyset, \quad \text{and} \quad [\beta > 0] \cap [\alpha < 1] = \emptyset.$$

Then, *a fortiori*

$$[\alpha > 0] \cap [\beta \leqq 0] = \emptyset, \quad \text{and} \quad [\beta > 0] \cap [\alpha \leqq 0] = \emptyset.$$

According to the first equality, (20) is impossible, and according to the second one, $W(\beta)(E) > 0$ and $W(\alpha)(E) \leqq 0$ also are incompatible. Thus $W(M\alpha) \dot{-} W(M\beta)$ and $W(M\beta) \dot{-} W(M\alpha)$ equal the constant 1^*. From this and (19) it then follows that $W(M\alpha \leftrightarrow M\beta) = 1^*$.

LEMMA 9. (a) *In every random model belonging to* $Z_Ł$, $W(\alpha \to M\alpha) \geqq 1$, *and* $W(L\alpha \to \alpha) \geqq 1$. (b) *There are random models* Γ_1 *and* Γ_2 *belonging to* $Z_Ł$ *such that in* Γ_1, $W(p \to Mp) < 1^*$, *and in* Γ_2, $W(Lp \to p) < 1^*$. (The mentioned formulas are theorems of Ł4 for any $\alpha \in F$, $p \in V$.)

PROOF. (a) According to (6), (3), and (2) we have:

$$[\alpha \to M\alpha < 1] = [\alpha > 0] \cap [M\alpha < 1] = [\alpha > 0] \cap [M\alpha = 0] = [\alpha > 0] \cap [\alpha \leqq 0] = \emptyset.$$

$$[L\alpha \to \alpha < 1] = [L\alpha > 0] \cap [\alpha < 1] = [\neg M \neg \alpha > 0] \cap [\alpha < 1] =$$

$$= [M \neg \alpha < 1] \cap [\alpha < 1] = [\neg \alpha \leqq 0] \cap [\alpha < 1] = [\alpha \geqq 1] \cap [\alpha < 1] = \emptyset.$$

(b) In the construction of Γ_1 put $W(p) = -\varepsilon/2$ (then $W(Mp) = 0$), and in the construction of Γ_2 put $W(p) = 1 + \varepsilon/2$ (then $W(Lp) = 1$). The proof runs analogously as in *Th.* 4. 2. (Remember that $-\varepsilon = {}^*0$.)

LEMMA 10. *Let Γ be a random model belonging to $Z_Ł$, and let w_1, w_2 be random variables on Γ. Then:* (a) *if $w_1 = 1^*$ and $w_1 \dot{-} w_2 = 1^*$, then $w_2 = 1^*$*; (b) *if $[w_1 < 1] = \emptyset$ and $[w_1 \dot{-} w_2 < 1] = \emptyset$, then $[w_2 < 1] = \emptyset$.*

PROOF. For (a): Using the assumption and (5) we get:

$$\emptyset = [w_1 \dot{-} w_2 < 1^*] = [w_1 > {}^*0] \cap [w_2 < 1^*] \cap \overline{[w_1 = 0] \cap [w_2 = 0]} \cap$$

$$\overline{[w_1 = 1] \cap [w_2 = 1]} = I \cap [w_2 < 1^*] \cap I = [w_2 < 1^*],$$

i.e. $\emptyset = [w_2 < 1^*]$, that is $w_2 = 1^*$.

For (b): use the assumptions and (6).

$$\emptyset = [w_1 \dot{-} w_2 < 1] = [w_1 > 0] \cap [w_2 < 1] = I \cap [w_2 < 1] = [w_2 < 1].$$

THEOREM 5. 1. *The rms $Z_Ł$ is a total* 1*-representation of the system Ł*1.

PROOF. Use *Lemmas* 5, 6, 7, 8, 9(a), and 10(b).

THEOREM 5. 2. *The rms $Z_Ł$ is a total* 1^**-representation of the system Ł*1^*.

PROOF. Use *Lemmas* 5, 6, 7, 8, and 10 (a).

REMARKS. (i) The system Ł1^* is a subsystem of the pure modal part of Ł4. (A formula α belongs to the pure modal part of a modal propositional calculus if any variable occurring in α stands in all its occurrences in the scope of a modal operator in α.) The definition of $T1^*$ shows that all theorems of Ł1^* belong to the pure modal part of Ł4, and *Th.* 4. 2 shows that Ł1^* is only a fragment of this pure modal part. The system Ł1^* does not include the "basic modal logic" of ŁUKASIEWICZ (see [5] or [6]) since $p \to Mp \notin T1^*$, and $Lp \to p \notin T1^*$ (see *Lemma* 9 (b)).

(ii) It can be seen easily that $T1^*$ is a proper part of $T1$, and that Ł1 includes the "basic modal logic", too.

(iii) It can be seen that all the sharply criticized theorems characteristic of ŁUKASIEWICZ's system with the only exception of „$M(p \wedge q) \leftrightarrow (Mp \wedge Mq)$" are theorems neither of Ł1^* nor of Ł1. This gives a new argument that it is reasonable to believe that Ł4 is only a "quasi" modal logic.

(iv) If we interpret the numbers ^{*0}, 0, 1, 1^* as the truth values "impossible" (or "necessarily false"), "false", "true" and "necessary" (or "necessarily true"), then we get the following interpretations of the systems Ł4, Ł1, and Ł1^*.

In Ł4 there are the mentioned four truth values only. All the theorems of Ł4 are necessarily true. As it is well-known no formula of the form $L\alpha$ or $\neg M\alpha$ is a theorem of Ł4. However, if α is a theorem of Ł4, then — according to the matrix **M4** — the only possible value of $L\alpha$ (i.e. of $\neg M \neg \alpha$) is 1. This means in our interpretation: if α is necessarily true, then $L\alpha$ is true, if α is necessarily false, then $\neg M\alpha$ is true. It seems to me that this is a better interpretation of the four-valued matrix **M4** than that of Łukasiewicz according to which 1^* means "true", ^{*0}, means "false", and 0 and 1 are intermediate and indefinite values.

If we assume that there may be truth values other than the mentioned four ones then it is conceivable to define a multi-valued or random logic $Ł = (V, O, F, T)$, where $Ł' = Ł4'$, and $\alpha \in T$ if and only if $\alpha \in F$, and in any random model belonging to $Z_Ł$, $[\alpha < 1] = \emptyset$. Furthermore, we may define a subset T^* of T by the following stipulation: $\alpha \in T^*$ if and only if in any random model belonging to $Z_Ł$, $W(\alpha) = 1^*$. Then T^* is the set of the necessarily true, and $T \setminus T^*$ is that of the "merely true" theorems of $Ł$. It is clear that $Ł1$ is an axiomatized subsystem of $Ł$, $T1$ is a subset of T, and $T1^*$ is a subset of T^*.

The possibility of a finite axiomatiozatin of the above outlined system $Ł$ is still an open question.

§ 6. A model of Lewis's system *S5*

The language of the classical modal system *S5* (due to LEWIS and LANGFORD, see [2]) can be given, in our notation, by

$$S5' = (V, O, F),$$

where $O = O_1 \cup O_2$, $O_1 = \{\neg, L\}$, and $O_2 = \{\wedge\}$. Defined operators are the same as in A_0 (with the same definitions), and, in addition, the modal operator M defined by

$$M\alpha = \neg L \neg \alpha.$$

To define the set T of *theorems* of *S5*, we may use either the axiom system due to GÖDEL (see e.g. in [7, Appendix I]), or the semantical notion of validity due to KRIPKE (see [1]). According to the first version, T is the smallest set satisfying the following conditions:

(a) If α is a theorem of A_0, $p_1, \ldots, p_n$ are all the variables occurring in α, $\beta_1, \ldots, \beta_n$ are elements of F, and α_1 arises from α by substituting β_i for p_i $(i = 1, \ldots, n)$, then $\alpha_1 \in T$. (b) If $\alpha \in F$, $\beta \in F$, then $L\alpha \to \alpha \in T$, $L(\alpha \to \beta) \to (L\alpha \to L\beta) \in T$, and $\neg L\alpha \to \to L \neg L\alpha \in T$. (c) If $\alpha \in T$, $\beta \in F$, and $\alpha \to \beta \in T$, then $\beta \in T$. (d) If $\alpha \in T$, then $L\alpha \in T$.

For the second version of the definition of T, we need the following notions.

By a *semantical model* μ of *S5* let us mean a pair $\mu = (I, \Psi)$, where I is a non-empty set, Ψ is a function on $F \times I$ with values in $\{0, 1\}$, and the following conditions (a) to (c) are satisfied.

(a) If $\alpha \in F$, $E \in I$, then $\Psi(\neg\alpha, E) = 1 - \Psi(\alpha, E)$. (b) If $\alpha \in F$, $\beta \in F$, $E \in I$, then $\Psi(\alpha \wedge \beta, E) = \min(\Psi(\alpha, E), \Psi(\beta, E))$. (c) If $\alpha \in F$, $E \in I$, then

$$\Psi(L\alpha, E) = \begin{cases} 1, & \text{if for all } \quad E' \in I, \quad \Psi(\alpha, E') = 1, \\ 0 & \text{otherwise.} \end{cases}$$

We say that a semantical model μ *satisfies* a formula α (of *S5*) if for some $E \in I$, $\Psi(\alpha, E) = 1$. α is called *satisfiable* if there is a semantical model which satisfies α. If no model satisfies α then α is said to be *unsatisfiable*. Finally, α is called *valid* if $\neg\alpha$ is unsatisfiable.

KRIPKE proved in [1] that α is valid if and only if α is derivable in the axiom system of GÖDEL (cf. his Theorem 8). Henceforth, we can define T as follows: $\alpha \in T$ if and only if α is valid (according to the definition above). In what follows we shall use both definitions of T.

We introduce an *rms* Z_m of *S5'* as follows:

$$Z_m = (0, 1, \theta),$$

where $\theta(\neg) = \dot{\neg}$ and $\theta(\wedge) = \dot{\wedge}$ are the same as in Z_0 (see §2, Part I), and $\theta(L) = \dot{L}$ is defined as follows.

$$(\dot{L}w)(E) = \begin{cases} 1, & \text{if } P(w = 1) = 1, \\ 0, & \text{if } P(w < \tfrac{1}{2}) > 0, \\ \tfrac{1}{2} & \text{otherwise.} \end{cases} \tag{21}$$

Here w is an arbitrary random variable. — Thus $\dot{L}w$ is always a *constant* random variable (0, $^1/_2$, or 1).

We introduce $\dot{M}$ by $\dot{M}w = \dot{\neg}\dot{L}\dot{\neg}w$. From this and (21) it follows that

$$\dot{M}w(E) = \begin{cases} 0, & \text{if } P(w = 0) = 1, \\ 1, & \text{if } P(w > \tfrac{1}{2}) > 0, \\ \tfrac{1}{2} & \text{otherwise.} \end{cases}$$

If the values of w belong to [0, 1], then (a) $\dot{M}w = 0$ if and only if the mathematical expectation of w is 0, (b) $\dot{M}w = {}^1/_2$ if and only if $P(w \leqq {}^1/_2) = 1$, and $P(w > 0) > 0$. — If w is discrete with respect to $\langle 0, 1\rangle$, then $\dot{L}w$ is 0 or 1, and $\dot{M}w$ is 1 or 0.

We see immediately that Z_m is a simple extension of Z_0, and that Z_m satisfies the conditions of *Def.* 3 (see Part I), thus Z_m *is* an *rms* of *S5'*. Furthermore, Z_m is discrete with respect to $\langle 0, 1\rangle$. From these it follows that Z_m is a total $^1/_2$-representation of A_0, and a discrete 1-representation of A_0 in $\langle 0, 1\rangle$.

LEMMA 11. *If α is a theorem of S5, then in any random model belonging to Z_m and discrete with respect to $\langle 0, 1\rangle$, $P(\alpha = 1) = 1$.*

PROOF. We shall use the first version of the definition of T, thus we have to consider the cases (a) to (d).

Case (a). If α is a theorem of A_0, then (by the preceding remark) $[\alpha < 1] = \emptyset$ in any random model Γ belonging to Z_m and discrete with respect to $\langle 0, 1\rangle$. From this it follows — by a slight modification of *Th.* 1 (Part I) — that $[\alpha_1 < 1] = \emptyset$ (where α_1 is the result of a substitution in α), and hence $P(\alpha_1 = 1) = 1$ in any such Γ.

Case (b). — (i). Assume that in Γ, $W(L\alpha \to \alpha)(E) = 0$[4]. Then $W(L\alpha)(E) = 1$, and $W(\alpha)(E) = 0$. These mean that $[L\alpha \to \alpha = 0] \subseteqq [\alpha = 0]$, and that $W(L\alpha)$ is the constant 1. By (21) we have that $P(\alpha = 1) = 1$, i.e. that $P(\alpha = 0) = 0$. Hence, $P(L\alpha \to \alpha = 0) = 0$, that is $P(L\alpha \to \alpha = 1) = 1$. —(ii). Assume that in Γ, $W(L(\alpha \to \beta) \to \to (L\alpha \to L\beta))(E) = 0$. Then $W(L(\alpha \to \beta))(E) = 1$, $W(L\alpha)(E) = 1$, and $W(L\beta)(E) = 0$. These mean that $W(L(\alpha \to \beta))$ and $W(L\alpha)$ are the constant 1, and $W(L\beta)$ is the constant 0. Then, by (21), we have that $P(\alpha \to \beta = 0) = 0$, $P(\alpha = 0) = 0$, $P(\beta = 0) > 0$. But

$$[\beta = 0] = ([\beta = 0] \cap [\alpha = 0]) \cup ([\beta = 0] \cap [\alpha = 1] \subseteqq [\alpha = 0] \cup [\alpha \to \beta = 0], \tag{22}$$

thus

$$P(\beta = 0) \leqq P(\alpha = 0) + P(\alpha \to \beta = 0) = 0, \tag{23}$$

[4] Here $\Gamma = (\Sigma = (I, \mathscr{A}, P), W)$ is a random model belonging to Z_m and discrete with respect to $\langle 0,1\rangle$; $E \in I$, $\alpha \in F$, $\beta \in F$.

which contradicts $P(\beta=0)>0$. Thus our assumption is false. — (iii). $W(\neg L\alpha)$ is constantly 0 or 1, and $W(L\neg L\alpha)$ is always the same constant. Thus $W(\neg L\alpha \to L^{-} L\alpha)=1$.

Case (c). Assume that in Γ, $P(\alpha\to\beta=0)=0$, and $P(\alpha=0)=0$.[4] Using (22) and (23) we get that $P(\beta=0)=0$.

Case (d). Assume that in Γ, $P(\alpha=1)=1$.[4] Then, by (21), $W(L\alpha)$ is the constant 1, thus $P(L\alpha=1)=1$. — Our proof is concluded.

LEMMA 12. *If α is not a theorem of S5, then there is a random model Γ belonging to Z_m and discrete with respect to $\langle 0, 1\rangle$ in which $P(\alpha=0)>0$.*

PROOF. Assume that α is not a theorem of *S5*. By the second version of the definition of T, this means that there is a semantical model $\mu=(I, \Psi)$ such that for some $E\in I$, $\Psi(\alpha, E)=0$. We may assume here that I is finite: $I=\{E_1, \ldots, E_n\}$, $n\geqq 1$ (see Theorem 8 in [1]). Let us consider the random model $\Gamma=(\Sigma=(I, \mathscr{A}, P), W)$ where I is the same as in μ, $\mathscr{A}$ is the σ-algebra on the power set of I, P is defined by $P(\{E_i\})=\frac{1}{n}$ for $i=1, \ldots, n$, and W is defined by $W(p)(E_i)=\Psi(p, E_i)$ for all $p\in V$, $E_i\in I$. (Clearly, Γ is discrete with respect to $\langle 0, 1\rangle$.) From this it follows easily that (a) for all $B\in\mathscr{A}$, $P(B)=0$ if and only if $B=\emptyset$, and $P(B)=1$ if and only if $B=I$; (b) for all $\beta\in F$, $W(\beta)(E_i)=\Psi(\beta, E_i)$. Henceforth $\Psi(\alpha, E)=0$ implies $W(\alpha)(E)=0$. Since $P(\{E\})=\frac{1}{n}$, we get that $P(\alpha=0)>0$; thus our proof is concluded.

In Part *I* we introduced the notions of the *total d-representation*, the *adequate total d*-representation, the *discrete d*-representation in $\langle e_1, \ldots, e_m\rangle$, the *adequate discrete d*-representation in $\langle e_1, \ldots, e_m\rangle$, and the *perfect discrete* representation in $\langle e_1, \ldots, e_m\rangle$ *(Definitions* 6, 7, 14, 15*)*. Let us modify these notions by replacing "$P(\alpha<d)=0$" for "$[\alpha<d]=\emptyset$" in their definitions. Then we get from the two preceding lemmas the following:

THEOREM 6. 1. *The rms Z_m is a perfect discrete representation of the system S5 in $\langle 0, 1\rangle$.*

LEMMA 13. *If $\Gamma_1=(\Sigma=(I, \mathscr{A}, P), W_1)$ is a random model belonging to Z_m, then there is a random model $\Gamma_2=(\Sigma, W_2)$ (with the same Σ as in Γ_1) such that Γ_2 is discrete with respect to $\langle 0, 1\rangle$, and for all $\alpha\in F$,*

$$(24)\qquad [W_1(\alpha)<{}^1/_2]\subseteqq[W_2(\alpha)=0], \quad \text{and} \quad [W_1(\alpha)>{}^1/_2]\subseteqq[W_2(\alpha)=1].$$

PROOF. Let us define W_2 as follows:

$$W_2(p)(E)=\begin{cases}1, & \text{if } W_1(p)(E)\geqq \frac{1}{2},\\ 0 & \text{otherwise}\end{cases}$$

for all $p\in V$, $E\in I$. This means immediately that (24) is true for α if α is atomic (i.e. if α is a variable). We have to prove that if (24) is true for α and β, then it is true for $\neg\alpha$, $L\alpha$, $\alpha\wedge\beta$.

Case of $\neg\alpha$. Using that $W(\neg\alpha)=1-W(\alpha)$, we get immediately that $[W_1(\alpha)<{}^1/_2]\subseteqq[W_2(\alpha)=0]$ implies $[W_1(\neg\alpha)>{}^1/_2]\subseteqq[W_2(\neg\alpha)=1]$, and $[W_1(\alpha)>{}^1/_2]\subseteqq[W_2(\alpha)=1]$ implies $[W_1(\neg\alpha)<{}^1/_2]\subseteqq[W_2(\neg\alpha)=0]$.

Case of $L\alpha$. Assume that $[W_1(L\alpha) < {}^1/_2] \neq \emptyset$. This means that $W_1(L\alpha) = 0$. Then, by (21), $P(W_1(\alpha) < {}^1/_2) > 0$. From our hypothesis $[W_1(\alpha) < {}^1/_2] \subseteq [W_2(\alpha) = 0]$ it then follows $P(W_2(\alpha) = 0) > 0$, consequently $W_2(L\alpha) = 0$. — Now assume that $[W_1(L\alpha) > {}^1/_2] \neq \emptyset$. This means that $W_1(L\alpha) = 1$, hence $P(W_1(\alpha) = 1) = 1$, and *a fortiori*, $P(W_1(\alpha) > {}^1/_2) = 1$. From the hypothesis $[W_1(\alpha) > {}^1/_2] \subseteq [W_2(\alpha) = 1]$ it then follows that $P(W_2(\alpha) = 1) = 1$, thus $W_2(L\alpha) = 1$.

Case of $\alpha \wedge \beta$. Assume that (24) is true for α and β. Then:

$$[W_1(\alpha \wedge \beta) < {}^1/_2] = [W_1(\alpha) < {}^1/_2] \cup [W_1(\beta) < {}^1/_2] \subseteq [W_2(\alpha) = 0] \cup$$

$$\cup [W_2(\beta) = 0] = [W_2(\alpha \wedge \beta) = 0]; \quad \text{and} \quad [W_1(\alpha \wedge \beta) > {}^1/_2] = [W_1(\alpha) > {}^1/_2] \cap$$

$$\cap [W_1(\beta) > {}^1/_2] \subseteq [W_2(\alpha) = 1] \cap [W_2(\beta) = 1] = [W_2(\alpha \wedge \beta) = 1].$$ — Our proof is concluded.

THEOREM 6. 2. *The rms* Z_m *is an adequate total* ${}^1/_2$*-representation of the system S5.*

PROOF. (a) If α is not a theorem of *S*5, then, by *Lemma* 12, there is a random model Γ in which $P(\alpha = 0) > 0$. (b) Assume, indirectly, that α is a theorem of *S5*, but there is a random model $\Gamma_1 = (\Sigma, W_1)$ belonging to Z_m in which $P(\alpha < {}^1/_2) > 0$. Then, by *Lemma* 13, there is a $\Gamma_2 = (\Sigma, W_2)$ (with the same Σ as Γ_1) which is discrete with respect to $\langle 0, 1 \rangle$, and $[W_1(\alpha) < {}^1/_2] \subseteq [W_2(\alpha) = 0]$. From these it follows that in Γ_2, $P(\alpha = 0) > 0$. But this contradicts *Lemma* 11 (using that Γ_2 is discrete with respect to $\langle 0, 1 \rangle$). Thus our indirect assumption is false.

REFERENCES

[1] KRIPKE, S. A.: A completeness theorem in modal logic. *The Journal of Symbolic Logic*, **24** (1959), 1—14.

[2] LEWIS, C. I. and LANGFORD, C. H.: *Symbolic Logic*. 2nd ed. Dover Publications, Inc. New York 1959.

[3] ŁUKASIEWICZ, J.: Philosophische Bemerkungen zu mehrwertigen Systemen des Aussagenkalküls. *Comptes Rendus des séances de la Société des Sciences et des Lettres de Varsovie, Cl. III.*, **23** (1930), 51—57.

[4] ŁUKASIEWICZ, J. and TARSKI, A.: Investigations into the sentential calculus. *Logic, Semantics, Metamathematics* (by A. Tarski). Oxford 1956. 28—59.

[5] ŁUKASIEWICZ, J.: A System of Modal Logic. *The Journal of Computing Systems*, **1**, 3. (1953), 111—149.

[6] ŁUKASIEWICZ, J.: A System of Modal Logic. *Actes du XIième Congrès Intern. de Phil., 1953.* XIV, 82—87.

[7] PRIOR, A. N.: *Formal Logic*. Oxford 1955.

Institute of Philosophy of the Hungarian Academy of Sciences, Budapest

(Received October 20, 1968.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 267—276.

RANK ORDER STATISTICS RELATED TO A GENERALIZED RANDOM WALK

by

S. G. MOHANTY and B. R. HANDA

1. Introduction and Summary

Let $X_1, \ldots, X_{\mu n}$ and $Y_1, \ldots, Y_n$ be two random samples from the same continuous distribution, having the empirical distribution functions as $F_{\mu n}(x)$ and $G_n(y)$ respectively. Arranging the combined $(\mu n+n)$ observations in increasing order as $\zeta_1 < \zeta_2 < \cdots < \zeta_{(\mu+1)n}$, and replacing each X observation by 1, and each Y observation by $-\mu$, we have corresponding to an ordered sequence $(\zeta_1, \zeta_2, \ldots, \zeta_{(\mu+1)n})$, a sequence of μn, $+1$'s and n, $-\mu$'s, which we shall term as a rank order indicator. The $\binom{(\mu+1)n}{n}$ possible rank order indicators are assumed to be equally likely.

A rank order statistic is a random variable defined on the ordered sequence $(\zeta_1, \ldots, \zeta_{(\mu+1)n})$, through the rank order indicator. Letting

$$H_{\mu,n}(u) = [F_{\mu n}(u) - G_n(u)]\, n\mu, \quad -\infty < u < \infty, \tag{1}$$

we note that statistics defined through $H_{\mu,n}(u)$, such as one sided KOLMOGOROV—SMIRNOV statistics,

$$D^+_{\mu,n} = \max_{-\infty<u<\infty} H_{\mu,n}(u),$$

can be treated as rank order statistics.

In [1], DWASS develops a new method (other than the combintorial one) based on the simple random walk, in order to derive the distributions of some rank order statistics, defined on $H_{1,n}(u)$. In this paper, the technique by DWASS is extended, so as to cover the case of general μ. For this purpose, we consider the random walk $S_0, S_1, \ldots,$ where

$$S_0 = W_0 = 0 \quad \text{and} \quad S_i = \sum_{j=0}^{i} W_j, \qquad W_1, W_2, \ldots,$$

being identically and independently distributed random variables, having the probability distributions as follows:

$$W_i = \begin{cases} 1, & \text{with probability } p, \\ -\mu, & \text{with probability } q, \end{cases}$$

$p+q = 1$, $i = 1, 2, \ldots$. A lattice path graph can be associated with the random walk in a familiar way, by starting from origin, and then by taking either, a unit horizontal step or, a unit vertical step at the ith stage, according as W_i is 1 or $-\mu$.

The following results are frequency used in the sequel: 1. The number of lattice paths from $(0, 0)$ to $(\mu n, n)$ lying below and never crossing the line $x = \mu y$ is

$$\frac{1}{(\mu+1)n+1}\binom{(\mu+1)n+1}{n}, \tag{2}$$

and of those lying below and never touching the line $x = \mu y$ except at the end points is

$$\frac{\mu}{(\mu+1)n-1}\binom{(\mu+1)n-1}{n-1} \tag{3}$$

(see Corollary on page 256 in [4]). 2. For any α and β,

$$\sum_{k=0}^{\infty} A_k(\alpha, \beta)\theta^k = x^{\alpha}, \tag{4}$$

where

$$A_k(\alpha, \beta) = \frac{\alpha}{\alpha+k\beta}\binom{\alpha+k\beta}{k}$$

$$\theta = \frac{x-1}{x^{\beta}} \quad \text{and} \quad |\theta| < \left|\frac{(\beta-1)^{\beta-1}}{\beta^{\beta}}\right|,$$

(see (7) in [3]).

2. The Random Walk

By the use of well-known results on recurrent events, it is remarked that, the event of returning to the origin is a transient recurrent event if $p \neq \frac{\mu}{\mu+1}$. Let $U(s)$ and $F(s)$ represent the generating functions for the return time to zero and the time for the first return to zero respectively. Clearly,

$$U(s) = \sum_{n=0}^{\infty}\binom{(\mu+1)n}{n} p^{\mu n} q^n s^{(\mu+1)n} = \frac{x}{[(\mu+1)-\mu x]} \tag{5}$$

by (9) in [3], where,

$$s^{\mu+1} p^{\mu} q = \frac{x-1}{x^{\mu+1}} \quad \text{and} \quad |s|^{\mu+1} p^{\mu} q < \frac{\mu^{\mu}}{(\mu+1)^{\mu+1}}.$$

Thus from Theorem 1 on page 285 in [2], we obtain

$$F(s) = \frac{U(s)-1}{U(s)} = (\mu+1)\, p^{\mu} q x^{\mu} s^{\mu+1}. \tag{6}$$

With the help of (4), (6) can be expressed as

$$(\mu+1)\, p^{\mu} q x^{\mu} s^{\mu+1} = \sum_{n=1}^{\infty} (\mu+1)\, A_{n-1}(\mu, \mu+1)\, p^{\mu n} q^n s^{(\mu+1)n}. \tag{7}$$

From (6), the probability of ever returning to zero is given by,

$$F(1) = (\mu+1)\,p^{\mu}qy^{\mu}, \tag{8}$$

where $$p^{\mu}q = (y-1)/y^{\mu+1} \quad \text{and} \quad p^{\mu}q < \mu^{\mu}/(\mu+1)^{\mu+1}.$$

Since $p^{\mu}q < \mu^{\mu}/(\mu+1)^{\mu+1}$ implies $p \neq \dfrac{\mu}{\mu+1}$ and in that case $F(1)<1$, (because of the remark at the beginning of this section), we observe that y, which is a positive root of the equation

$$p^{\mu}qy^{\mu+1} - y + 1 = 0, \tag{9}$$

must be less than $\dfrac{\mu+1}{\mu}$. Furthermore, from the discussion on page 302 in [2], we state that there are only two positive roots, $\dfrac{1}{p}$ and y (say),

whence, $y < \dfrac{\mu+1}{\mu} < \dfrac{1}{p}$ if $p < \dfrac{\mu}{\mu+1}$, and $\dfrac{1}{p} < \dfrac{\mu+1}{\mu} < y$ if $p > \dfrac{\mu}{\mu+1}$.

Thus without loss of generality, we may assume that $p < \dfrac{\mu}{\mu+1}$, in which case the probability of ever returning to zero is $(\mu+1)p^{\mu}qy^{\mu}$, where y is the smallest positive root of (9).

Some auxiliary results: From (8) the probability of being at zero, for more than k times is

$$[(\mu+1)\,p^{\mu}qy^{\mu}]^{k}, \tag{10}$$

the initial visit at time zero, being counted.

Let $G(s, k)$, denote the generating function for the time to reach k. Then

$$G(s,1) = \sum_{n=0}^{\infty} A_n(1, \mu+1)\,p^{\mu n+1}q^{n}s^{(\mu+1)n+1} \tag{11}$$

$$= psx, \quad \text{(by (4))}$$

and the probability of ever reaching 1 is given by

$$G(1,1) = py \tag{12}$$

Using (11),

$$G(s,k) = (psx)^{k}, \qquad k = 1, 2, \ldots, \tag{13}$$

and hence the probability of ever reaching k is

$$G(1,k) = (py)^{k}, \qquad k \geq 1. \tag{14}$$

Let $D^{+} = \max(0, S_1, S_2, \ldots)$, and $Q =$ number of indices i for which $S_i = D^{+}$. Then, using a similar argument as in Appendix (11), in [1], we have

$$\mathsf{P}(D^{+} = k, Q = r) = (py)^{k}[q(py)^{\mu}]^{r-1}(1-(\mu+1)qp^{\mu}y^{\mu}), \tag{15}$$

$$k = 0, 1, \ldots, \quad r = 1, 2, \ldots .$$

Next we state and prove

$$G(1, -1) = \mu p^{\mu-1} q y^{\mu}, \tag{16}$$

which is used subsequently.

It is seen that

$$G(s, -1) = \sum_{n=1}^{\infty} (a_n - b_n) p^{\mu n-1} q^n s^{(\mu+1)n-1}, \tag{17}$$

where

b_n = number of lattice paths from $(-1, 0)$ to $(\mu n - 1, n)$ lying above the line $x = \mu y - 1$, and never touching it except at the end points;

and

a_n = number of lattice paths from $(-1, 0)$ to $(\mu n - 1, n)$ never passing through a lattice point on the line $x = \mu y - 1$, except at the end points.

Then using (3) and (7), we can write

$$a_n - b_n = (\mu+1) A_{n-1}(\mu, \mu+1) - A_{n-1}(\mu, \mu+1) = \mu A_{n-1}(\mu, \mu+1).$$

A further simplification of (17), with the help (4) yields

$$G(s, -1) = \mu p^{\mu-1} q x^{\mu} s^{(\mu+1)n-1}$$

which readly gives (16).

Again let $N^+(r)$ = the number of indices i for which $S_{i-1} = r$, $S_i = r+1$. Following the steps in [1] and using (16), we obtain

$$\mathsf{P}(N^+(0) \geqq k) = [py \cdot \mu p^{\mu-1} q y^{\mu}]^k, \tag{18}$$

and

$$\mathsf{P}(N^+(r) \geqq k) = (py)^r \, \mathsf{P}(N^+(0) \geqq k) = (py)^{r+\mu k} \mu^k q^k y^k. \tag{19}$$

For given $i_2 > i_1 \geqq 0$, if $S_{i_1} = 0$, $S_i > 0$ for all $i_1 < i < i_2$ and $S_{i_2} \leqq 0$, we say that there is a positive sojourn, whereas, if $-\mu < S_{i_1} \leqq 0$, $S_i < 0$ for all $i_1 < i < i_2$, and $S_{i_2} = 0$, we say that there is a negative sojourn. Denote by N^+, the number of positive sojourns. Between two consecutive zeros of the random walk, there can be at most one positive and one negative sojourn. The conditional distribution of N^+, given the number of returns to zero to be r, is binomial with parameters r and $\frac{\mu}{\mu+1}$. Hence by a similar argument as for II, section 4 in [1], we have

$$\begin{aligned} \mathsf{P}(N^+ = k) &= \sum_{r=k}^{\infty} \binom{r}{k} \left(\frac{\mu}{\mu+1}\right)^k \left(\frac{1}{\mu+1}\right)^{r-k} [(\mu+1) p^{\mu} q y^{\mu}]^r \left(1 - (\mu+1) p^{\mu} q y^{\mu}\right) = \\ &= \mu^k (q p^{\mu} y^{\mu})^k (1 - q p^{\mu} y^{\mu})^{-k-1} \left(1 - (\mu+1) p^{\mu} q y^{\mu}\right) = \\ &= \mu^k (q p^{\mu})^k \, y^{(\mu+1)k+1} \left(1 - (\mu+1) p^{\mu} q y^{\mu}\right), \end{aligned} \tag{20}$$

the last expression being due to (7).

3. Distributions of rank order statistics

At the outset, we state the modified forms of the main results in [1], which play vital role for finding the distributions of rank order statistics.

LEMMA 1. *For any* p *in* $(0, 1)$ *the conditional distribution of* $W_1, W_2, \ldots, W_{(\mu+1)n}$ *given* $W_1 + W_2 + \cdots + W_{(\mu+1)n} = 0$, *assigns equal probabilities to each of* $\binom{(\mu+1)n}{n}$ *possible sequences of* μn, $+1$'s *and* n, $-\mu$'s.

In other words, the distribution is exactly the same as that of rank order indicators described in section-1.

Let $p < \dfrac{\mu}{\mu+1}$ and T be the time for the last return to zero in the random walk. Define V as a function on the random walk which is completely determined by $W_1, \ldots, W_T$, whenever $T > 0$.

LEMMA 2. (a) *Conditional distribution of* V, *given that* $T = (\mu+1)n$ *is exactly that of a rank order satistic.*

(b) *Conversely, if* $V_{\mu,n}$ *is a rank order statistic defined for* $n = 1, 2, \ldots$, *then there is a function* V, *such that the distribution of* V, *given that* $T = (\mu+1)n$ *is exactly the distribution of* $V_{\mu,n}$.

Because of (b), we say that V is the function on the random walk corresponding to $V_{\mu,n}$.

THEOREM. *Suppose,* $V_{\mu,n}$ *is a rank order statistic for every* n *and* V *is the corresponding function defined on the random walk. Define*

$$h(p) = E(V), \quad p < \frac{\mu}{\mu+1}, \tag{21}$$

Then we have the following power series (in powers of $p^\mu q$*) expansion:*

$$\frac{h(p)}{1-(\mu+1)p^\mu q y^\mu} = \sum_{n=0}^{\infty} E(V_{\mu,n}) \binom{(\mu+1)n}{n} p^{\mu n} q^n. \tag{22}$$

The proofs are in the lines of [1] and hence omitted.

We present another useful result on the power series expansion:

$$\frac{p^k}{1-(\mu+1)p^\mu q y^\mu} = \sum_{n=\left\{\frac{k}{\mu}\right\}}^{\infty} \binom{(\mu+1)n-k}{n} p^{\mu n} q^n, \quad k > 0, \tag{23}$$

where $\{\zeta\}$ is the smallest integer greater than or equal to ζ.

For the proof we observe that,

$$\mathsf{P}(W_1 = 1, \ldots, W_k = 1) = p^k,$$

$$\mathsf{P}(W_1 = 1, \ldots, W_k = 1) = \sum_{n=\left\{\frac{k}{\mu}\right\}}^{\infty} [\mathsf{P}(W_1 = 1, \ldots, W_k = 1 | T = (\mu+1)n) \cdot \cdot \mathsf{P}(T = (\mu+1)n)]$$

$$\mathsf{P}(W_1 = 1, \ldots, W_k = 1 | T = (\mu+1)n) = \binom{(\mu+1)n-k}{n} \Big/ \binom{(\mu+1)n}{n},$$

and

$$\mathsf{P}(T = (\mu+1)) = \binom{(\mu+1)n}{n} p^{\mu n} q^n (1-(\mu+1)p^\mu q y^\mu).$$

The following is the list of rank order statistics and their distributions:

(I) $N_{\mu,n}$ = the number of indices i for which

$$H_{\mu,n}(\zeta_i) = 0, \quad i = 0, 1, \ldots, (\mu+1)n, \quad \text{where} \quad \zeta_0 = -\infty;$$

$$\mathsf{P}(N_{\mu,n} > k) = (\mu+1)^k \binom{(\mu+1)n-k}{n-k} \Big/ \binom{(\mu+1)n}{n}, \quad \text{for} \quad k = 0, 1, \ldots, n. \tag{24}$$

(II) $N_{\mu,n}(r)$ = the numben of indices i for which

$$H_{\mu,n}(\zeta_i) = r, \quad r \geqq 0, \; i = 0, 1, \ldots, (\mu+1)n;$$

$$\mathsf{P}(N_{\mu,n}(r) > k) = \frac{(\mu+1)^k \sum_{i=\left\{\frac{r}{\mu}\right\}}^{n-k} \binom{(\mu+1)i-r}{i} A_{n-i-k}(r+\mu k, \mu+1)}{\binom{(\mu+1)n}{n}}, \tag{25}$$

for $k = 0, 1, \ldots, n - \left\{\frac{r}{\mu}\right\}$.

(III) $N^+_{\mu,n}(r)$ = the number of indices i for which

$$H_{\mu,n}(\zeta_i) = r+1 \quad \text{and} \quad H_{\mu,n}(\zeta_{i-1}) = r,$$
$$r = 0, 1, \ldots, \quad i = 1, \ldots, (\mu+1)n;$$

$$\mathsf{P}(N^+_{\mu,n}(r) \geqq k) = \frac{\mu^k \sum_{i=\left\{\frac{r}{\mu}\right\}}^{n-k} \binom{(\mu+1)i-r}{i} A_{n-i-k}(r+(\mu+1)k, \mu+1)}{\binom{(\mu+1)n}{n}}, \tag{26}$$

for $k = 0, 1, \ldots, n - \left\{\frac{r}{\mu}\right\}$.

(IV) $$D^+_{\mu,n} = \max_{0 \leqq i \leqq (\mu+1)n} H_{\mu,n}(\zeta_i);$$

(27) $$\mathsf{P}(D^+_{\mu,n} \geqq k) = \frac{\sum_{i=\left\{\frac{k}{\mu}\right\}}^{n} \binom{(\mu+1)i-k}{i} A_{n-i}(k, \mu+1)}{\binom{(\mu+1)n}{n}},$$

for $k = 0, 1, \ldots, \mu n.$

(V) $Q_{\mu,n}$ = the number indices i for which

$$H_{\mu,n}(\zeta_i) = D^+_{\mu,n}, \quad 0 \leqq i \leqq (\mu+1)n;$$

(28)
$$\mathsf{P}(Q_{\mu,n} = r, D^+_{\mu,n} = k) =$$

$$= \begin{cases} \binom{(\mu+1)\left\{\frac{k}{\mu}\right\}-k}{\left\{\frac{k}{\mu}\right\}} \Big/ \binom{(\mu+1)n}{n}, & \text{when} \quad n = \left\{\frac{k}{\mu}\right\}+r-1, \\ \dfrac{\binom{(\mu+1)(n-r+1)-k}{n-r+1}}{\binom{(\mu+1)n}{n}} + \sum_{i=\left\{\frac{k}{\mu}\right\}}^{n-r} \binom{(\mu+1)i-k}{i} \cdot \\ \quad \cdot \dfrac{(A_{n-r+1-i}(k+(r-1)\mu, \mu+1) - (\mu+1)A_{n-k-i}(k+r\mu, \mu+1)}{\binom{(\mu+1)n}{n}}, \\ \quad \text{when} \quad n \geqq \left\{\frac{k}{\mu}\right\}+r, \end{cases}$$

for $r = 1, \ldots, n-\left\{\frac{k}{\mu}\right\}+1, \quad k = 0, 1, \ldots, \mu n.$

(VI) $Q_{\mu,n,k}$ = index i for which $H_{\mu,n}(\zeta_i) = 0$,

$0 \leqq i \leqq (\mu+1)n$, for the k th time, i. e. the position of the k th zero;

(29) $$\mathsf{P}(Q_{\mu,n,k} = (\mu+1)i, N_{\mu,n} \geqq k) = \frac{(\mu+1)^k A_{i-k}(\mu k, \mu+1)\binom{(\mu+1)(n-i)}{n-i}}{\binom{(\mu+1)n}{n}},$$

for $i = k, k+1, \ldots, n, k = 0, 1, \ldots, n.$

(VII) $R^+_{\mu,n}$ = smallest i such that

$$H_{\mu,n}(\zeta_i) = D^+_{\mu,n},\ D^+_{\mu,n} > 0, \quad 0 \leqq i \leqq (\mu+1)n;$$

$$\mathsf{P}(R^+_{\mu,n} = \mu n,\ D^+_{\mu,n} = \mu n) = 1 \Big/ \binom{(\mu+1)n}{n}, \tag{30}$$

and

$$\mathsf{P}(R^+_{\mu,n} = k+(\mu+1)i,\ D^+_{\mu,n} = k) = A_i(k, \mu+1)\left[\binom{(\mu+1)(n-i)-k}{n-i} - \sum_{j=\left[\frac{k}{\mu}\right]+1}^{n-i} \binom{(\mu+1)j-k-1}{j} A_{n-i-j}(1, \mu+1)\right] \Big/ \binom{(\mu+1)n}{n},$$

for $i = 0, 1, \ldots, n - \left\{\frac{k}{\mu}\right\} - 1, \quad k = 0, 1, \ldots, \mu n - 1,$

(here $[x]$ is the largest integer less than or equal to x).

(VIII) For any indices $0 \leqq i_1 < i_2 \leqq (\mu+1)n$, if $H_{\mu,n}(\zeta_{i_1}) = 0$, $H_{\mu,n}(\zeta_i) > 0$ for all $i_1 < i < i_2$, and $H_{\mu,n}(\zeta_{i_2}) \leqq 0$, we say that there is a positive sojourn. Let $N^+_{\mu,n}$, denote the number of positive sojourn, then

$$\mathsf{P}(N^+_{\mu,n} \geqq k) = \mu^k \binom{(\mu+1)n}{n-k} \Big/ \binom{(\mu+1)n}{n}, \tag{31}$$

for $k = 0, 1, \ldots, n$.

PROOF of the results: In most cases expansion formulae (4) and (23) are frequently used.

(I) and (II): Let, $N(r)$ = number of indices i for which $S_i = r$, $i \geqq 0$. Using (10) and (14), we can write

$$\mathsf{P}(N(r) > k) = (py)^r\,[(\mu+1)qp^\mu y^\mu]^k, \qquad k \geqq 0.$$

Thus

$$\frac{h(p)}{(1-(\mu+1)p^\mu q y^\mu)} = \frac{\mathsf{P}(N(r) > k)}{(1-(\mu+1)p^\mu q y^\mu)} = \frac{p^r}{(1-(\mu+1)p^\mu q y^\mu)}\,(\mu+1)^k (p^\mu q)^k y^{r+\mu k}.$$

Expanding the last member in the power series of $p^\mu q$, with help of (4) and (23), we get

$$\sum_{i=\left\{\frac{r}{\mu}\right\}}^{\infty} \binom{(\mu+1)i-r}{i} p^{\mu i} q^i \cdot (\mu+1)^k (p^\mu q)^k \sum_{j=0}^{\infty} A_j(r+\mu k, \mu+1)(p^\mu q)^j$$

$$= (\mu+1)^k \sum_{i=\left\{\frac{r}{\mu}\right\}}^{\infty} \sum_{n=i+k}^{\infty} \binom{(\mu+1)i-r}{i} A_{n-i-k}(r+\mu k, \mu+1)(p^\mu q)^n$$

$$= (\mu+1)^k \sum_{n=\left\{\frac{r}{\mu}\right\}}^{\infty} \left(\sum_{i=\left\{\frac{r}{\mu}\right\}}^{n-k} \binom{(\mu+1)i-r}{i} \right) A_{n-i-k}(r+\mu k, \mu+1)(p^\mu q)^n.$$

An application of the theorem then yields the required result (25).

When, $r=0$

$$\sum_{i=0}^{n-k}\binom{(\mu+1)i}{i}A_{n-k-i}(\mu k,\mu+1)=\binom{(\mu+1)n-k}{n-k} \tag{32}$$

by (11) in [3]. This completes the proof of (I).

(III): Using the expression in (19) and following the same steps as for (II), we obtain (26). For $\mu=1$,

$$\mathsf{P}\left(N^{+}_{1,n}(r)\geqq k\right)=\frac{\sum_{i=r}^{n-k}\binom{2i-r}{i-r}\dfrac{r+2k}{r+2k+2(n-i-k)}\binom{r+2k+2(n-i-k)}{n-i-k}}{\binom{2n}{n}} \tag{33}$$

the right hand side simplifies to $\binom{2n}{n-k-r}\Big/\binom{2n}{n}$, with the help of summation formula (11) in [3].

This verifies IV in section 4 of [1].

For $r=0$, one gets

$$\mathsf{P}\left(N^{+}_{r,n}(0)\geqq k\right)=\mu^{k}\binom{(\mu+1)n}{n-k}\Big/\binom{(\mu+1)n}{n}. \tag{34}$$

The proofs of (IV) and (V) employ no new technique, and therefore are omited. As a corollary of (V), we state that

$$\mathsf{P}(Q_{\mu,n}=r, D^{+}_{\mu,n}=0)=\begin{cases}1\Big/\binom{(\mu+1)n}{n}, & \text{for} \quad n=r-1\\ \dfrac{A_{n-r-1}(\mu(r-1),\mu+1)}{\binom{(\mu+1)n}{n}}, & \text{for} \quad n\geqq r.\end{cases} \tag{35}$$

(VI): Let $N=$number of indices i for which $S_i=0$, $i\geqq 0$, and $\Theta_k=$the position of kth zero, i.e. it is the index for which $S_i=0$ for the kth time. Then by using (6) and (8) we have

$$\begin{aligned}h(p)&=E(s^{\Theta_k};\ N=k+r)\\&=\sum_{i=k}^{\infty}\mathsf{P}\left(\Theta_k=(\mu+1)i,\ N=k+r\right)s^{(\mu+1)i}\\&=[(\mu+1)p^{\mu}qs^{\mu+1}x^{\mu}]^{k}[(\mu+1)p^{\mu}qy^{\mu}]^{r}\left(1-(\mu+1)p^{\mu}qy^{\mu}\right).\end{aligned}$$

Hence

$$\frac{h(p)}{1-(\mu+1)p^{\mu}qy^{\mu}}=(\mu+1)^{r+k}(p^{\mu}q)^{r+k}s^{(\mu+1)k}x^{\mu k}y^{r\mu}.$$

Expanding right hand side as a power series in $p^{\mu}q$ with the help of (4), we get

$$(p^{\mu}q)^{r+k}(\mu+1)^{r+k}s^{(\mu+1)k}\sum_{l=0}^{\infty}A_l(\mu k,\mu+1)(p^{\mu}qs^{\mu+1})^l\sum_{j=0}^{\infty}A_j(\mu r,\mu+1)(p^{\mu}q)^j=$$

$$=(\mu+1)^{r+k}\sum_{i=k}^{\infty}\sum_{n=i+r}^{\infty}A_{i-k}(\mu k,\mu+1)A_{n-r-i}(\mu r,\mu+1)(p^{\mu}q)^n s^{(\mu+1)i},$$

and the coefficient of $s^{(\mu+1)i}(p^{\mu}q)^n\binom{(\mu+1)n}{n}$ gives

$$\text{(36)}\quad \mathsf{P}(Q_{\mu,n,k}=(\mu+1)i,\,N_{\mu,n}=r+k)=(\mu+1)^{r+k}\frac{A_{i-k}(\mu k,\mu+1)A_{n-r-i}(\mu r,\mu+1)}{\binom{(\mu+1)n}{n}}.$$

Using the relation

$$A_{n-r-i}(\mu r,\mu+1)=\binom{(\mu+1)(n-i)-r}{n-r-i}-(\mu+1)\binom{(\mu+1)(n-i)-r-1}{n-r-i-1}$$

and summing (36) over r, $0\leqq r\leqq n-i$ we get (29).

(VII): Let R^+, be the index at which the maximum of the random walk is first achieved. A simple argument and the application of (12) and (13) yield

$$h(p)=E(s^{R^+};D^+=k)=\sum_i \mathsf{P}(R^+=k+(\mu+1)i,\,D^+=k)s^{(\mu+1)i+k}=$$

$$=(psx)^k(1-py).$$

Thus

$$\text{(37)}\qquad \frac{h(p)}{1-(\mu+1)p^{\mu}qy^{\mu}}=s^kx^k\left[\frac{p^k}{1-(\mu+1)p^{\mu}qy^{\mu}}-\frac{p^{k+1}}{1-(\mu+1)p^{\mu}qy^{\mu}}\cdot y\right].$$

When the right hand side of (37) is expanded with the help of (4) and (23), the coefficient of $s^{k+(\mu+1)i}(p^{\mu}q)^n\binom{(\mu+1)n}{n}$ would give rise to the required result (30).

The proof for (VIII) is also similar by keeping (20) and the theorem in mind.

Derivation of distribution of statistics corresponding to $N_n^*(r)$, D_n and L_n in [1], does not seem to be simple.

REFERENCES

[1] DWASS, M.: Simple random walk and rank order statistics, *Ann. Math. Statistics,* **38** (1967), 1042—53.

[2] FELLER, W.: *An introduction to probability theory and its applications* (2nd Edit) Wiley, New York (1957).

[3] GOULD, H. W.: Some generalization of Vondermonde's convolution. *Amer. Math. Monthly* **63** (1956), 84—91.

[4] MOHANTY, S. G. and NARAYANA, T. V.: Some properties of compositions and their application to probability theory-1, *Biometrische Zeitschrift,* **3** (1961), 252—58.

McMaster University, Hamilton, Canada; *Indian Institute of Technology, New Delhi*

(Received November 8, 1968.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 277—281.

NOTES ON THE RATE OF CONVERGENCE OF THE INFORMATION PROVIDED BY AN EXPERIMENT

by

T. NEMETZ

1 §. Introduction

Let us consider a statistical space $\{X, \mathscr{B}, \mathsf{P}_\Theta\}$, where Θ is an abstract set (the parameter space), $\{X, \mathscr{B}\}$ is a measurable space, and $\mathsf{P}_\Theta = \{\mathsf{P}_\vartheta, \vartheta \in \Theta\}$ is a set of probability measures on $\mathscr{B}$. Let $\mathscr{T}$ denote a σ-algebra, containing the subsets of Θ. In this paper we shall assume, that $\Theta = \{\vartheta_1, \ldots, \vartheta_r\}$ is finite and $\mathscr{T}$ contains all subsets of Θ. There are given probability measures P resp. Q defined on $\mathscr{B}$ resp. $\mathscr{T}$, so that P_ϑ is absolutely continuous ($\mathsf{P}_\vartheta \ll \mathsf{P}$) with respect to P for all $\vartheta \in \Theta$. Without loss of generality we can assume that $\mathsf{Q}(\vartheta_i) = q_i > 0$, $i = 1, \ldots, r$. For the simplicity we introduce the notations $\mathsf{P}_i = \mathsf{P}_{\vartheta i}$ and $p_i = \frac{d\mathsf{P}_i}{d\mathsf{P}}$, where p_i's are Radon-Nikodym derivatives.

The ordered quadruple $\mathscr{E} = \{X, \mathscr{B}, \Theta, \mathsf{P}_\Theta\}$ characterizes an experiment $\mathscr{E}$, which will result in an observation $x \in X$. The measure of information provided by an experiment is defined as follows (see LINDLEY [2]);

DEFINITION 1:

$$\mathscr{I}(\mathscr{E}) = \mathscr{I}(\mathscr{E}, \mathsf{Q}) = \mathsf{E}\{\mathscr{I}_0 - \mathscr{I}_1(x)\} \tag{1.1}$$

where $\mathscr{I}_0$ is the Shannon-entropy of parameter

$$\mathscr{I}_0 = \sum_{i=1}^{r} q_i \log \frac{1}{q_i}$$

and $\mathscr{I}_1(x)$ is the posterior information of parameter, having the observed value x

$$\mathscr{I}_1(x) = -\sum_{i=1}^{r} p(\vartheta_i | x) \log p(\vartheta_i | x)$$

($\mathsf{E}\{.\}$ denotes the expectation).

The measure of missing information $\mathsf{R}(\mathscr{E})$ after an experiment $\mathscr{E}$ is defined by the

DEFINITION 2:

$$\mathsf{R}(\mathscr{E}, \Theta) = \mathscr{I}_0 - \mathscr{I}(\mathscr{E}) \tag{1.2}$$

The following variant of a RÉNYI's theorem [3] holds:

THEOREM 1:

$$\mathsf{R}(\mathscr{E}, \Theta) \leqq B \cdot \max_{i \neq j} \Lambda_{i,j} \tag{1.3}$$

where the constant B depends on the distribution of the parameter only, and $\Lambda_{i,j}$ denotes the Hellinger-distance of measures P_i and P_j, i.e.

$$\Lambda_{i,j} = \int_X \sqrt{p_i(x)\, p_j(x)}\, d\mathsf{P}. \tag{1.4}$$

I. VAJDA [5] considers the case of independent experiments. Let us give the experiments $\mathscr{E}_i = \{X_k, \mathscr{B}_k, \Theta, \mathsf{P}_{\Theta,k}\}$, $k = 1, 2, \ldots$, with same parameter space and consider the product experiments $\mathscr{E}^n = \{X^n, \mathscr{B}^n, \Theta, \mathsf{P}_\Theta^n\}$ where

$$\{X^n, \mathscr{B}^n\} = \bigotimes_{k=1}^{n} \{X_k, \mathscr{B}_k\}$$

and

$$\mathsf{P}_i^n = \bigotimes_{k=1}^{n} \mathsf{P}_{i,k}$$

His main result in the paper [5] is

THEOREM 2: *If*

$$\lim \inf_{n=1,2,\ldots} \frac{1}{n} \sum_{k=1}^{n} \Delta(\mathsf{P}_{i,k}; \mathsf{P}_{j,k}) > 0 \tag{1.5}$$

for every $i \neq j$, then there exists number $A > 0$ and $0 < \lambda < 1$ such that

$$\mathsf{R}(\mathscr{E}^n, \Theta) < A \cdot \lambda^n. \tag{1.6}$$

(Here $\Delta(\mu; \nu)$ denotes the variation distance of the measures μ and ν.)

The aim of this paper is to prove that this theorem follows from Theorem 1. We shall see this in § 3. In § 2. we are going to give a proof of Theorem 1.

2 §. Prof of Theorem 1

The proof is based on a lemma of RÉNYI [4] which we repeat now.

LEMMA R. *For arbitrary probability distribution $\{q_1, q_2, \ldots, q_N\}$ one has*

$$\sum_{k=1}^{N} q_k \log \frac{1}{q_k} \leqq C \sum_{k=2}^{N} \sqrt{q_k} \tag{2.1}$$

where the positive constant C does not depend either on N or on the distribution $\{q_1, q_2, \ldots, q_N\}$.

PROOF of lemma R. Evidently both

$$\frac{x \log \frac{1}{x}}{\sqrt{x}} \quad \text{and} \quad \frac{(1-x) \log \frac{1}{1-x}}{\sqrt{x}}$$

are continuous in the closed interval [0, 1]. Thus there exist positive numbers C_1 and C_2 such that

$$x \log \frac{1}{x} \leqq C_1 \sqrt{x}$$

and

$$(1-x) \log \frac{1}{1-x} \leqq C_2 \sqrt{x};$$

from this we obtain

$$\sum_{k=1}^{N} q_k \log \frac{1}{q_k} \leqq C_1 \cdot \sum_{k=2}^{N} \sqrt{q_k} + C_2 \sqrt{\sum_{k=2}^{N} q_k} \leqq (C_1 + C_2) \sum_{k=2}^{N} \sqrt{q_k}$$

which proves the lemma with $C = C_1 + C_2$.

Theorem 1 can be deduced from this lemma, as follows

$$\mathsf{R}(\mathscr{E}, \Theta) = \mathsf{E}\left\{\sum_{i=1}^{r} \mathsf{P}(\vartheta = \vartheta_i | x) \log \frac{1}{\mathsf{P}(\vartheta = \vartheta_i | x)}\right\} \leqq$$

$$\leqq C \sum_{k=1}^{r} \sum_{i \neq k}^{r} q_k \cdot \int \sqrt{\mathsf{P}(\vartheta = \vartheta_i | x)}\, d\mathsf{P}_k$$

Applying the inequality

$$\int \sqrt{\mathsf{P}(\vartheta = \vartheta_i | x)}\, d\mathsf{P}_k = \int \sqrt{\frac{q_i \cdot \frac{d\mathsf{P}_i}{d\mathsf{P}}}{\sum_{j=1}^{r} q_j \frac{d\mathsf{P}_j}{d\mathsf{P}}}}\, d\mathsf{P}_k \leqq \int \sqrt{\frac{q_i p_i}{q_k p_k}}\, d\mathsf{P}_k =$$

$$= \sqrt{\frac{q_i}{q_k}} \int \sqrt{p_i(x) p_k(x)}\, d\mathsf{P} = \sqrt{\frac{q_i}{q_k}}\, \Lambda_{i,k}$$

we obtain

$$\mathsf{R}(\mathscr{E}, \Theta) \leqq C \sum_{k=1}^{r} \sum_{i \neq k} \sqrt{q_i q_k} \cdot \Lambda_{i,k} \leqq B \cdot \max_{i \neq k} \Lambda_{i,k}.$$

Thus the theorem is valid with

$$B = C \cdot \sum_{k=1}^{r} \sum_{i \neq k} \sqrt{q_i q_k}.$$

We note that the theorem is clearly valid also with

$$B^* = C \cdot r(r-1)$$

and this constant does not depend on $\{q_k\}$ either.

3§. Proof of Theorem 2

For the proof we need the next simple

LEMMA: *Let* $a_1, a_2, \ldots$ *arbitrary sequence of real numbers so that* $0 \leqq a_k \leqq 1$, *and*

$$\liminf_{(n)} \frac{1}{n} \sum_{k=1}^{n} a_k \geqq c > 0 \tag{3.1}$$

Then in any case of $0 < c_1 < c$ *there exist values* $\alpha = \alpha(c_1) > 0$ *and* $N(c_1)$ *such that the number of* $a_k's$, $k = 1, 2, \ldots, n$, *for which* $a_k > c_1$, *is greater than* $\alpha \cdot n$, *provided* $n > N(c_1)$.

PROOF of the lemma: Let us consider a number c^* with $c > c^* > c_1$, and let us make the new sequence of b_k's:

$$b_k = \begin{cases} 1 & \text{if} \quad a_k > c_1 \\ c_1 & \text{if} \quad a_k \leqq c_1. \end{cases}$$

Then if $N = N(c_1, c^*)$ is large enough, for all $n > N$

$$\frac{1}{n} \sum_{k=1}^{n} b_k \geqq c^*,$$

Let n^* denote the number of a_k's, $k = 1, 2, \ldots, n$ for which $a_k > c_1$. We get

$$\frac{n^* + (n - n^*) c_1}{n} \geqq c^*.$$

So we obtain

$$\frac{n^*}{n} \geqq \frac{c^* - c_1}{1 - c_1} > 0$$

which proves the lemma with $\alpha = \dfrac{c^* - c_1}{1 - c_1}$.

Let $\lambda_{i,j}(k)$ denote the Hellinger-distance of the measures $\mathsf{P}_{i,k}$ and $\mathsf{P}_{j,k}$. It is well known that for the Hellinger-distance $\Lambda_{i,j}^n$ of P_i^n and P_j^n

$$\Lambda_{i,j}^n = \prod_{k=1}^{n} \lambda_{i,j}(k)$$

From Theorem 1 it follows

$$\mathsf{R}(\mathscr{E}^n, Q) \geqq A \cdot \max_{i \neq j} \prod_{k=1}^{n} \lambda_{i,j}(k)$$

A theorem of CSISZÁR concerning the connection of the $\mathscr{F}$-divergences and the variation-distance (see e.g. [1], theorem 2. 1) implies that there exist number $A_1 > 0$ and $1 > r_0 > 0$ such that

$$\Delta(\mathsf{P}_{i,k}; \mathsf{P}_{j,k}) \leqq A_1 \cdot \sqrt{1 - \lambda_{i,j}(k)} \tag{3.2}$$

if

$$1 - \lambda_{i,j}(k) \leqq r_0$$

From this fact and from the lemma it is easy to see that there is $1 > c_2 > 0$ so that the number of k's, $k = 1, 2, \ldots, n$ for which the inequality

$$\lambda_{i,j}(k) \leqq c_2$$

holds, greater than $\alpha \cdot n$. Putting

$$\lambda^*_{i,j}(k) = \begin{cases} 1 & \text{if} \quad \lambda_{i,j}(k) > c_2 \\ c_2 & \text{if} \quad \lambda_{i,j}(k) \leqq c_2 \end{cases}$$

we obtain

$$\mathsf{R}(\mathcal{E}^n, \Theta) \leqq A \cdot \max_{i \neq j} \prod_{k=1}^{n} \lambda^*_{i,j}(k) \leqq A \cdot c_2^{\alpha \cdot n} = A \cdot \lambda^n$$

where $\lambda = c_2^{\alpha} < 1$

We remark that the assumption (1. 5) is not necessary for the validity of relation (1. 6).

REFERENCES

[1] CSISZÁR, I.: Eloszlások eltérésének információ-típusú mértékszámai, *Magyar Tud. Akad.* III. *Oszt. Közleményei,* **17** (1967), 123—149.

[2] LINDLEY, D. V.: On a measure of information provided by an experiment, *Ann. Math. Stat.* **27** (1956), 986—1005.

[3] RÉNYI, A.: On some basic problems of statistics from the point of view of information theory, *Proc. of the 5-th Berkeley Symposium.*

[4] RÉNYI, A.: Statistics based on information theory, *Proc. of European Meeting of Statisticians,* 1966.

[5] VAJDA, I.: Rate of convergence of the information in a sample concerning a parameter, *Czech. Math. Journ.* **17** (1967), 225—231.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received November 20, 1968.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 283—287.

SCHRANKEN FÜR DEN DIAMETER EINES GRAPHEN

von

F. KRAMER und H. KRAMER

Es sei $G=(X, U)$ ein ungerichteter, endlicher, zusammenhängender schlingenloser Graph, in dem jedes Knotenpunktepaar höchstens durch eine Kante verbunden ist; $X=\{x_1, x_2, \ldots, x_n\}$ sei die Knotenpunktemenge und U die Kantenmenge des Graphen G. Dem Graphen G wird folgenderweise eine reelle $n\times n$ Matrix $A=(a_{ij})$ zugeordnet:

$$a_{ij}=\begin{cases}1 & \text{falls } (x_i, x_j)\in U,\\ 0 & \text{falls } i=j \text{ ist},\\ \infty & \text{falls } (x_i, x_j)\notin U \quad \text{und} \quad i\neq j \text{ ist.}\end{cases}$$

Anschliessend führen wir sowohl für reelle Zahlen, als auch für reelle Matrizen die bekannten Operationen $\dot{+}$ und $\dot{\times}$ ein ([1] S. 132). Sind r_1, r_2 zwei reelle Zahlen, so sei:

$$r_1 \dot{+} r_2 = \min\{r_1, r_2\}$$

$$r_1 \dot{\times} r_2 = r_1 + r_2.$$

Sind $A=(a_{ij})$ und $B=(b_{ij})$ zwei reelle $n\times n$ Matrizen, so sei $A\dot{+}B = S = (s_{ij})$ ebenfalls eine reelle $n\times n$ Matrix, deren Elemente durch die Beziehung $s_{ij} = a_{ij}\dot{+}b_{ij}$ definiert sind, und $A\dot{\times}B = P = (p_{ij})$ ebenfalls eine reelle $n\times n$ Matrix, deren Elemente durch die Beziehung:

$$p_{ij} = (a_{i1}\dot{\times}b_{1j})\dot{+}(a_{i2}\dot{\times}a_{2j})\dot{+}\ldots\dot{+}(a_{in}\dot{\times}b_{nj})$$

definiert sind. Es sei: $A\dot{\times}A = A^2, \ldots, A^{k-1}\dot{\times}A = A^k$. Wir betrachten nun die Matrix $A^k=(a_{ij}^{(k)})$. Sind die beiden Knotenpunkte x_i und x_j durch einen Weg verbunden, dessen Länge $\leq k$ ist, so ist $a_{ij}^{(k)}$ gleich der Distanz $d(x_i, x_j)$, andernfalls ist $a_{ij}^{(k)}=\infty$.

Ein bekannter Satz [1] besagt, dass $A^{n-1}=A^n$ ist. Das Element $a_{ij}^{(n)}$ der Matrix A^n ist somit gleich der Distanz $d(x_i, x_j)$ d.h. es gilt $a_{ij}^{(n)}=d(x_i,x_j)$.

Der Diameter δ eines Graphen G ist durch die Beziehung

$$\delta = \max_{x_i, x_j\in X} d(x_i, x_j)$$

definiert. Aber um den Diameter eines Graphen zu berechnen (falls die Länge eines Weges gleich der Kantenanzahl des Weges genommen wird, was bei uns der Fall ist) muss die Berechnung der Matrizen A^k, $k=1, 2, \ldots$ im allgemeinen nicht bis zu der Matrix A^n durchgeführt werden. Ist k_0 die kleinste natürliche Zahl, so

dass $A^{k_0} = A^{k_0+1}$, so haben wir: $\delta = k_0$. Dann ist aber k_0 (folglich auch δ) die kleinste natürliche Zahl derart dass die Matrix A^{k_0} nur endliche Elemente besitzt.

Die Berechnung der Matrizen A^k, $k = 1, 2, \ldots, k_0$ benötigt im Falle grosser Graphen einen grossen Rechenaufwand. In der vorliegenden Arbeit wird gezeigt, wie man bei jedem Schritt der Berechnungen der Matrizen $A, A^2, A^3, \ldots$ je eine obere Schranke für den Diameter des Graphen erhalten kann. In der Arbeit [2] haben wir den nun folgenden Satz bewiesen. Nachträglich haben wir erfahren, dass die in diesem Satz gegebene Schranke für den Diameter eines Graphen bereits von J. W. MOON [3] angegeben wurde.

SATZ 1. *Es sei $G = (X, U)$ ein endlicher, schlingenloser, zusammenhängender Graph, in dem jedes Knotenpunktepaar höchstens durch eine Kante verbunden ist. Gilt für den Grad $g(x)$ eines jeden Knotenpunktes $x \in X$ die Ungleichung*

$$g(x) \geqq \left[\frac{|X|}{h}\right],$$

wobei h eine natürliche Zahl ist, $2 \leqq h \leqq \dfrac{|X|+3}{3}$, so haben wir für den Diameter δ des Graphen G die Ungleichung:

$$\delta \leqq 3h - 4.$$

Mit $|X|$ haben wir die Anzahl der Knotenpunkte des Graphen G bezeichnet, mit $[r]$ dass grösste Ganze der reellen Zahl r. Der Grad eines Knotenpunktes $x_i \in X$ ist gleich der Anzahl der Knotenpunkte, die mit x_i durch Kanten verbunden sind, und ist somit gleich der Anzahl der von Null verschiedenen endlichen Elemente der i-ten Zeile der Matrix A, die in in der obigen Weise dem Graphen G zugeordnet wurde.

DEFINITION. Wir bezeichnen mit $g_k(x_i)$ die Anzahl der Knotenpunkte y aus X für welche die Ungleichung $1 \leqq d(x_i, y) \leqq k$ gilt. Die Zahl $g_k(x_i)$ ist somit gleich der Anzahl der von Null verschiedenen, endlichen Elemente der i-ten Zeile der Matrix A^k.

Es sei $g_k = \min\limits_{i=1,2,\ldots,n} g_k(x_i)$. Dann gibt es eine natürliche Zahl h, $h \geqq 2$, so dass die Ungleichung:

$$\left[\frac{|X|}{h}\right] \leqq g_k < \left[\frac{|X|}{h-1}\right]$$

erfüllt ist.

SATZ 2. *Es sei $G = (X, U)$ ein ungerichteter, endlicher, zusammenhängender, schlingenloser Graph, in dem jedes Knotenpunktepaar höchstens durch eine Kante verbunden ist, h die kleinste natürliche Zahl, $2 \leqq h \leqq \dfrac{|X|}{2k+1} + 1$, so dass die Ungleichung:*

$$g_k \geqq \left[\frac{|X|}{h}\right]$$

erfüllt ist. Dann gilt gür den Diameter δ des Graphen G die Ungleichung:

$$\delta \leqq (2k+1)(h-1)-1.$$

BEWEIS. Wir nehmen an es sei $\delta > (2k+1)(h-1)-1$. Dann gibt es zwei Knotenpunkte x_0' und $x'_{(2k+1)(h-1)}$ so dass

$$d(x_0', x'_{(2k+1)(h-1)}) = (2k+1)(h-1)$$

ist. Es seien $x_0', x_1', x_2', \ldots, x'_{(2k+1)(h-1)}$ die Knotenpunkte auf einem der kürzesten Wege zwischen x_0' und $x'_{(2k+1)(h-1)}$. Für $i=1, 2, \ldots, h$ bezeichnen wir mit C_i die Menge der Knotenpunkte $x \in X$, für welche die Ungleichung $d(x'_{(2k+1)(i-1)}, x) \leqq k$ gilt. Laut Voraussetzung des Satzes ist für alle Knotenpunkte $x \in X$ die Ungleichung $g_k(x) \geqq g_k \geqq \left[\frac{|X|}{h}\right]$ erfüllt. Es folgt, dass jede der Mengen C_i mindestens g_k+1 verschiedene Knotenpunkte enthält und folglich mindestens $\left[\frac{|X|}{h}\right]+1$ verschiedene Knotenpunkte.

Wir beweisen nun, dass für $i \neq j$, $1 \leqq i \leqq h$, $1 \leqq j \leqq h$, $C_i \cap C_j = \emptyset$ ist. Wir nehmen an es sei für ein $i<j$ $C_i \cap C_j \neq \emptyset$. Es sei dann y ein Knotenpunkt, der den beiden Mengen C_i und C_j angehört. Es folgt

$$d(x'_{(2k+1)(i-1)}, y) \leqq k \quad \text{und} \quad d(x'_{(2k+1)(j-1)}, y) \leqq k.$$

In einem ungerichteten Graphen ist aber $d(x, y) = d(y, x)$ für jedes Knotenpunktepaar x, y. Folglich ist:

$$(1) \quad d(x'_{(2k+1)(i-1)}, x'_{(2k+1)(j-1)}) \leqq d(x'_{(2k+1)(i-1)}, y) + d(y, x'_{(2k+1)(j-1)}) \leqq 2k.$$

Anderseits befinden sich die Knotenpunkte $x'_{(2k+1)(i-1)}$ und $x'_{(2k+1)(j-1)}$ auf einem der kürzesten Wege zwischen den Knotenpunkten x_0' und $x'_{(2k+1)(h-1)}$. Es folgt $d(x'_{(2k+1)(i-1)}, x'_{(2k+1)(j-1)}) = (2k+1)(j-i) \geqq 2k+1$. Damit sind wir zu einem Widerspruch gelangt. Folglich ist $C_i \cap C_j = \emptyset$ für $i \neq j$. Die Vereinigungsmenge $\bigcup_{i=1}^{h} C_i$ enthält somit wenigstens $h\left(\left[\frac{|X|}{h}\right]+1\right)$ verschiedene Knotenpunkte. Da aber $h\left(\left[\frac{|X|}{h}\right]+1\right) > |X|$ ist, sind wir wieder zu einem Widerspruch gelangt. Damit ist die Behauptung unseres Satzes $\delta \leqq (2k+1)(h-1)-1$ bewiesen.

Wir werden nun anhand von drei Beispielen zeigen, dass die in Satz 2 gegebene obere Schranke für den Diameter eines Graphen genau ist in dem Sinne, dass für die drei folgenden Graphen die Voraussetzungen unseres Satzes erfüllt sind und für ihren Diameter die Gleichheit $\delta = (2k+1)(h-1)-1$ gilt.

Beispiel 1. Der Graph G bestehe allein aus dem folgenden Weg

$$a_1, (a_1, a_2), a_2, (a_2, a_3), a_3, \ldots, a_{2k}, (a_{2k}, a_{2k+1}), a_{2k+1}.$$

Es ist für diesen Graph $|X| = 2k+1$, $g_k = g_k(a_1) = g_k(a_{(2k+1)}) = k$, $\delta = 2k$. Es gilt somit für $h=2$ die Ungleichung $g_k \geqq \left[\frac{|X|}{h}\right]$ und es ist $\delta = (2k+1)(h-1)-1$.

Beispiel 2. Der Graph G bestehe allein aus dem Kreis:

$$a_1, (a_1, a_2), a_2, (a_2, a_3), a_3, \ldots, a_{4k-1}, (a_{4k-1}, a_{4k}), a_{4k}, (a_{4k}, a_1), a_1.$$

Es ist $|X| = 4k$, $g_k(a_i) = 2k$ $i = 1, 2, \ldots, 4k$, folglich $g_k = 2k$, und $\delta = 2k$. Für $h = 2$ sind die Voraussetzungen von Satz 2 erfüllt und es ist $\delta = (2k+1)(h-1)-1$. Es sei noch bemerkt, dass dieser Graph keine Schnittpunkte besitzt.

Beispiel 3. Es sei h eine beliebige natürliche Zahl, $h \geqq 3$. Der Graph G (Abb. 1) bestehe aus den beiden Kreisen

$$a_1, (a_1, a_2), a_2, (a_2, a_3), a_3, \ldots, a_{2k}, (a_{2k}, a_{2k+1}), a_{2k+1}, (a_{2k+1}, a_1), a_1$$

und

$$c_1, (c_1, c_2), c_2, (c_2, c_3), c_3, \ldots, c_{2k}, (c_{2k}, c_{2k+1}), c_{2k+1}, (c_{2k+1}, c_1), c_1$$

sowie aus dem Weg

$$a_{k+1}, (a_{k+1}, b_1), b_1, (b_1, b_2), b_2, \ldots, b_{(2k+1)(h-2)-1}, (b_{(2k+1)(h-2)-1}, c_{k+1}), c_{k+1}.$$

Es ist $|X| = 2kh + h - 1$, $g_k = g_k(a_1) = g_k(c_1) = 2k$, $\left[\frac{|X|}{h}\right] = \left[\frac{2kh+h-1}{h}\right] = 2k$ und $\delta = d(a_1, c_1) = (2k+1)(h-1)-1$.

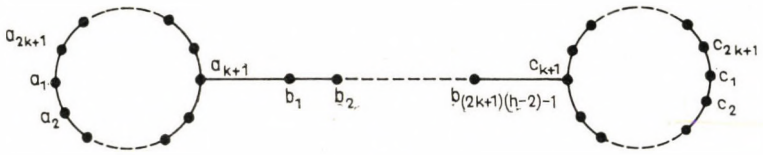


Abb. 1

Anschliessend betrachten wir dasselbe Problem für gerichtete Graphen. Zu diesem Zwecke führen wir den Begriff des k-symmetrischen Graphen ein.

DEFINITION. Ist k eine natürliche Zahl, so sagen wir der gerichtete Graph $G = (X, U)$ sei k-symmetrisch, falls für jedes Knotenpunktepaar x, y, für welches $d(x, y) \leqq k$ gilt, auch $d(y, x) \leqq k$ folgt.

Für $k = 1$ erhalten wir die Definition eines symmetrischen Graphen im gewöhnlichen Sinne. Offensichtlich ist ein stark-zusammenhängender Graph stets δ-symmetrisch, wobei δ der Diameter des Graphen ist. Ein gerichteter zusammenhängender k-symmetrischer Graph ist stark-zusammenhängend. Der in Abb. 2 dargestellte Graph ist 2-symmetrisch, ist aber nicht 1-symmetrisch.

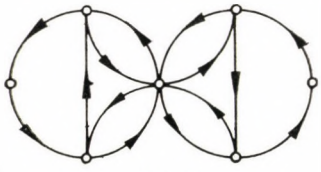


Abb. 2

Wir bezeichnen mit $g_k(x)$ wiederum die Anzahl der Knotenpunkte y aus X des gerichteten Graphen $G = (X, U)$, für welche die Ungleichung $1 \leqq d(x, y) \leqq k$ gilt, und mit $g_k = \min_{x \in X} g_k(x)$.

SATZ 3. *Gilt in einem gerichteten, endlichen, schlingenlosen, zusammenhängenden, k-symmetrischen Graphen* $G=(X, U)$ *die Ungleichung*

$$g_k \geqq \left[\frac{|X|}{h}\right], \qquad 2 \leqq h \leqq \frac{|X|}{2k+1}+1,$$

so erfüllt der Diameter δ *des Graphen* G *die Ungleichung:*

$$\delta \leqq (2k+1)(h-1)-1.$$

Der Beweis dieses Satzes verläuft ebenso wie der des Satzes 2. Der einzige Unterschied besteht in der Ableitung der Ungleichung (1). Und zwar folgt aus der Ungleichung $d(x'_{(2k+1)(j-1)}, y) \leqq k$ auf Grund der k-Symmetrie des Graphen G: $d(y, x'_{(2k+1)(j-1)}) \leqq k$ und dann die Ungleichung (1).

Die in Satz 3 erhaltene obere Schranke für den Diameter eines gerichteten Graphen ist ebenfalls genau. Um sich davon zu überzeugen braucht man nur die in Beispiel 1 bis 3 gegebenen ungerichteten Graphen in gerichtete Graphen umzuwandeln indem man jede ungerichtete Kante (a, b) durch zwei gerichtete Kanten (a, b) und (b, a) ersetzt. Auf diese Weise erhält man gerichtete symmetrische (folglich auch k-symmetrische) Graphen, für welche die Voraussetzungen von Satz 3 erfüllt sind und $\delta = (2k+1)(h-1)-1$ gilt.

LITERATUR

[1] BERGE, C.: *Théorie des graphes et ses applications.* Dunod, Paris 1963.

[2] KRAMER, F. und KRAMER, H.: Ein Färbungsproblem der Knotenpunkte eines Graphen bezüglich der Distanz p, *Revue Roumaine de Math. Pures et Appl.* **XIV** (1969) 7, 1031—1038.

[3] MOON, J. W.: On the diameter of a graph, *The Michigan Mathematical Journal* **12** (1965) 3, 349—351.

Akademie der S. R. Rumänien, Zweigstelle Cluj, Recheninstitut

(Eingegangen: 11 Dezember, 1968.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 289—298.

ON THE OPERATIONAL SOLUTION OF CERTAIN NON-LINEAR INTEGRAL EQUATIONS

by

T. FÉNYES

Introduction

In [1], the integral equation

$$(1) \qquad (t-a)f(t)-\int_0^t f(\tau)g(t-\tau)\,d\tau = b\int_0^t f(\tau)f(t-\tau)\,d\tau \qquad (0 \leqq t < \infty),\ a \geqq 0$$

has been solved by the application of the operational calculus of MIKUSIŃSKI. The operational solutions of the corresponding algebraic differential equation of the Bernoulli-type

$$(2) \qquad Df+(a+\{g\})f = -bf^2$$

were found. D denotes the symbol of the so-called algebraic derivative. The existence problem of the locally integrable solutions of (2) was also considered.

In the present paper we shall be dealing with the operational solution of non-linear integral equations of the following types:

$$(3) \qquad (t-a)f(t)+(t-a)\int_0^t f(\tau)g(t-\tau)\,d\tau = b\int_0^t f(\tau)f(t-\tau)\,d\tau, \qquad a \geqq 0,\ t \geqq 0$$

$$(4) \qquad (t+a)f(t)-\int_0^t f(\tau)g(t-\tau)\,d\tau = b\int_0^t f(\tau)f(t-\tau)\,d\tau-\frac{a}{b}g(t), \qquad a > 0,\ t \geqq 0$$

$$(5) \qquad (t+a)f(t)+(t-a)\int_0^t f(\tau)g(t-\tau)\,d\tau = b\int_0^t f(\tau)f(t-\tau)\,d\tau+\frac{a}{b}(t-a)g(t), \qquad a > 0,\ t \geqq 0$$

Here $g(t)$ is given and defined on $\langle 0, \infty)$ and is locally integrable in Lebesgue's sense, $b \neq 0$ a given real number. The unknown functions $f(t)$ are assumed to be real.

It is obvious that equations (1), (3) and (4), (5) are of quite different types. The first two equations are singular, the last two are regular ones. However, it appears the interesting circumstance that between the equations (1) and (4), moreover between (3) and (5) there is a close connection in the Mikusiński operator field when $a > 0$.

The notations and symbols of [2] will be generally used. Moreover, we consider two locally integrable functions $a(t)$, $b(t)$ to be equal, if and only if $\int_0^t a(\tau)d\tau = \int_0^t b(\tau)d\tau$ for every $t > 0$.

1. The solution of the equations (3), (5)

Let us consider the integral equation (3)

$$(1.1)\quad (t-a)f(t)+(t-a)\int_0^t f(\tau)\,g(t-\tau)\,d\tau = b\int_0^t f(\tau)f(t-\tau)\,d\tau, \qquad a \geqq 0$$

having the operational form

$$(1.2)\qquad Df(1+\{g\})+f(a+D\{g\}+a\{g\}) = -bf^2.$$

This is an algebraic Bernoulli equation and by introducing the substitution

$$u = \frac{1}{f}, \qquad f \neq 0$$

it can be reduced to the linear equation

$$(1.3)\qquad Du(1+\{g\})-(a+a\{g\}+D\{g\})u = b.$$

By Gesztelyi's theorem on algebraic integration, the general (operational) solution of (1. 3) can be determined very easily (see [3], [4]). This it is of the form

$$(1.4)\quad \begin{aligned} u &= (1+\{g\})\left[Ce^{as}-\frac{b}{a}-b\left\{\frac{H(t)}{t+a}\right\}\right], \quad \text{if} \quad a>0,\\ u &= (1+\{g\})\left[C+bs-bs\left\{\int_\varepsilon^t \frac{H(\tau)}{\tau}\,d\tau\right\}\right], \quad \text{if} \quad a=0. \end{aligned}$$

Here $\varepsilon>0$ is arbitrary, C is an arbitrary number (we assume it to be real) and $\{H(t)\}$ is defined by

$$(1.5)\qquad \frac{1}{(1+\{g\})^2} = \sum_{\nu=0}^{\infty}\binom{-2}{\nu}\{g\}^\nu = 1+\{H(t)\}.$$

The general (nontrivial) solution of (1. 3) is

$$(1.6)\qquad f=\frac{1}{u}=\begin{cases} \dfrac{1}{(1+\{g\})\left[Ce^{as}-\dfrac{b}{a}-b\left\{\dfrac{H(t)}{t+a}\right\}\right]}, & \text{if} \quad a>0,\\[2ex] \dfrac{1}{(1+\{g\})\left[C+bs-bs\left\{\displaystyle\int_\varepsilon^t \frac{H(\tau)}{\tau}\,d\tau\right\}\right]}, & \text{if} \quad a=0. \end{cases}$$

The operator (1. 6) is locally integrable function if and only if $a=0$. This holds for every value of $-\infty < C < \infty$ as we shall show it as follows.

If for $a>0$ the operator (1. 6) were a function then the following identity would hold:

$$[\{f\}+\{f\}\{g\}]\left[C-\frac{b}{a}e^{-as}-be^{-as}\left\{\frac{H(t)}{t+a}\right\}\right]=e^{-as}$$

and it would be obtained

$$C\{f\}-\frac{b}{a}e^{-as}\{f\}-be^{-as}\{f\}\left\{\frac{H(t)}{t+a}\right\}+C\{f\}\{g\}-$$

(1. 7)

$$-\frac{b}{a}e^{-as}\{f\}\{g\}-be^{-as}\{f\}\{g\}\left\{\frac{H(t)}{t+a}\right\}=e^{-as}.$$

This relation is impossible since on the left of (1. 7) stands a locally integrable function and on the right the translation operator. Thus (1. 6) is not a function for $a>0$.

Introducing the function $\{K(t)\}$ by the formula

$$\frac{1}{1+\{g\}}=\sum_{\nu=0}^{\infty}(-1)^{\nu}\{g\}^{\nu}=1+\{K(t)\}, \tag{1. 8}$$

in the case $a=0$ (1. 6) reads as

$$f=\frac{1+\{K(t)\}}{C+bs-bs\left\{\int_{\varepsilon}^{t}\frac{H(\tau)}{\tau}\,d\tau\right\}}=\frac{\left\{\frac{1}{b}+\frac{1}{b}\int_{0}^{t}K(\tau)\,d\tau\right\}}{1+\left\{\frac{C}{b}-\int_{\varepsilon}^{t}\frac{H(\tau)}{\tau}\,d\tau\right\}}=$$

(1. 9)

$$=\left\{\frac{1}{b}+\frac{1}{b}\int_{0}^{t}K(\tau)\,d\tau\right\}\sum_{\nu=0}^{\infty}(-1)^{\nu}\left\{\frac{C}{b}-\int_{\varepsilon}^{t}\frac{H(\tau)}{\tau}\,d\tau\right\}^{\nu}.$$

Thus (1. 6) is a function for $a=0$.

Now let us discuss the integral equation (5)

(1. 10)

$$(t+a)f(t)+(t-a)\int_{0}^{t}f(\tau)g(t-\tau)\,d\tau=b\int_{0}^{t}f(\tau)f(t-\tau)\,d\tau+\frac{a}{b}(t-a)g(t),\qquad a>0$$

having the operational form

$$Df(1+\{g\})+f(-a+D\{g\}+a\{g\})=-bf^{2}+\frac{a}{b}D\{g\}+\frac{a^{2}}{b}\{g\}. \tag{1. 11}$$

This is an algebraic Riccati equation and is very easily solvable in the operator

field. Observe that the number $q=\frac{a}{b}$ is a particular solution of (1. 11). By introducing the substitution

$$f=F+\frac{a}{b} \tag{1.12}$$

we get

$$DF(1+\{g\})+\left(F+\frac{a}{b}\right)(-a+D\{g\}+a\{g\})=-b\left(F+\frac{a}{b}\right)^2+\frac{a}{b}D\{g\}+\frac{a^2}{b}\{g\},$$

which leads to the following Bernoulli equation

$$DF(1+\{g\})+F(a+D\{g\}+a\{g\})=-bF^2, \tag{1.13}$$

which is the same as (1. 2). In this way the integral equations (1. 1) and (1. 10), which are of entirely different types in the classical analysis if $a>0$, can be regarded as "similar" in the operational sense i.e. the general solution of (1. 11) is different from that of (1. 2) only by the number $q=\frac{a}{b}$.

So with (1. 6) we obtain every solution of (1. 11) as follows:

$$\begin{aligned} &f=\frac{a}{b}, \\ &f=\frac{a}{b}+\frac{1}{(1+\{g\})\left(Ce^{as}-\frac{b}{a}-b\left\{\frac{H(t)}{t+a}\right\}\right)}. \end{aligned} \qquad (a>0) \tag{1.14}$$

It will be shown that the operator (1. 14) is locally integrable function if and only it $C=0$. (The particular solution $q=\frac{a}{b}$ is not a function). For $C=0$ (1. 14) reads as

$$\begin{aligned} f&=\frac{a}{b}-\frac{1}{(1+\{g\})\left(\frac{b}{a}+b\left\{\frac{H(t)}{t+a}\right\}\right)}=\frac{a}{b}-\frac{a(1+\{K(t)\})}{b\left(1+\left\{\frac{aH(t)}{t+a}\right\}\right)}=\frac{a}{b}\left[1-\frac{1+\{K(t)\}}{1+\left\{\frac{aH(t)}{t+a}\right\}}\right]= \\ &=\frac{a}{b}\left[\frac{\left\{\frac{aH(t)}{t+a}-K(t)\right\}}{1+\left\{\frac{aH(t)}{t+a}\right\}}\right]=\frac{a}{b}\left\{\frac{aH(t)}{t+a}-K(t)\right\}\sum_{\nu=0}^{\infty}(-1)^{\nu}a^{\nu}\left\{\frac{H(t)}{t+a}\right\}^{\nu}. \end{aligned} \tag{1.15}$$

Thus (1. 15) is locally integrable.

Now let $C \neq 0$. If (1. 14) were a function $\left(f \neq \frac{a}{b}\right)$, then the following identity would hold:

$$(\{f\}+\{f\}\{g\})\left(C-\frac{b}{a}e^{-as}-be^{-as}\left\{\frac{H(t)}{t+a}\right\}\right)=$$

$$=\left(\frac{a}{b}+\frac{a}{b}\{g\}\right)\left(C-\frac{b}{a}e^{-as}-be^{-as}\left\{\frac{H(t)}{t+a}\right\}\right)+e^{-as},$$

whence

$$C\{f\}+C\{f\}\{g\}-\frac{b}{a}e^{-as}\{f\}-\frac{b}{a}e^{-as}\{f\}\{g\}-$$

$$-be^{-as}\{f\}\left\{\frac{H(t)}{t+a}\right\}-be^{-as}\{f\}\{g\}\left\{\frac{H(t)}{t+a}\right\}= \tag{1.16}$$

$$=\frac{a}{b}C-e^{-as}\left\{\frac{aH(t)}{t+a}\right\}+\frac{a}{b}C\{g\}-e^{-as}\{g\}-e^{-as}\{g\}\left\{\frac{aH}{t+a}\right\}.$$

However the obtained relation is impossible since this can be written briefly as

$$\{M(t)\}=\frac{a}{b}C+\{N(t)\},$$

where $\{M(t)\}$, $\{N(t)\}$ are locally integrable and the number $\frac{a}{b}C$ does not vanish!

It holds the following

THEOREM 1. *Let $g(t)$ be locally integrable on $\langle 0, \infty)$, $b \neq 0$ real, $a \geqq 0$ in the case of* (1. 1), *$a > 0$ in the case of* (1. 10).

The integral equation (1. 1) *has the operational form* (1. 2) *being an algebraic differential equation of Bernoulli-type.* (1. 2) *has a family of nontrivial solutions depending on an arbitrary constant given by the formulas* (1. 6), (1. 9) *by the aid of* (1. 5), (1. 8). *These solutions are locally integrable solutions of* (1. 1) *if and only if $a=0$. So* (1. 1) *and* (1. 2) *are equivalent if and only if $a=0$.*

The integral equation (1. 10) *has the operational form* (1. 11) *which is an algebraic Riccati equation. This Riccati equation has a particular solution $q=\frac{a}{b}$. The other solutions of* (1. 11) *can be written as the sum of $\frac{a}{b}$ and the nontrivial solutions of the Bernoulli equation* (1. 2). *The integral equation* (1. 10) *has one and only one locally integrable solution given by* (1. 15).

2. The solution of the eqation (4)

Between the integral equations (1) and (4) there is the same operational connection as between (3) and (5). The integral equation

$$(2.1)\quad (t+a)f(t)-\int_0^t f(\tau)\,g(t-\tau)\,d\tau = b\int_0^t f(\tau)f(t-\tau)\,d\tau-\frac{a}{b}g(t), \qquad a>0$$

has the operational form

$$(2.2)\qquad Df+f(-a+\{g\}) = -bf^2+\left\{\frac{a}{b}g(t)\right\}.$$

This Riccati equation has the particular solution $q=\frac{a}{b}$. By introducing the substitution

$$(2.3)\qquad f = F+\frac{a}{b}$$

we get

$$DF+\left(F+\frac{a}{b}\right)(-a+\{g\}) = -b\left(F+\frac{a}{b}\right)^2+\left\{\frac{a}{b}g(t)\right\},$$

whence

$$(2.4)\qquad DF+(a+\{g\})F = -bF^2.$$

However (2. 4) is the operational form of the integral equation

$$(t-a)F(t)-\int_0^t F(\tau)\,g(t-\tau)\,d\tau = b\int_0^t F(\tau)F(t-\tau)\,d\tau$$

detailed in [1] (see also the introduction of this paper). By the formula (3. 3) of [1] and (2. 3), the general (operational) solution of (2. 2) reads as follows:

$$(2.5)\qquad f=\frac{a}{b}, \qquad (a>0)$$

$$f=\frac{a}{b}+$$

$$+\frac{e^{-as}s^{-g(0)}(1+\{G\})}{C+be^{-as}s^{-g(0)}\left[\left\{g(0)a^{g(0)-1}(t+a)^{-g(0)-1}-\frac{G(t)}{t+a}+g(0)(t+a)^{-g(0)-1}\int_0^t\frac{G(\tau)\,d\tau}{(\tau+a)^{1-g(0)}}\right\}-\frac{1}{a}\right]}$$

provided that $\lim_{t\to+0} g(t)=g(0)$ exists and

$$\frac{g(t)-g(0)}{t}$$

is locally integrable (see [1]). $\{G(t)\}$ is defined by

$$\{G(t)\} = \sum_{k=1}^{\infty} \frac{1}{k!} \left\{ \frac{g(t)-g(0)}{t} \right\}^k.$$

Similarly to the case treated in the preceding paragraph, it can be shown that (2. 5) is locally integrable function if and only if $C=0$. (The particular solution $f=\frac{a}{b}$ is not a function.)

For $C=0$ (2. 5) reads as

$$f=\frac{a}{b}\left[1-\frac{1+\{G\}}{1+\left\{\frac{aG(t)}{t+a}-g(0)\,a^{g(0)}(t+a)^{-g(0)-1}-ag(0)(t+a)^{-g(0)-1}\int_0^t \frac{G(\tau)\,d\tau}{(\tau+a)^{1-g(0)}}\right\}}\right],$$

and after simple calculation it can be written as

(2. 6)

$$f=-\frac{a}{b}\left\{g(0)\,a^{g(0)}(t+a)^{-g(0)-1}+\right.$$

$$\left.+\frac{tG(t)}{t+a}+ag(0)(t+a)^{-g(0)-1}\int_0^t \frac{G(\tau)\,d\tau}{(\tau+a)^{1-g(0)}}\right\}\sum_{\nu=0}^{\infty}(-1)^{\nu}\left\{-g(0)a^{g(0)}(t+a)^{-g(0)-1}+\right.$$

$$\left.+\frac{aG(t)}{t+a}-ag(0)(t+a)^{-g(0)-1}\int_0^t \frac{G(\tau)\,d\tau}{(\tau+a)^{1-g(0)}}\right\}^{\nu}.$$

Thus (2. 6) is locally integrable. For $C\neq 0$, $g(0)>0$ (2. 5) is the sum of the number $\frac{a}{b}$ and a locally integrable function (see also Theorem 2 of [1]). Consequently, the operator (2. 5) itself cannot be a function for $C\neq 0$, $g(0)>0$.

If $C\neq 0$, $g(0)\leqq 0$ we write briefly (2. 5) in the following form:

$$f=\frac{a}{b}+\frac{e^{-as}s^{-g(0)}(1+\{G\})}{C+e^{-as}s^{-g(0)}\left(\{\varrho(t)\}-\frac{b}{a}\right)}$$

where $\varrho(t)$ is a certain function. Multiplying both sides of this equation by the operator $s^{g(0)}$ the formula

$$fs^{g(0)}=\frac{a}{b}s^{g(0)}+\frac{e^{-as}(1+\{G\})}{C+e^{-as}s^{-g(0)}\left(\{\varrho(t)\}-\frac{b}{a}\right)}$$

is obtained and by repeated multiplication with

$$C+e^{-as}s^{-g(0)}\left(\{\varrho(t)\}-\frac{b}{a}\right)$$

we get the following relation

$$(2.7)\qquad Cfs^{g(0)}+e^{-as}f\{\varrho(t)\}-\frac{b}{a}e^{-as}f=C\frac{a}{b}s^{g(0)}+\frac{a}{b}e^{-as}\{\varrho\}+e^{-as}\{G\}.$$

Let f be locally integrable.

We distinguish the cases $g(0)=0$ and $g(0)<0$. For $g(0)=0$ we obtain from (2.7)

$$(2.8)\qquad C\{f\}+e^{-as}\{f\}\{\varrho\}-\frac{b}{a}e^{-as}\{f\}=C\frac{a}{b}+\frac{a}{b}e^{-as}\{\varrho(t)\}+e^{-as}\{G\}$$

which is impossible since every term of (2.8) is a function except the nonvanishing number $C\frac{a}{b}$.

Taking into account the meaning of the operator e^{-as} and the formula

$$s^{g(0)}=\left\{\frac{t^{-g(0)-1}}{\Gamma(-g(0))}\right\},\quad g(0)<0,$$

from (2.7) it would follow

$$(2.9)\qquad \int_0^t f(\tau)\frac{(t-\tau)^{-g(0)-1}}{\Gamma[-g(0)]}d\tau=\frac{at^{-g(0)-1}}{b\Gamma[-g(0)]}$$

in the interval $0\leqq t\leqq a$. However, a Volterra integral equation cannot have an eigen solution. This is a contradiction.

Thus (2.5) is not locally integrable for $C\neq 0$. Thus we obtain

THEOREM 2. *Let $g(t)$ be locally integrable on $\langle 0,\infty)$, $b\neq 0$ real, $a>0$.*

The integral equation (2.1) *has the operational form* (2.2) *being an algebraic Riccati equation. It has the particular solution $q=\frac{a}{b}$. The other solutions of* (2.2) *can be written as the sum of $\frac{a}{b}$ and the nontrivial solutions of the Bernoulli equation* (2.4). *The integral equation* (2.1) *has one and only one locally integrable solution given by* (2.6).

As an example let us solve

$$(2.10)\qquad (t+1)f(t)+\int_0^t f(\tau)e^{t-\tau}d\tau=\int_0^t f(\tau)f(t-\tau)d\tau+e^t$$

(Here $g(t)=-e^t$, $a=b=1$). (2.10) has the operational form

$$(2.11)\qquad Df-\frac{s}{s-1}f=-f^2-\frac{1}{s-1}.$$

We know that the number $q = \frac{a}{b} = 1$ is a particular solution of (2. 11). By introducing the substitution

$$u = \frac{1}{f-1}, \qquad f = \frac{1}{u} + 1$$

we obtain the linear equation

$$Du - \frac{s-2}{s-1} u = 1. \tag{2.12}$$

The general solution of the homogeneous equation is

$$u = C \exp\left[\int \frac{s-2}{s-1} ds\right] = C \exp \int \left[1 - \frac{1}{s-1}\right] ds = Ce^s e^{-\log(s-1)} = C \frac{e^s}{s-1}.$$

It will be determined by the method of variation of parameters a particular solution of the inhomogeneous equation (2. 12).

$$u_{\text{part}} = C(s) \frac{e^s}{s-1},$$

where

$$C(s) = \int (s-1) e^{-s} ds = -se^{-s}$$

and

$$u_{\text{part}} = -\frac{s}{s-1}.$$

The general solution of (2. 12) is

$$u = \frac{Ce^s - s}{s-1},$$

and that of (2. 11)

$$f = 1,$$

(2. 13)

$$f = \frac{1}{u} + 1 = \frac{s-1}{Ce^s - s} + 1.$$

The only locally integrable solution of (2. 10) is obtained from (2. 13) by the choice $C = 0$

$$f = \frac{s-1}{-s} + 1 = -1 + \frac{1}{s} + 1 = \frac{1}{s} = \{1\}.$$

Remarks

1. The existence formulas (1. 9), (1. 15), (2. 6) of the locally integrable solutions of (3), (4), (5) are very complicated. Sometimes in the solution process these complicated formulas are omitted. Occasionally it is more convenient to begin with the operational notations and express $g(t)$ as a function of the operator s.

In this manner in many cases the solutions appear in a very simple closed form (see the above example).

On the other hand, it is already known that the Laplace transformation method is a part of MIKUSIŃSKI's operational calculus [5]. By the use of the inverse transformation based on the FOURIER—MELLIN theorem, useful solution formulae can be obtained.

The crucial advantage of the MIKUSIŃSKI's operational method lies in its generality and simplicity.

2. We have assumed $a \geqq 0$ in the investigation of the equation (3). However the case $a < 0$ is not interesting because it is obvious that if $a < 0$, (3) has only the trivial solution (see also [1]).

On the other hand it is required to make a further investigation of the equations (4), (5) when $a < 0$.

REFERENCES

[1] FÉNYES, T.: On the operational solution of certain non-linear singular integral equations, *Studia Sci. Math. Hung.* **4** (1969), 69—91.

[2] MIKUSIŃSKI, J.: *Operational alculus*, Pergamon Press-Panstwowe 1959.

[3] GESZTELYI, E.: Anwendung der Operatorenrechnung auf lineare Differentialgleichungen mit Polynom-Koeffizienten, *Publicationes Mathematicae.* **10** (1963), 215—243.

[4] FÉNYES, T.: A note on the solution of integral equations of convolution type of the third kind by application of the operational calculus of Mikusiński, *Studia Sci. Math. Hung.* **2** (1967), 81—89.

[5] BERG, L.: *Einführung in die Operatorenrechnung*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1965.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received December 16, 1968.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 299—301.

ON REMOVING A POINT OF A DIGRAPH

by

B. MANVEL, P. K. STOCKMEYER and D. J. A. WELSH

In this note we study the effect which removing a point or a line from a digraph has on its connective class, thus extending some results of HARARY and ROSS [2] and HARARY, NORMAN, and CARTWRIGHT [1].

The following terminology is taken from [2]. If an *arc* (directed line) goes from u to v in a digraph D then u is *adjacent to* v and v is *adjacent from* u. The *out-degree*, od (v), of a point v is the number of points adjacent from v, the *indegree*, id (v), is the number of points adjacent to v, and the *total degree*, td (v), is id $(v)+$ $+$ od (v). For any point v of D, the digraph $D-v$ is obtained by removing v, and all arcs adjacent to or from v, from the digraph D. A *walk* is an alternating sequence of points and arcs, $v_0, x_1, v_1, \ldots, x_n, v_n$, in which each arc x_i is from v_{i-1} to v_i. A walk is *spanning* if it contains all the points of D and is *closed* if $v_0 = v_n$. If there is a walk in D from u to v then v is said to be *reachable* from u. If either v is reachable from u or u is reachable from v, u and v are said to *communicate* in D. A digraph is *strong* if every pair of points are mutually reachable and is *unilateral* if every pair of points communicate. For any digraph D the *underlying graph* $G(D)$ is the graph having the same point set as D with points u and v joined by a line in $G(F)$ if u is adjacent to or from v in D. The digraph D is *weak* if $G(D)$ is a connected graph. It is easy to see that D is strong if and only if it has a spanning closed walk and is unilateral if and only if it has a spanning walk [1, p. 66].

Clearly, if D is weak there is a point v (in fact, two points) such that $D-v$ is weak, since every graph has at least two non-cut points. The following result is slightly less obvious.

THEOREM 1. *If D is unilateral there is a point v such that $D-v$ is unilateral.*

PROOF. Let $v_0, x_1, v_1, \ldots, x_n, v_n$ be a spanning walk of D which is minimal with respect to its number of arcs. Then clearly $v_n \neq v_i$, $i<n$. Thus $v_0, x_1, v_1, \ldots,$ x_{n-1}, v_{n-1} is a spanning walk of $D-v_n$, which is therefore unilateral.

Note that in the above proof $D-v_0$ is also unilateral, so that any unilateral digraph has at least two points whose removal results in a unilateral digraph.

By considering digraphs consisting of a closed spanning walk through distinct points we see that a corresponding result does not hold for arbitrary strong digraphs. However, it is tempting to conjecture that if D is strong and $G(D)$ has high enough point connectivity there exists a point v such that $D-v$ is strong. In fact this is not the case.

THEOREM 2. *For arbitrary $n>0$, there exists a strong digraph D such that $G(D)$ is n-connected but $D-v$ is not strong for any point v.*

PROOF. The complete bipartite graph $K_{n,n}$ on two sets U and V of n points each has lines from every point of U to every point of V, and is n-connected as a graph. If its points are labeled $u_1, u_2, \ldots, u_n$ and $v_1, v_2, \ldots, v_n$, then orienting the lines $u_i v_i$ from U to V and the lines $u_i v_j$, $i \neq j$, from V to U produces a digraph with the desired property. This orientation of $K_{3,3}$ is illustrated in Figure 1.

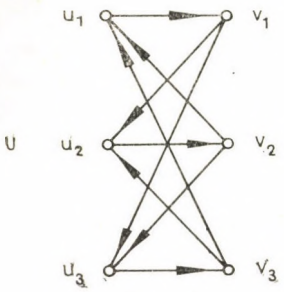


Figure 1

From Theorem 2 it follows that for arbitrary n there exists a strong digraph D with the total degree of each vertex not less than n but $D-v$ not strong for any v. In view of this the following mild restrictions on D are interesting.

THEOREM 3. *If D is strong and either for every point v, $\operatorname{od}(v) \geq 2$, or for every v, $\operatorname{id}(v) \geq 2$, then there exists a point v such that $D-v$ is strong.*

PROOF. Let D' be a maximal strong proper subdigraph of D, and let W be a shortest walk beginning and ending in D' and containing a point of $D-D'$. Such a walk exists because D is strong and D' is proper. Clearly W must contain all points of $D-D'$, for otherwise we would have a larger strong proper subdigraph. If there is only one point in $D-D'$ we are done; if not, let v_1 be the first point of W not in D', and let v_2 be the last such point of W. Say every point has outdegree at least 2, so there is another arc out of v_1. But this is impossible since if it went to a point of D' it would contradict the maximality of D', and if it went to a point of W it would contradict the minimality of W. Similarly, if $\operatorname{id}(v) \geq 2$, for all v in D, we get a contradiction at point v_2. Thus $D-D'$ must have only one point, and deletion of this point yields a strong subdigraph.

If e is an arc of a digraph D then the digraph $D-e$ is defined in the obvious way. The following theorem describes what may happen in deleting an arc.

THEOREM 4[1]. (a) *If D is a weak digraph then $D-e$ is weak for some arc e of D if and only if $G(D)$ contains a cycle.*

(b) *If D is a unilateral digraph and for each point v, $\operatorname{td}(v) \geq 2$, then there is an arc e such that $D-e$ is unilateral.*

(c) *If D is strong and for every point v, $\operatorname{td}(v) \geq 3$, there exits an edge e such that $D-e$ is strong.*

PROOF. Statement (a) is obvious. To prove (b), let $W = v_0, x_1, v_1, \ldots, x_n, v_n$ be a spanning walk of D which is minimal with respect to its number of arcs. By virtue of its minimality W passes through v_n exactly once. Taking e to be the arc of D which has v_n as an endpoint and is not x_n proves (b). To prove (c), for each arc e_i of D let X_i denote a strong component of $D-e_i$ which is maximal with respect to its number of points. Choose e_1 so that the number of points in X_1 is not less than the number of points of X_i, $(i \geq 2)$. Suppose u is a point of $D-e_1$ which is not a point of X_1. Let W_1 denote a shortest walk in D from X_1 to u and let W_2 be such a walk from u to X_1. Clearly the arc e_1 must belong to one or both of W_1

[1] Part (c) of this theorem was discovered independently by D. P. GELLER.

and W_2 since otherwise u would belong to X_1. Moreover since W_1 and W_2 are minimal, there is an arc $e_j \neq e_1$ of D having u as an endpoint which does not belong to either W_1 or W_2. Clearly $D-e_j$ has a strong component which contains X_1 and the point u. This contradicts the choice of e_1 and proves the theorem.

Interesting but apparently difficult probelms are to find necessary and sufficient conditions on a digraph D for each of the following to hold:

(1) D is strong and $D-v$ is strong for all points v of D.

(2) D is unilateral and $D-v$ is unilateral for all points v of D.

The following theorem gives some information about deletion of an arbitrary point.

THEOREM 5. *If D is strong (unilateral) then for every point v there exits a point $u(v) \neq v$ such that v can reach (communicate with) every point in $D-u(v)$.*

PROOF. Let D be strong and choose u_0 to be a point such that v can reach a maximal number of points in $D-u_0$. Suppose there is a point w such that there is no walk from v to w in $D-u_0$. Then every walk in D from v to w must pass through u_0. But in $D-w$, v can reach u_0 and also all points which it can reach in $D-u_0$. This contradicts the choice of u_0. A similar proof holds for unilateral digraphs.

REFERENCES

[1] HARARY, F., NORMAN, R. Z. and CARTWRIGHT, D.: *Structural Models: an introduction to the theory of directed graphs*. Wiley, New York, 1965.

[2] HARARY, F. and ROSS, I. C.: A description of strengthening and weakening members of a group. *Sociometry* **22** (1959), 139—147.

University of Michigan and Merton College, Oxford

(Received November 20, 1968.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 303—315.

SOME THEOREMS ON COVERINGS

by

M. AIGNER

I. Introduction

Recently several papers have appeared discussing covering and related problems, [6, 7, 10, 11]. In this note, we treat these topics using graph-theoretical notions. Let $S_n = \{1, 2, \ldots, n\}$ and define for $1 \leqq l \leqq r$ the following two finite undirected graphs $L^l_{r,n}$, $T^l_{r,n}$: The vertex-set of $L^l_{r,n}(T^l_{r,n})$ consists of all ordered (unordered) r-tuples of elements of S_n with repetitions (without repetitions), and two vertices in $L^l_{r,n}(T^l_{r,n})$ are adjacent iff the corresponding r-tuples have at least l coordinates (symbols) in common.

We quote three sets of problems concerning the graphs $L^l_{r,n}$ and $T^l_{r,n}$ that bear special significance for coverings and designs.

(a) Evaluation of the internal stability number α.

(b) Evaluation of the external stability number β.

(c) Characterization problems.

Of these, we propose to discuss (a) and (b) in this paper. As to (c), attention has been focused on related association schemes and uniqueness in terms of their parameters. For reference, see e. g. [1, 3, 5, 8, 12], in particular [2].

II. The internal stability number α

Let us first investigate $L^l_{r,n}$. An internally stable set M will consist of ordered r-tuples in which no ordered l-tuple appears more than once. An orthogonal array OA (r, l, λ, n) with r constraints, strength l, index λ and n levels is an $r \times \lambda \cdot n^l$-matrix where each $l \times \lambda \cdot n^l$ submatrix contains all possible $l \times 1$ column-vectors with the same frequency λ.

THEOREM 1: $\alpha(L^l_{r,n}) \leqq n^l$ *with equality iff there exists an orthogonal array* OA $(r, l, 1, n)$.

PROOF: Let M be an internally stable set of $L^l_{r,n}$ and let f denote the number of times an arbitrary ordered $(l-1)$ — tuple appears in M. Clearly $f \leqq n$. Considering all possible $(l-1)$ — tuples we obtain the inequality

$$\binom{r}{l-1} \cdot |M| \leqq \binom{r}{l-1} \cdot n^{l-1} \cdot n.$$

Equality in the above formula then means that every ordered l-tuple appears exactly once, hence the theorem.

Examination of the frequency with which an element or coordinate can appear leads to the following two bounds for $\alpha(L^l_{r,n})$.

THEOREM 2: $$\alpha(L^l_{r,n}) \leqq n \cdot \alpha(L^{l-1}_{r-1,n}),$$

$$\alpha(L^l_{r,n-1}) \geqq \alpha(L^l_{r,n}) - \sum_{i=1}^{l-1} \binom{r}{i} \alpha(L^{l-i}_{r-i,n}) - \binom{r}{l}.$$

PROOF: Given a maximal internally stable set M for $L^l_{r,n}$, it is clear that an arbitrary element can appear as, say, first coordinate at most $\alpha(L^{l-1}_{r-1,n})$ times. As to the second inequality, the quantity on the right hand side which is being subtracted plainly constitutes an upper bound for the number of r-tuples which contain a single element at least once.

Let us now evaluate $\alpha(L^l_{r,n})$ for small values of l.

$l=1$. $\alpha(L^1_{r,n})=n$, the set $(1, \ldots, 1), \ldots (n, \ldots, n)$ being a maximal internally stable set.

$l=2$. It is well known that the existence of an OA $(r, 2, 1, n)$ is equivalent to the existence of $r-2$ mutually orthogonal Latin squares of order n. Thus we have $\alpha(L^2_{r,n})=n^2$ for $r \leqq \min(p^t)+1$, where $n=\prod p^t$, and $\alpha(L^2_{r,n})<n^2$ for $r > n+1$. For $r=4$ more information is available, namely, $\alpha(L^2_{r,n})=n^2$ except for $n=2, 6$.

From the theory of orthogonal arrays (see [4]), we know that $r \leqq n+l-1$, and in case $l \geqq 3$, n odd that $r \leqq n+l-2$. Furthermore, it is known that OA $(n+1, 3, 1, n)$ exists for $n=p^t$ (p prime) and OA $(n+2, 3, 1, n)$ for $n=2^t$.

$l=3$. $\alpha(L^3_{r,n})<n^3$ for $r > n+2$ and $r > n+1$ when n is odd, $\alpha(L^3_{r,n})=n^3$ for $r \leqq n+2$ and $n=2^t$, $\alpha(L^3_{r,n})=n^3$ for $r \leqq n+1$ and $n=p^t$.

The corresponding theorems for the graph $T^l_{r,n}$ read as follows, [9, 11].

THEOREM 1*a*:

$$\alpha(T^l_{r,n}) \leqq \frac{\binom{n}{l}}{\binom{r}{l}},$$

where now $n \geqq r$, *with equality iff there exists a tactical system* TS (r, l, n).

THEOREM 2*a*:

$$\alpha(T^l_{r,n}) \leqq \left[\frac{n}{r} \cdot \alpha(T^{l-1}_{r-1,n-1})\right],$$

whence by induction

$$\alpha(T^l_{r,n}) \leqq \left[\frac{n}{r}\left[\frac{n-1}{r-1}\left[\ldots\left[\frac{n-l+1}{r-l+1}\right]\ldots\right]\right]\right],$$

$$\alpha(T^l_{r,n}) \geqq \alpha(T^l_{r,n+1}) - \alpha(T^{l-1}_{r-1,n}).$$

The known values of α for small l are:

$$l=1. \quad \alpha(T^1_{r,n}) = \left[\frac{n}{r}\right].$$

$$l=2. \quad \alpha(T^2_{r,n}) = \left[\frac{n}{r}\left[\frac{n-1}{r-1}\right]\right] \quad \text{for} \quad r=2, \quad r=3 \quad \text{and} \quad n \not\equiv 5 \mod 6,$$

$r=4$ and $n\equiv 0, 1, 3, 4 \bmod 12$, $r=5$ and $n\equiv 0, 1, 4, 5 \bmod 20$, and

$$\alpha(T^2_{r,n}) = \left[\frac{n}{r}\left[\frac{n-1}{r-1}\right]\right]-1$$

for $r=3$ and $n\equiv 5 \bmod 6$.

$$l=3. \quad \alpha(T^3_{r,n}) = \left[\frac{n}{r}\left[\frac{n-1}{r-1}\left[\frac{n-2}{r-2}\right]\right]\right] \quad \text{for} \quad r=3, \quad r=4 \quad \text{and}$$

$n\equiv 1, 2, 3, 4 \bmod 12$.

III. The number $\bar{\alpha}$

Let us define the numbers $\bar{\alpha}(L^l_{r,n})$ and $\bar{\alpha}(T^l_{r,n})$ as the minimal cardinality of a set N of ordered (unordered) r-tuples of elements taken from S_n such that each ordered (unordered) l-tuple appears at least once in N. This notion will prove, apart from its own interest, particularly important with regard to the evaluation of the external stability number β, since by definition of β, we will be concerned with minimal sets P of r-tuples such that an arbitrary r-tuple coincides in at least one l-tuple with some member of P.

It is easy to derive the following theorems analogous to theorems 1—2a.

THEOREM 3: $\bar{\alpha}(L^l_{r,n}) \geqq n^l$ *with equality iff there exists an orthogonal array* OA $(r, l, 1, n)$.

THEOREM 4: $\bar{\alpha}(L^l_{r,n}) \geqq n \cdot \bar{\alpha}(L^{l-1}_{r-1,n})$,

$$\bar{\alpha}(L^l_{r,n}) \leqq \bar{\alpha}(L^l_{r,n-1}) + \sum_{i=1}^{l-1} \binom{r}{i} \bar{\alpha}(L^{l-i}_{r-i,n-1}) + 1.$$

THEOREM 3a: $\bar{\alpha}(T^l_{r,n}) \geqq \dfrac{\binom{n}{l}}{\binom{r}{l}}$ *with equality iff there exists a tactical system* TS (r, l, n).

THEOREM 4a: $\bar{\alpha}(T^l_{r,n}) \geqq \left\{\dfrac{n}{r}\,\bar{\alpha}(T^{l-1}_{r-1,n-1})\right\}^{(1*)}$, *whence, by induction*

$$\bar{\alpha}(T^l_{r,n}) \geqq \left\{\frac{n}{r}\left\{\frac{n-1}{r-1}\left\{\cdots\left\{\frac{n-l+1}{r-l+1}\right\}\cdots\right\}\right\}\right\},$$

$$\bar{\alpha}(T^l_{r,n}) \leqq \bar{\alpha}(T^l_{r,n-1}) + \bar{\alpha}(T^{l-1}_{r-1,n-1}).$$

[1*] $\{u\}$ stands for the smallest integer not smaller than u.

Some known values, [6, 10]:

$$l = 1. \qquad \bar{\alpha}(L^1_{r,n}) = n, \quad \bar{\alpha}(T^1_{r,n}) = \left\{\frac{n}{r}\right\}.$$

$$l = 2. \quad r = 3: \bar{\alpha}(L^2_{3,n}) = n^2, \quad \bar{\alpha}(T^2_{3,n}) = \left\{\frac{n}{3}\left\{\frac{n-1}{2}\right\}\right\}.$$

$$r = 4: \bar{\alpha}(L^2_{4,n}) = n^2 \quad \text{for} \quad n \neq 2, 6.$$

$$\bar{\alpha}(T^2_{4,n}) = \left\{\frac{n}{4}\left\{\frac{n-1}{3}\right\}\right\} \quad \text{for} \quad n = 1, 2, 4, 5 \bmod 12.$$

$$l = 3. \quad r = 4: \bar{\alpha}(T^3_{4,n}) = \left\{\frac{n}{4}\left\{\frac{n-1}{3}\left\{\frac{n-2}{2}\right\}\right\}\right\} \quad n \equiv 2, 3, 4, 5 \bmod 6.$$

IV. The external stability number β

An externally stable set in an undirected graph is a set of vertices with the property that any vertex of the graph either lies in the set or is adjacent to at least one member of the set. With regard to our special graphs $L^l_{r,n}$, $T^l_{r,n}$, we are then faced with the problem of constructing sets of r-tuples which cover an arbitrary r-tuple in at least an l-tuple. Since the number of vertices adjacent to an arbitrary vertex of $L^l_{r,n}(T^l_{r,n})$ clearly is

$$\sum_{i=l}^{r-1} \binom{r}{i}(n-1)^{r-i} \quad \text{and} \quad \sum_{i=l}^{r-1} \binom{r}{i}\binom{n-r}{r-i}, \quad \text{respectively, we obtain}$$

THEOREM 5:

$$\frac{n^r}{\sum_{i=l}^{r} \binom{r}{i}(n-1)^{r-i}} \leqq \beta(L^l_{r,n}) \leqq n^l, \quad (2^*)$$

$$\frac{\binom{n}{r}}{\sum_{i=l}^{r} \binom{r}{i}\binom{n-r}{r-i}} \leqq \beta(T^l_{r,n}) \leqq \frac{\binom{n}{l}}{\binom{r}{l}}.$$

Two more inequalities analogous to Theorems 4 and 4a are spelled out in

THEOREM 6:

$$\beta(L^l_{r,n}) \leqq n\beta(L^{l-1}_{r-1,n}),$$

$$\beta(T^l_{r,n}) \leqq \beta(T^l_{r,n-1}) + \beta(T^{l-1}_{r-1,n-1}).$$

2* The inequality $\beta \leqq \alpha$ is true for all graphs.

$l = 1$. $\beta(L^1_{r,n}) = n, \beta(T^1_{r,n}) = \left[\frac{n}{r}\right]$, since for smaller values of β we cover at most $n^r - 1$ and $\binom{n}{r} - 1$ vertices, respectively.

$l = 2$.

THEOREM 7: $$\beta(L^2_{r,n}) \geqq \frac{n^2}{r-1}.$$

PROOF: We use induction on r. For $r = 2$, the theorem is trivially satisfied since no two vertices are adjacent. Let us take now a minimal externally stable set P in $L^2_{r,n}$ and let us assume $|P| \leqq \frac{n^2}{r-1}$. Let x be an element that appears with minimum frequency f as a coordinate, say the first coordinate. We then have

$$f \leqq \frac{|P|}{n} \leqq \frac{n}{r-1}.$$

Denoting by $s_2, s_3, \ldots, s_r$ the numbers of distinct elements appearing in the second, third, ..., r-th position of the f r-tuples headed by x, we obtain $s_i \leqq f$ $(i = 2, \ldots, r)$. Suppose without loss of generality $s_2 \geqq s_j$ $(j \neq 2)$ and $s_2 = \frac{n}{r-1} - t$ $(t \geqq 0)$. Since f was the minimal frequency of any coordinate, we note that the s_2 distinct elements appear in a combined total of at least $f \cdot s_2 \geqq s_2^2$ r-tuples of P. Further, any ordered $(r-1)$-tuple with first coordinate distinct from the s_2 elements from above, with second coordinate distinct from the s_3 elements, etc., must coincide in at least one ordered pair with some $(r-1)$-tuple of 2 nd, ..., r-th coordinates taken from the subset of P not containing x as first coordinate. Since we have at least $n - s_2$ i-th coordinates $(i \geqq 2)$ other than the s_i specified above, we obtain the following inequality by the induction hypothesis

$$|P| \geqq s_2^2 + \frac{(n-s_2)^2}{r-2} = \frac{n^2}{r-1} + \frac{r-1}{r-2} t^2,$$

thus proving the theorem.

COROLLARY: *Let* $n \equiv u \bmod r-1$, $0 \leqq u < r-1$ *and suppose orthogonal arrays* $\mathrm{OA}\left(r, 2, 1, \left[\frac{n}{r-1}\right]\right)$ *and* $\mathrm{OA}\left(r, 2, 1, \left\{\frac{n}{r-1}\right\}\right)$ *exist, then*

$$\beta = \left\{\frac{n^2}{r-1}\right\},$$

provided that $u(r-1-u) \leqq v$, *where* $n^2 \equiv -v \bmod r-1$, $0 \leqq v < r-1$.

PROOF: We split the set S_n into $r-1$ components, u of these containing $\frac{n-u+r-1}{r-1}$ elements, $r-1-u$ containing $\frac{n-u}{r-1}$ elements. For each one of these, we construct an orthogonal array, thus obtaining

$$u\left(\frac{n-u+r-1}{r-1}\right)^2+(r-1-u)\left(\frac{n-u}{r-1}\right)^2=\frac{n^2}{r-1}+\frac{u(r-1-u)}{r-1}\leqq\frac{n^2+v}{r-1}=$$

$$=\left\{\frac{n^2}{r-1}\right\}\ r\text{-tuples.}$$

Now, since an arbitrary r-tuple must have at least two coordinates in the same component, it follows that the hereby constructed set constitutes an externally stable set, thus establishing the corollary.

Remark: The problem of evaluating the external stability number was first formulated in group theory-language by TAUSSKY—TODD [13] for the case $L^l_{l+1,n}$, the general problem seems not to have been treated so far.

For $r\leqq p\leqq n$, let β_p be the minimal cardinality of a set P of unordered p-tuples taken from S_n, such that every r-tuple has at least a pair of elements in common with some member of P. Clearly we have $\beta_r=\beta(T^2_{r,n})$.

THEOREM 8:

$$\beta_p\geqq\frac{n(n-r+1)}{p(p-1)(r-1)}.$$

PROOF: For $r=2$, the condition of the theorem states that every element must appear with every other element in some p-tuple of P, thus the minimal frequency of an arbitrary element is at least $\frac{n-1}{p-1}$. Since every p-tuple contributes p appearances to the count, we obtain the inequality

$$\beta_p\geqq\frac{n(n-1)}{p(p-1)}.$$

Suppose the assertion is correct for $r-1$, and let P be a set of minimal cardinality for r, $|P|\leqq\frac{n(n-r+1)}{p(p-1)(r-1)}$. Let x be an element of minimal frequency f, then

$$f\leqq\frac{n-r+1}{(p-1)(r-1)}.$$

If s denotes the number of distinct elements appearing together with x in some p-tuple, we have

$$s\leqq(p-1)f\leqq\frac{n-r+1}{r-1},\quad\text{or}$$

$$s=\frac{n}{r-1}-1-t,\quad t\geqq 0,$$

$$f\geqq\frac{n-r+1-t(r-1)}{(r-1)(p-1)}.$$

Now clearly every $(r-1)$-tuple taken from the set of elements different from the s elements and x must coincide in at least one pair with some p-tuple in P, and further everyone of the s elements appears at least f times. These two counts, together with the induction hypothesis, now yield

(1)

$$|P| \geqq \frac{1}{p} \cdot (s+1) \cdot f + \frac{(n-s-1)(n-s-r+1)}{p(p-1)(r-2)} \geqq \frac{n(n-r+1)}{p(p-1)(r-1)} + \frac{(r-1)t^2}{p(p-1)(r-2)}.$$

COROLLARY: $\beta(T^2_{r,n}) \geqq \dfrac{n(n-r+1)}{r(r-1)^2}$ *with equality iff* $\dfrac{n}{r-1}$ *is an integer and a tactical system* $\mathrm{TS}\left(r, 2, \dfrac{n}{r-1}\right)$ *exists.*

PROOF: From the argument used in Theorem 8, it follows that equality in (1) implies $t = 0$, i.e., $s = \dfrac{n}{r-1} - 1 = (r-1) \cdot f$. Hence, $\dfrac{n}{r-1}$ must be an integer and by using induction on r again, the necessity-part is easily established. As to the sufficiency, we split S_n into $r-1$ components of $\dfrac{n}{r-1}$ elements each, and construct $r-1$ tactical systems $\mathrm{TS}\left(r, 2, \dfrac{n}{r-1}\right)$, thus obtaining

$$(r-1) \cdot \frac{\dfrac{n}{r-1}\left(\dfrac{n}{r-1} - 1\right)}{r(r-1)} = \frac{n(n-r+1)}{r(r-1)^2}$$

r-tuples altogether. Since every r-tuple must have at least two elements in the same component, the corollary follows.

V. $\beta(L^2_{3,n})$ and $\beta(T^2_{3,n})$.

To give an application of Theorems 7 and 8, the external stability number of $L^2_{3,n}$ and $T^2_{3,n}$ shall be evaluated, and properties of the minimal externally stable sets discussed.

THEOREM 9: $\beta(L^2_{3,n}) = \left\{\dfrac{n^2}{2}\right\}$ *and* P *is a minimal externally stable set iff it can be split into two components of* $\left[\dfrac{n}{2}\right]^2$ *and* $\left\{\dfrac{n}{2}\right\}^2$ *elements, respectively, such that the two components have no coordinates in common and form orthogonal arrays* $\mathrm{OA}\left(3, 2, 1, \left[\dfrac{n}{2}\right]\right)$, $\mathrm{OA}\left(3, 2, 1, \left\{\dfrac{n}{2}\right\}\right)$ *upon suitable relabeling of the elements.*

PROOF: Since Latin squares exist for all n, the corollary to Theorem 7 immediately establishes β. If we are given two Latin squares of orders $\left[\dfrac{n}{2}\right], \left\{\dfrac{n}{2}\right\}$, respectively, then by invoking the same corollary, we find P to be a minimal externally stable set. (The statement of the theorem simply means that, since repetitions are permitted, we may relabel the coordinates arbitrarily, making sure that the Latin

squares remain coordinate-disjoint.) Let P now be an arbitrary minimal externally stable set and let x be an element that appears with minimum frequency f as a coordinate. Using the same notation as in Theorem 7, it follows that

$$|P| \geqq f \cdot s_2 + (n - s_2)(n - s_3),$$

hence

$$f = s_2 = s_3 = \left[\frac{n}{2}\right].$$

Let us split P into two components, the first containing the s_2 and s_3 coordinates which appear together with x, the second comprising all remaining triples. The cardinalities of these two sets are then $\left[\frac{n}{2}\right]^2, \left\{\frac{n}{2}\right\}^2$ respectively, and the same minimal frequency argument applied to one of the s_2 second coordinates (appearing with frequency f) establishes the desired decomposition.

To compute $\beta(T^2_{3,n})$, it appears convenient to treat the cases n even and n odd separately.

THEOREM 10*a*: *Let n be even, then*

$$\beta(T^2_{3,n}) = \left\{\frac{n(n-2)}{12}\right\} \quad \text{for} \quad n \equiv 2, 4, 6 \bmod 12$$

(2)

$$= \left\{\frac{n(n-2)}{12}\right\} + 1 \quad n \equiv 0, 8, 10 \bmod 12.$$

Any minimal externally stable set P consists of two element-disjoint components A, B with A containing $\frac{n}{2}$, B containing $\frac{n}{2}$ elements for $n \equiv 2, 6 \bmod 12$; A containing $\frac{n}{2} - 1$, B containing $\frac{n}{2} + 1$ elements for $n \equiv 0, 4, 8 \bmod 12$; A containing $\frac{n}{2}$, B containing $\frac{n}{2}$ elements or A containing $\frac{n}{2} - 1$, B containing $\frac{n}{2} + 1$ elements for $n \equiv 10 \bmod 12$, such that A, B contain all possible pairs of their respective sets of elements at least once, and are minimal with respect to this property.

PROOF: It is a simple computation to show that the cardinality of a set P as described in the theorem attains the bound (2) by quoting the result of Section III, [6]. By the argument employed in the Corollary to Theorem 8, these sets are readily seen to be externally stable.

To give one example, suppose $n = 12k + 10$. Let us split S_n into two sets containing $\frac{n}{2}$ elements each. It then follows that

$$|A| + |B| = 2 \cdot \left\{\frac{n}{2}\left\{\frac{\frac{n}{2} - 1}{2}\right\}\right\} = 12k^2 + 18k + 8 = \left\{\frac{n(n-2)}{12}\right\} + 1.$$

It remains to verify that β can not be smaller than (2), and that every minimal set P is of the specified form. For $n \equiv 2, 6 \bmod 12$, (2) is the exact bound of Theorem 8, and so the corollary applies. Hence we may assume $n \equiv 0, 4, 8$ or $10 \bmod 12$. Let P now be such a minimal set and f the minimal frequency. Since we already know an upper bound for β, we must have $f \leqq \frac{n-2}{4}$.

Case A. $n \equiv 0 \bmod 4$.

Here we have $f \leqq \frac{n}{4} - 1$, and thus $s \leqq 2f \leqq \frac{n}{2} - 2$ (using the same notation as in Theorem 8).

Let $s = \frac{n}{2} - 2 - t \quad (t \geqq 0)$, $v = n - s - 1$, then

$$|P| \geqq \frac{1}{3}\left((s+1)f + v \cdot \left\{\frac{v-1}{2}\right\}\right), \tag{3}$$

since the expression in parantheses gives a lower bound for the sum of the frequencies.

Let us denote the subset of S_n consisting of the s elements appearing with x plus the element x S_1, the complementary set S_2. A triple of P containing elements of both S_1 and S_2 shall be called a mixed triple. (3) now becomes

$$\begin{aligned} |P| &\geqq \frac{1}{3}\left(\left(\frac{n}{2} - 1 - t\right) \cdot \left(\frac{n}{4} - 1 - \left[\frac{t}{2}\right]\right) + \left(\frac{n}{2} + 1 + t\right) \cdot \left(\frac{n}{4} + \left\{\frac{t}{2}\right\}\right)\right) = \\ &\frac{n(n-2)}{12} + \frac{2n}{12}\left(\left\{\frac{t}{2}\right\} - \left[\frac{t}{2}\right]\right) + \frac{4(t+1)}{12}\left(\left\{\frac{t}{2}\right\} + \left[\frac{t}{2}\right] + 1\right). \end{aligned} \tag{4}$$

Since for $t \geqq 1$, (4) exceeds the known bound (2), we must have $t = 0$, i.e.,

$$|P| \geqq \frac{n(n-2)}{12} + \frac{4}{12}. \tag{5}$$

For $n \equiv 4 \bmod 12$, (5) is an integer, and since we clearly cannot have any mixed triples (we would add at least two appearances), the result follows.

For $n \equiv 0, 8 \bmod 12$, we have to add $\frac{2}{3}$, i.e., two appearances to (5), in order to obtain an integer. Suppose we had mixed triples, then they are of the form $(ab'b'')$ or $(a'a''b)$ $(a \in S_1, b \in S_2)$, but since both $|S_1|$ and $|S_2|$ are odd and $|S_1| \geqq 3$ $(n \geqq 8)$, we would obtain 3 additional appearances. (The additional summand $\frac{2}{3}$ to (5) is accounted for by the fact that one of the two sets S_1 or S_2 has cardinality $C \equiv 5 \bmod 6$, and that a covering of such sets (see [6]) contains exactly two elements appearing $\frac{C+1}{2}$ times, whereas all the others appear $\frac{C-1}{2}$ times.)

Case B. $n \equiv 10 \bmod 12$.

In this case, $f \leqq \frac{n-2}{4}$, $s \leqq 2f \leqq \frac{n-2}{2}$.

With $s = \frac{n}{2} - 1 - t$, $v = n - s - 1$, we obtain

$$|P| \geqq \frac{1}{3}\left(\left(\frac{n}{2}-t\right)\left(\frac{n-2}{4}-\left[\frac{t}{2}\right]\right)+\left(\frac{n}{2}+t\right)\cdot\left(\frac{n-2}{4}+\left\{\frac{t}{2}\right\}\right)\right) =$$

$$= \frac{n(n-2)}{12}+\frac{2n}{12}\left(\left\{\frac{t}{2}\right\}-\left[\frac{t}{2}\right]\right)+\frac{4t}{12}\left(\left\{\frac{t}{2}\right\}+\left[\frac{t}{2}\right]\right). \tag{6}$$

Comparing (6) with the previously established bound (2), we readily infer $t=0$ or 2. Suppose $t=0$, then $|S_2| \equiv 5 \bmod 6$, and quoting [6] again, we see that in this case (2) cannot be improved, and by a similar argument as before no mixed triple can occur. For $t=2$, the bound (2) is attained exactly, no additional appearances are possible, hence the second possibility cited in the theorem results.

THEOREM 10*b*: *For n odd*

$$\beta(T^2_{3,n}) = \left\{\frac{n(n-1)}{12}\right\} \qquad \textit{for} \quad n \equiv 1, 3, 5, 7, 11 \bmod 12$$

$$= \left\{\frac{n(n-1)}{12}\right\}+1 \qquad n \equiv 9 \bmod 12 \tag{7}$$

Any minimal externally stable set P consists of two element-disjoint components A, B with element sets S_1 and S_2, respectively, such that A and B contain all pairs taken from their element sets at least once and are minimal with respect to this property, where

$$|S_1| = \frac{n-1}{2}, \quad |S_2| = \frac{n+1}{2} \qquad \textit{for} \quad n \equiv 1, 5, 7 \bmod 12$$

$$\left.\begin{aligned} |S_1| &= \frac{n-1}{2}, \quad |S_2| = \frac{n+1}{2} \\ \textit{or} \qquad & \\ |S_1| &= \frac{n-3}{2}, \quad |S_2| = \frac{n+3}{2} \end{aligned}\right\} \quad \textit{for} \quad n \equiv 3, 9, 11 \bmod 12. \tag{8}$$

If one of the two sets S_1, S_2 in (8) has cardinality $C \equiv 2$ or $4 \bmod 6$, P may contain one mixed triple with two elements from the set of cardinality C, one from the other.

PROOF: Sets P as specified in (8) have cardinality (7) and thus provide an upper bound for β. Using the notation of Theorem 10a, we then have

$$f \leqq \frac{n-1}{4}.$$

Case A. $n \equiv 1 \bmod 4$.

Writing $s = \frac{n-1}{2} - t$, $t \geqq 0$, $f \geqq \frac{n-1}{4} - \left[\frac{t}{2}\right]$, $v = \frac{n-1}{2} + t$, we obtain

$$|P| \geqq \frac{1}{3}\left(\left(\frac{n+1}{2} - t\right)\left(\frac{n-1}{4} - \left[\frac{t}{2}\right]\right) + \left(\frac{n-1}{2} + t\right)\left(\frac{n-1}{4} + \left[\frac{t}{2}\right]\right)\right) = \tag{9}$$

$$= \frac{n(n-1)}{12} + \frac{8t-4}{12}\left[\frac{t}{2}\right].$$

In order not to exceed bound (7), $t = 0$, 1 or 2. Since for $t = 1$, v equals the value of $s+1$ for $t = 0$ and vice versa, it suffices to consider one of the two possibilities, e.g., $t = 1$. For $n \equiv 1 \bmod 12$, (9) is an integer and, since we cannot have any mixed triples (i. e., no additional appearances), the decomposition (8) results. If $n = 12k+5$, we have $|S_1| = 6k+2$, $|S_2| = 6k+3$, and since $\left\{\frac{n(n-1)}{12}\right\} = \frac{n(n-1)}{12} + \frac{4}{12}$, we have one additional appearance at our disposal. If there are no mixed triples, we are led to (8), if we had a triple $(ab'b'')$, $a \in S_1$, $b', b'' \in S_2$, three additional appearances would occur. Suppose now we have a triple $(a'a''b)$. Since b must occur with all other elements of S_2, we have to construct a set of triples A', which contains all possible pairs of elements, taken from S_1, except $a'a''$, such that $|A'| = \frac{(n-1)^2 - 16}{24}$. This is done as follows. $S_1'' = S_1 - \{a''\}$ contains $\frac{n-3}{2} = 6k+1$ elements. We construct a Steiner triple system A'' on S_1'' and then define A' to be:

$$A' = A'' \cup (a_1, a_2, a'') \cup (a_3, a_4, a'') \cup \ldots \cup (a_{6k-1}, a_{6k}, a''),$$

where $S_1'' = \{a_1, a_2, \ldots, a_{6k}, a'\}$.

Now A' plainly covers all pairs distinct from $a'a''$, and $|A'| = \frac{(n-3)}{2} \cdot \frac{(n-5)}{12} + \frac{n-5}{4} = \frac{(n-1)^2 - 16}{24}$. Since we clearly cannot have more than one mixed triple, the desired result follows.

The case $n = 12k+9$ can be dealt with similarly, noting that $|S_1| = 6k+4$, $|S_2| = 6k+5$ and that, since neither S_1 nor S_2 permits a Steiner triple system, the bound (7) cannot be improved.

For $t = 2$, (9) becomes $\frac{n(n-1)}{12} + 1 > \left\{\frac{n(n-1)}{12}\right\}$ hence we only have to consider the case $n = 12k+9$. Here

$$|S_1| = 6k+3, \quad |S_2| = 6k+6.$$

No additional appearances are possible, i.e., no mixed triples can occur, and we arrive at (8).

Case B. $n \equiv 3$ (4).

Here $f \leqq \dfrac{n-3}{4}$, $s = \dfrac{n-3}{2} - t$, $v = \dfrac{n+1}{2} + t$, and

$$|P| \geqq \frac{1}{3}\left(\left(\frac{n-1}{2} - t\right)\left(\frac{n-3}{4} - \left[\frac{t}{2}\right]\right) + \left(\frac{n+1}{2} + t\right)\left(\frac{n+1}{4} + \left[\frac{t}{2}\right]\right)\right) = \tag{10}$$

$$= \frac{n(n-1)}{12} + \frac{(4t+2)\left(2\left[\frac{t}{2}\right]+1\right)}{12}.$$

Since in this case $\dfrac{n(n-1)}{12}$ is not an integer, for (10) not to exceed the already established bound (7), $t=0$ or 1. We are then faced with three possiblities as to whether $n \equiv 3, 7, 11 \bmod 12$, and since the argument follows the pattern of Case *A*, the proof has been omitted.

Let us conclude with an example to illustrate (8), $n=15$. $\beta(T^2_{3,15}) = 18$ and according to (8), there are essentially three different decompositions of a minimal externally stable set *P*.

(a) $|S_1| = 7$, $|S_2| = 8$, no mixed triples.

$A = ((1, 2, 3), (1, 4, 5), (1, 6, 7), (2, 4, 6), (2, 5, 7,) (3, 4, 7), (3, 5, 6))$,

$B = ((8, 9, 10), (8, 11, 12), (8, 13, 14), (9, 11, 13), (9, 12, 14), (10, 11, 14),$
$(10, 12, 13), (8, 9, 15), (10, 11, 15), (12, 13, 15), (13, 14, 15))$.

(b) $|S_1| = 7$, $|S_2| = 8$, one mixed triple.

A as in (a), *B* as in above minus (13, 14, 15), mixed triple (1, 14, 15).

(c) $|S_1| = 6$, $|S_2| = 9$, no mixed triple.

$A = ((1, 2, 3), (1, 2, 4), (3, 4, 5), (3, 4, 6), (5, 6, 1), (5, 6, 2))$,

$B = ((7, 8, 9), (7, 10, 11), (7, 12, 13), (7, 14, 15), (8, 10, 12), (8, 11, 14),$
$(8, 13, 15), (9, 10, 15), (9, 11, 13), (9, 12, 14), (10, 13, 14), (11, 12, 15))$.

REFERENCES

[1] AIGNER, M.: The uniqueness of the cubic lattice graph, *J. Combinatorial Theory* **6** (1969), 282—297.
[2] BOSE, R. C.: Strongly regular graphs, partial geometries and partially balanced designs, *Pac. J. Math.* **13** (1963), 389—419.
[3] BOSE, R. C. and LASKAR, R.: A characterization of tetrahedral graphs, *J. Combinatorial Theory* **3** (1967), 366—385.
[4] BOSE, R. C.: *Combinatorial problems of experimental designs,* Vol. I, John Wiley and Sons, New York, to appear.
[5] DOWLING, T.: A characterization of the T_m-graph, *J. Combinatorial Theory* **6** (1969), 251—263.
[6] FORT, M. K., JR., and HEDLUND, G. A.: Minimal coverings of pairs by triples, *Pac. J. Math.* **8** (1958), 709—719.
[7] KALBFLEISCH, J. G. and STANTON, R. G.: Maximal and minimal coverings of $(k-1)$-tuples by k-tuples, *Pac. J. Math.* **26** (1968), 131—140.
[8] LASKAR, R.: A characterization of cubic lattice graphs, *J. Combinatorial Theory* **3** (1967), 386—401.
[9] di PAOLA, J. W.: Blockdesigns and graphtheory, *J. Combinatorial Theory,* **1** (1966), 132—148.
[10] SCHÖNHEIM, J.: On coverings, *Pac. J. Math.* **14** (1964), 1405—1411.
[11] SCHÖNHEIM, J.: On maximal systems of k-tuples, *Studia Sci. Math. Hung.* **1** (1966), 363—368.
[12] SHRIKHANDE, S. S.: The uniqueness of the L_2-association scheme, *Ann. Math. Stat.* **30** (1959), 781—798.
[13] TAUSSKY, O. and TODD, J.: Covering theorems for groups, *Ann. Soc. Pol. d. Math.* **21** (1948), 303—305.

The University of North Carolina at Chapel Hill

(Received January 20, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 317—325.

LIMES SUPERIOR SÄTZE FÜR DIE WARTEMODELLE

von

P. BÁRTFAI

In dieser Arbeit werden wir uns mit einem der wichtigsten Wartemodelle, mit dem Modell GI/G/1 beschräftigen. Wie im Satz vom iterierten Logarithmus bei den Irrfahrten taucht die Frage auch hier auf: mit welcher monotonen Funktion $f(n)$ soll man die Wartezeit w_n verteilen, dass der $\overline{\lim} \frac{w_n}{\log n}$ endlich sei? Dieser Analogon des Satzes vom iterierten Logarithmus, der hier lieber als Satz vom Limes superior genannt wird, gibt natürlich wichtige Aufklärungen über die Natur der Folge w_n. Wir werden uns mit den ähnlichen Fragen für die Länge der Warteschlange und für die Arbeitsperiode beschäftigen.

Ohne die Beschränkung der Allgemeinheit können wir annehmen (der Beweis kann ebenso durchgeführt werden), dass der erste Kunde im Zeitpunkt Null ankommt, und seine Wartezeit Null ist. Führen wir die folgenden Bezeichungen ein:

η_k — die Bedienungszeit der k-ten Kunden ($\eta_k \geqq 0$),

τ_k — die k-ten Pausezeit ($\tau_k \geqq 0$),

$T(x)$ — die Verteilungsfunktion von τ_k,

$\xi_k = \eta_k - \tau_k$,

$F(x)$ — die Verteilungsfunktion von ξ_k ($k = 1, 2, \ldots$),

w_k — die Wartezeit der k-ten Kunden, $w_k = [w_{k-1} + \xi_{k-1}]_+$, $w_1 = 0$,

$W_k(x)$ — die Verteilungsfunktion von w_k,

ϑ_k — die grösste Wartezeit in der k-ten Arbeitsperiode,

$G(x)$ — die Verteilungsfunktion von ϑ_k,

ζ_k — die Zahl der Wartenden in dem Zeitpunkt der Ankunft des k-ten Kunden,

ζ_k^* — die Zahl des Kunden, die während des Wartens des k-ten Kunden ankommen,

F_k, G_k, T_k — die k-te Faltungspotenz von F, G, T.

Die Arbeitsperiode ist in gewohnter Weise definiert: Es bezeichnet $\nu_1 > \nu_0 = 0$ die kleinste Zahl, für welche $w_{\nu_1} = 0$ ist, bezeichnet $\nu_2 > \nu_1$ die kelinste Zahl für welche $w_{\nu_2} = 0$ ist u. s. w., so definieren wir die Zufallsveränderlichen ϑ_k und λ_k mit den Relationen

$$\vartheta_k = \max_{\nu_{k-1} < j \leqq \nu_k} w_j,$$

$$\lambda_k = \nu_k - \nu_{k-1}.$$

Nehmen wir die folgende Bedingungen für die Zufallsveränderlichen $\xi_1, \xi_2, \ldots$ an:

1. sie sind unabhängig und sie haben dieselbe Verteilung mit der Verteilungsfunktion $F(x)$,
2. $\mathsf{M}(\xi_1) = m < 0$,
3. $\mathsf{P}(\xi_1 > 0) > 0$,
4. die momenterzeugende Funktion $R(t) = \mathsf{M}(e^{t\xi_1})$ existiert in einer kelinen Umgebung des Ursprungs.

In den Beweisen werden wir den Satz über die grossen Abweichungen anwenden, in Verbindung damit führen wir den Begriff der Chernoffschen Funktion von $F(x)$ ein:

$$\varrho(x) = \inf_t e^{-tx} R(t),$$

dann gilt der Satz von CHERNOFF

$$\sqrt[n]{1 - F_n(nx)} \to \varrho(x).$$

5. Für uns ist es noch notwendig die Existenz der Zahl $m_0 \neq m$ anzunehmen*, die mit der Gleichung

$$\frac{\varrho'(m_0)}{\varrho(m_0)} = \frac{\log \varrho(m_0)}{m_0} \tag{3}$$

eindeutig bestimmt ist (siehe die Abbildung 1.), nämlich gilt $\log \varrho(m) = 0$, $\log \varrho(x) < 0$ für $x \neq m$, ferner ist $\log \varrho(x)$ eine konkave Funktion.

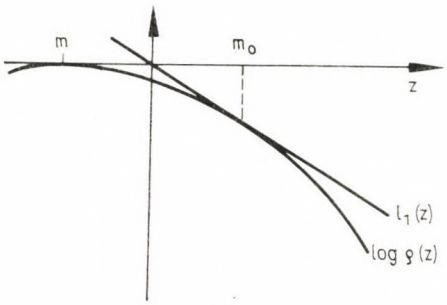


Abb. 1

SATZ 1. *Es gilt unter den Bedingungen* 1—5. *mit Wahrscheinlichkeit* 1

$$\overline{\lim_{n\to\infty}} \frac{w_n}{\log n} = A \tag{4}$$

und

$$A = -\frac{\varrho(m_0)}{\varrho'(m_0)}. \tag{5}$$

Der Wert von A lässt sich auch anders bestimmen.

SATZ 1'. *Ist* $t = t_0 \neq 0$ *eine Wurzel der Gleichung*

$$R(t) = 1, \tag{6}$$

so gilt

$$A = \frac{1}{t_0}. \tag{7}$$

Dieser Wert t_0 tritt auch bei dem asymptotischen Verhalten der Grenzvertilung von w_k auf (s. CRAMÉR [2]).

* Es ist möglich unter der Bedingung 5 freizukommen, siehe [1].

Der Satz 1 lässt sich verschärfen.

SATZ 2. *Es gilt mit Wahrscheinlichkeit* 1

$$\lim_{n\to\infty} \frac{\max\limits_{k\leq n} w_k}{\log n} = A. \tag{8}$$

Ähnliche Sätze sind gültig für λ_k und ζ_k.

SATZ 3. *Es gilt mit Wahrscheinlichkeit* 1

$$\overline{\lim_{n\to\infty}} \frac{\lambda_n}{\log n} = -\frac{1}{\log \varrho(0)}. \tag{9}$$

SATZ 4. *Ist die Zufallsveränderlichen* $\eta_1, \tau_1, \eta_2, \tau_2, \ldots$ *vollständig unabhängig und erfüllen die Bedingungen* 1—5., *so gilt mit Wahrscheinlichkeit* 1

$$\overline{\lim_{n\to\infty}} \frac{\zeta_n}{\log n} = -\frac{1}{\log R_\tau(-t_0)}. \tag{10}$$

wobei $R_\tau(t) = \mathsf{M}(e^{t\tau})$ *ist.*

Wir bemerken noch, dass die vorigen Sätze für mehrere modifizierten Irrfahrten unverändert gültig sind, z. B. für die Irrfahrten mit einer spiegelnden Wand, für die Irrfahrten, deren Schritte eine andere Verteilung mit positivem Erwartungswert haben, wenn die Summe negativ ist.

Zum Schluss wird der Wert A für einige Verteilungen gegeben. Ist $\xi_1 = a$ mit Wahrscheinlichkeit p und $\xi_1 = -a$ mit Wahrscheinlichkeit $q = 1-p$ $(q>p)$, so ist

$$A = \frac{a}{\log \frac{q}{p}}.$$

Hat ξ_1 eine normale Verteilung mit den Erwartungswert $m<0$ und mit der Streuung σ, so ist $A = -\frac{\sigma^2}{m}$. Ist $\xi_1 = \eta_1 - \tau_1$ der Unterschied der zwei Zufallsveränderlichen, die exponentielle Verteilung haben, so gilt $A = \frac{1}{\mu - \lambda}$, wo $\mathsf{M}(\eta_1) = \lambda < \mu = \mathsf{M}(\tau_1)$ ist.

Beweise

Der BEWEIS *von Satz* 3. Die Zufallsveränderlichen $\lambda_1, \lambda_2, \ldots$ sind unabhängig und sie haben dieselbe Verteilung. Ist

$$\pi_n = \mathsf{P}(\lambda_k = n)$$

und

$$\varphi(s) = \sum_1^\infty \pi_n s^n,$$

so gilt die von SPARRE—ANDERSEN stammende Identität (s. zB. FELLER [3] S. 395)

$$\log \frac{1-\varphi(s)}{1-s} = \sum_1^\infty \frac{s^n}{n}\left(1 - F_n(+0)\right).$$

Der Konvergenzradius der an der rechten Seite stehenden Potenzreihe ist

$$R = \left(\lim \sqrt[n]{\frac{1}{n}(1-F_n(+0))}\right)^{-1} = \frac{1}{\varrho(0)},$$

das ist zugleich auch der Konvergenzradius von $\varphi(s)$.

Bedeutet $\{x\}$ die kleinste natürliche Zahl, die nicht kleiner als x ist, dann ist

$$\sum_{n=1}^{\infty} \mathsf{P}(\lambda_n \geqq c \log n) = \sum_{n=1}^{\infty} \sum_{k=\{c \log n\}}^{\infty} \pi_k = \sum_{k=1}^{\infty} [e^{k/c}]\pi_k.$$

Daraus kann man sehen, dass die Reihe $\sum_n \mathsf{P}(\lambda_n \geqq c \log n)$ danach konvergent oder divergent ist, dass $e^{\frac{1}{c}} < \frac{1}{\varrho(0)}$ oder $e^{\frac{1}{c}} > \frac{1}{\varrho(0)}$, dh. $c > -\frac{1}{\log \varrho(0)}$ oder $c < -\frac{1}{\log \varrho(0)}$ ist. Mit Rücksicht auf den Borel-Cantellischen Lemma folgt daraus die Behauptung des Satzes.

Bemerkung. Aus der ersten Hälfte des Beweises folgt, dass die momenterzeugende Funktion $\mathsf{M}(e^{t\lambda_1})$ existiert, falls $t < -\log \varrho(0)$.

HILFSSATZ 1.

$$\lim_{x \to \infty} \frac{1}{x} \log(1 - G(x)) = -\frac{1}{A}. \tag{11}$$

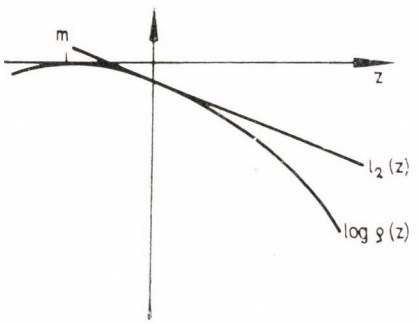


Abb. 2

BEWEIS. Wir werden zwei Schätzungen für $\varrho(z)$ anwenden, beide beruhen auf der Konkavität von $\log \varrho(z)$. Der Abbildung 1 nach

$$\log \varrho(z) \leqq l_1(z) = \frac{\varrho'(m_0)}{\varrho(m_0)} z_1$$

d.h.

$$\varrho(z) \leqq e^{\frac{\varrho'(m_0)}{\varrho(m_0)} z} = e^{-\frac{z}{A}} \tag{12}$$

und der Abbildung 2 nach

$$\log \varrho(z) \leqq l_2(z) = \log \varrho(0) - cz$$

d.h.

$$\varrho(z) \leqq \varrho(0)\, e^{-cz} \qquad (c > 0). \tag{13}$$

Ist $K = \left[\frac{x}{m_0} \cdot \frac{\log \varrho(m_0)}{\log \varrho(0)}\right]$, so folgt aus (12) und (13)

$$1 - G(x) = \mathsf{P}(\vartheta_1 \geqq x) \leqq \sum_{j=1}^{\infty} \mathsf{P}(\xi_1 + \cdots + \xi_j \geqq x) \leqq \sum_{j=1}^{\infty} \varrho^j\left(\frac{x}{j}\right) =$$

$$= \sum_{j=1}^{k} \varrho^j\left(\frac{x}{j}\right) + \sum_{j=k+1}^{\infty} \varrho^j\left(\frac{x}{j}\right) \leqq K e^{-\frac{x}{A}} + \varrho^{k+1}(0)\, e^{-cx} \frac{1}{1-\varrho(0)} \leqq P_1(x)\, e^{-\frac{x}{A}}, \tag{14}$$

wobei $P_1(x)$ ein Polynom mit dem ersten Grad bedeutet. Anderseits

$$\mathsf{P}\left(\max \vartheta_1, \vartheta_2, \ldots, \vartheta_{\left[\frac{x}{m_0}\right]} \leqq m_0 \left[\frac{x}{m_0}\right]\right) \leqq 1 - \mathsf{P}\left(\xi_1 + \cdots + \xi_{\left[\frac{x}{m_0}\right]} \geqq m_0 \left[\frac{x}{m_0}\right]\right)$$

und

$$\mathsf{P}\left(\max \vartheta_1, \vartheta_2, \ldots, \vartheta_{\left[\frac{x}{m_0}\right]} \leqq m_0 \left[\frac{x}{m_0}\right]\right) = \left(G\left(m_0 \left[\frac{x}{m_0}\right]\right)\right)^{\left[\frac{x}{m_0}\right]} \geqq$$

$$\geqq 1 - \left[\frac{x}{m_0}\right]\left(1 - G\left(m_0 \left[\frac{x}{m_0}\right]\right)\right) \geqq 1 - \frac{x}{m_0}(1 - G(x - m_0)),$$

aus diesen Ungleichungen folgt, dass

$$1 - G(x) \geqq \frac{m_0}{x + m_0}\left\{1 - F_{\left[\frac{x}{m_0}\right]+1}\left(m_0\left(\left[\frac{x}{m_0}\right]+1\right)\right)\right\}. \tag{15}$$

Aus (14) und (15) mit der Hilfe des Chernoffschen Satzes folgt (11).

HILFSSATZ 2. *Es gilt mit Wahrscheinlichkeit* 1

$$\frac{\log \nu_k}{\log k} \to 1 \qquad (k \to \infty). \tag{16}$$

BEWEIS. Nach dem Gesetz der grossen Zahlen

$$\frac{1}{k}\nu_k = \frac{1}{k}\sum_1^k \lambda_j \to \mathsf{M}(\lambda_j)$$

($\mathsf{M}(\lambda_j)$ ist endlich, da die momenterzeugende Funktion $\mathsf{M}(e^{t\lambda_j})$ existiert), und daraus folgt (16).

Der BEWEIS *von Satz* 1. Die Reihe $\sum_n \mathsf{P}(\vartheta_n \geqq a \log n)$ ist konvergent oder divergent danach, dass $a > A$ oder $a < A$ ist, nähmlich es gilt nach dem Hilfssatz 1

$$\log \mathsf{P}(\vartheta_n \geqq a \log n) \sim -\frac{a}{A} \log n.$$

Aus dem Borel-Cantellischen Lemma ergibt sich, dass

$$\overline{\lim} \frac{\vartheta_n}{\log n} = A$$

mit Wahrscheinlichkeit 1 gilt.

Da $\dfrac{w_n}{\log n} \leqq \dfrac{\vartheta_k}{\log(k-1)}$ im Falle $\nu_{k-1} \leqq n < \nu_k$ ist,

folgt

$$\overline{\lim} \frac{w_n}{\log n} \leqq A. \tag{17}$$

Wählen wir die Teilfolge $w_{n_1}, w_{n_2}, \ldots$ so, dass $v_{k-1} \leqq n_k \leqq v_k$ und $w_{n_k} = \vartheta_k$ $(k = 1, 2, \ldots)$ erfüllen sollen, dann gilt wegen (16)

$$\overline{\lim} \frac{w_n}{\log n} \geqq \overline{\lim} \frac{w_{n_k}}{\log n_k} \geqq \overline{\lim} \frac{\vartheta_k}{\log k} : \frac{\log v_k}{\log k} = A. \tag{18}$$

Mit (16) und (17) ist der Satz 1 bewiesen.

Der BEWEIS *von Satz* 1′. Es ist genug, wegen der Relationen (5), (6) und (7), zu beweisen, dass

$$R\left(-\frac{\varrho'(m_0)}{\varrho(m_0)}\right) = 1 \tag{19}$$

ist. Nach (1)

$$\varrho(x) = \inf_t e^{-tx} R(t),$$

aber, wie man leicht berechnen kann, die Minimumstelle von $e^{-tx}R(t)$ bei einem gegebenen Wert x ist

$$t = -\frac{\varrho'(x)}{\varrho(x)}$$

falls $\varrho(x) > 0$ d.h. (da $\varrho(m_0) \neq 0$ ist) wir bekommen die Gleichung

$$\varrho(m_0) = e^{-\frac{\varrho'(m_0)}{\varrho(m_0)} m_0} R\left(-\frac{\varrho'(m_0)}{\varrho(m_0)}\right).$$

Daraus und aus (3) folgt (19).

Der BEWEIS *von Satz* 2. Bezeichnet k_n die Zahl, für welche $w_{k_n} = \max_{k \leqq n} w_k$, so folgt aus (17)

$$\overline{\lim_{n \to \infty}} \frac{\max_{k \leqq n} w_k}{\log n} = \overline{\lim_{n \to \infty}} \frac{w_{k_n}}{\log n} \leqq \overline{\lim_{n \to \infty}} \frac{w_{k_n}}{\log k_n} \leqq A. \tag{20}$$

Es sei $n_k = [k^\beta]$. Wir können eine Schätzung mit der Hilfe (11) durchführen. Für $k \geqq k_0$ ist es gültig

$$\begin{aligned} p_k = \mathsf{P}\left(\max_{n_k < j \leqq n_{k+1}} \vartheta_j < (A-\varepsilon)\beta \log k\right) &= \left(G((A-\varepsilon)\beta \log k)\right)^{n_{k+1}-n_k} \leqq \\ &\leqq \exp\left\{-(n_{k+1}-n_k)\left(1-G((A-\varepsilon)\beta \log k)\right)\right\} \leqq \\ &\leqq \exp\left\{-\frac{1}{2} k^{\beta-1} e^{-\left(\frac{1}{A}+\frac{\varepsilon}{A^2}\right)(A-\varepsilon)\beta \log k}\right\} = \\ &= \exp\left\{-\frac{1}{2} k^{-1+\beta\frac{\varepsilon^2}{A^2}}\right\}. \end{aligned}$$

Wählen wir die Zahl β so, dass $\beta \dfrac{\varepsilon^2}{A^2} > 1$ sei, so ist $\sum_1^\infty p_k < \infty$, also gilt die Relation

$$\underline{\lim} \frac{\max_{n_{k-1} < j \leqq n_k} \vartheta_j}{\log n_{k-1}} \geqq A - \varepsilon$$

mit Wahrscheinlichkeit 1. Daraus erhalten wir, dass

$$\underline{\lim} \frac{\max\limits_{j \leq n} w_j}{\log n} \geqq \underline{\lim} \frac{\max\limits_{j \leq \nu_{n_k}} w_j}{\log \nu_{n_k+1}} \geqq \underline{\lim} \frac{\max\limits_{n_{k-1} < j \leq n_k} \vartheta_j}{\log n_{k-1}} \geqq A - \varepsilon$$

ist, diese Relation und (20) geben die Behauptung von Satz 2.

HILFSSATZ 3.

$$1 - W_n(x) \leqq e^{-t_0 x} \qquad (n = 2, 3, \ldots). \tag{21}$$

BEWEIS mit vollständiger Induktion. Im Falle $n=2$ aus (6) folgt

$$1 = \int_{-\infty}^{\infty} e^{t_0 y} dF(y) \geqq \int_{x}^{\infty} e^{t_0 y} dF(y) \geqq e^{t_0 x}\big(1 - F(x)\big)$$

d.h. für $x > 0$

$$1 - W_2(x) = 1 - F(x) \leqq e^{-t_0 x},$$

im Falle $x = 0$ ist die Ungleichung trivial.

Im allgemeinen ist

$$1 - W_{n+1}(x) = \int_{-\infty}^{x} \big(1 - W_n(x-y)\big) dF(y) + 1 - F(x) \leqq$$

$$\leqq e^{-t_0 x} \int_{-\infty}^{x} e^{t_0 y} dF(y) + 1 - F(x) \leqq$$

$$\leqq e^{-t_0 x} - e^{-t_0 x} \int_{x}^{\infty} e^{t_0 y} dF(x) + 1 - F(x) \leqq e^{-t_0 x}.$$

Der BEWEIS *von Satz* 4.

Zuerst werden wir den Satz vom Limes superior für ζ_n^* beweisen. Mit der Anwendung (21) ergibt sich

$$\mathsf{P}(\zeta_n^* \geqq l) = \int_{0}^{\infty} \mathsf{P}(\zeta_n^* \geqq l | w_n = x) dW_n(x) =$$

$$= \int_{0}^{\infty} \mathsf{P}(\tau_1 + \ldots + \tau_l < x) dW_n(x) = \int_{0}^{\infty} T_l(x) dW_n(x) =$$

$$= \int_{0}^{\infty} \big(1 - W_n(x)\big) dT_l(x) \leqq \int_{0}^{\infty} e^{-t_0 x} dT_l(x) = R_\tau^l(-t_0),$$

daraus folgt, dass $\sum\limits_n \mathsf{P}(\zeta_n^* \geqq c \log n)$ konvergent ist, falls $c > -\dfrac{1}{\log R_\tau(-t_0)}$. Aus dem Borel-Cantellischen Lemma folgt, dass

$$\overline{\lim} \frac{\zeta_n^*}{\log n} \leqq -\frac{1}{\log R_\tau(-t_0)} \tag{22}$$

mit Wahrscheinlichkeit 1 gilt.

Anderseits, wählen wir die Folge $n_1, n_2, \ldots$ sowie im (18) —, dass $v_{k-1} \leqq n_k \leqq v_k$ und $w_{n_k} = \vartheta_k$ $(k = 1, 2, \ldots)$ erfüllen sollen, so ist

$$\mathsf{P}(\zeta_{n_k}^* \geqq l) = \int_0^\infty \mathsf{P}(\zeta_{n_k}^* \geqq l | \vartheta_k = x)\, dG(x) =$$

$$= \int_0^\infty T_l(x)\, dG(x) = \int_0^\infty \big(1 - G(x)\big)\, dT_l(x).$$

Nun wenden wir das Hilfssatz 1 an, wenn $x > x_0$, so gilt

$$1 - G(x) \geqq e^{-x(1+\varepsilon)t_0},$$

und daraus folgt

$$\mathsf{P}(\zeta_{n_k}^* \geqq l) \geqq \int_{x_0}^\infty e^{-x(1+\varepsilon)t_0}\, dT_l(x) =$$

$$= R_\tau^l\big(-t_0(1+\varepsilon)\big) - \int_0^{x_0} e^{-t_0(1+\varepsilon)x}\, dT_l(x) \geqq R_\tau^l\big(-t_0(1+\varepsilon)\big) - T_l(x_0).$$

Für beliebiges $\delta > 0$ gilt

$$\mathsf{P}(\zeta_{n_k}^* \geqq l) \geqq R_\tau^l\big(-t_0(1+\varepsilon)\big) - T_l(l\delta) \geqq R_\tau^l\big(-t_0(1+\varepsilon)\big) - \varrho_\tau^l(\delta),$$

wenn l genügend gross ist. $\varrho_\tau(x)$ bezeichnet hier die Chernoffschen Funktion von $T(x)$. Da

$$\varrho_\tau(+0) = \mathsf{P}(\tau_1 = 0) < R_\tau(x)$$

ist, lässt $\delta > 0$ sich so erwählen, dass

$$\varrho_\tau(\delta) < R_\tau\big(-t_0(1+\varepsilon)\big)$$

sei, aber dann gilt für genügend grosses l

$$\mathsf{P}(\zeta_{n_k}^* \geqq l) \geqq \frac{1}{2} R_\tau^l\big(-t_0(1+\varepsilon)\big).$$

Daraus kann man sehen, dass

$$\sum_n \mathsf{P}(\zeta_{n_k}^* \geqq c \log k) = +\infty$$

ist, falls

$$c < -\frac{1}{\log R_\tau\big(-t_0(1+\varepsilon)\big)},$$

also gilt mit Wahrscheinlichkeit 1

$$\overline{\lim}\, \frac{\zeta_{n_k}^*}{\log k} \geqq -\frac{1}{\log R_\tau(-t_0)}.$$

Aus (16) ergibt sich, dass

$$\overline{\lim}\, \frac{\zeta_n^*}{\log n} \geqq \overline{\lim}\, \frac{\zeta_{n_k}^*}{\log n_k} = \overline{\lim}\, \frac{\zeta_{n_k}^*}{\log k} \geqq -\frac{1}{\log R_\tau(-t_0)}. \tag{23}$$

Wir haben bewiesen, dass — wie aus (22) und (23) folgt —

$$\overline{\lim}\frac{\zeta_n^*}{\log n} = -\frac{1}{\log R_\tau(-t_0)} \tag{24}$$

mit Wahrscheinlichkeit 1 gilt.

Im zweiten Teil des Beweises werden wir zeigen, dass

$$\overline{\lim}\frac{\zeta_n}{\log n} = \overline{\lim}\frac{\zeta_n^*}{\log n}. \tag{25}$$

Ist $\zeta_n = k$, so folgt $\zeta_{n-k}^* \geqq k$, anderseits, ist $\zeta_n^* = l$ so folgt $\zeta_{n+l} \geqq l$. Daraus und aus der Relation $n - \zeta_n \to \infty$ folgt

$$\overline{\lim}\frac{\zeta_n}{\log n} \leqq \overline{\lim}\frac{\zeta_{n-\xi_n}^*}{\log(n-\zeta_n)} \leqq \overline{\lim}\frac{\zeta_n^*}{\log n}$$

und

$$\overline{\lim}\frac{\zeta_n}{\log n} \geqq \overline{\lim}\frac{\zeta_{n+\xi_n^*}}{\log(n+\zeta_n^*)} \geqq \overline{\lim}\frac{\zeta_n^*}{\log n}\cdot\frac{\log n}{\log(n+\zeta_n^*)}.$$

Wegen (24)

$$\frac{\log n}{\log(n+\zeta_n^*)} \to 1$$

mit Wahrscheinlichkeit 1, so ist (25) und damit der Satz 4. bewiesen.

LITERATUR

[1] BÁRTFAI, P.: Large deviations in the queueing theory, *Periodica Mathematica*, **1** (1971) (in print)
[2] CRAMÉR, H.: On some questions connected with mathematical risk, *Univ. Calif. Publ. in Statistics* **2** (1954), 99—125.
[3] FELLER, W.: *An Introduction to Probability Theory and its Applications*. Vol. II. Wiley, New York 1966.

Mathematisches Institut der Ungarischen Akademie der Wissenschaften, Budapest

(Eingegangen: 2. April, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 327—335.

ÜBER DIE CHARAKTERISIERUNG GEWISSER FUNKTIONENKLASSEN DURCH APPROXIMATION MIT RIESZSCHEN MITTELN VON FOURIERREIHEN*

von

D. KRÁLIK

1. Einleitung

Der Gedanke, einen Teil der Approximationstheorie auf einfache reihentheoretische Sätze zu gründen, war schon bisher fruchtbar. G. ALEXITS scheint diese Idee als erster in die approximationstheoretischen Untersuchungen eingeführt zu haben, als er die Funktionenklasse $\mathrm{Lip}(1, p)$ $(1 \leqq p \leqq +\infty)$[1] charakterisierte [1]. ALEXITS, sowie andere Forscher haben mit derselben Methode weitere nennenswerte Ergebnisse erzielt. Neuerdings gelang es uns auf diese Weise die Klasse derjenigen 2π-periodischen Funktionen approximationstheoretisch zu charakterisieren, deren Derivierten beliebiger Ordnung zur Klasse $\mathrm{Lip}(1, p)$ gehören [5]. Dieses Ergebnis wurde zwar auf andere Weise schon früher erzielt ([3], [6]), jedoch nicht in lokalisierter Form, weshalb die Ausdehnung auf nichtperiodische Funktionen nicht möglich war.

Wir werden im folgenden die Charakterisierung für eine recht allgemeine Differenzierbarkeitsklasse in lokalisierter Form angeben, wobei wir als Approximationspolynome die Rieszschen Mittel der Fourierreihe verwenden. Unsere Ergebnisse lassen sich sehr einfach auch auf den Fall nichtperiodischer Funktionen übertragen; statt der trigonometrischen Approximationspolynome treten dann die Tschebyscheffschen Polynome auf. Allen unseren Betrachtungen liegen nur zwei einfache reihentheoretische Sätze zugrunde.

Wir führen zunächst die folgende Funktion ein: Sei $\lambda(x)$ eine für positive x-Werte definierte positive, monoton abnehmende Funktion mit $\lim\limits_{x \to +\infty} \lambda(x) = 0$, für welche geeignete Zahlen $0 < \beta \leqq \gamma < 1$ derart existieren, daß bei genügend großen x-Werten die Monotonie-Eigenschaften

$$\lambda(x)\, x^{\gamma} \uparrow \quad \text{bzw.} \quad \lambda(x)\, x^{\beta} \downarrow$$

bestehen. Wir sagen, die 2π-periodische Funktion $f(x)$ gehöre zur Funktionenklasse $L^p_\lambda = L^p_\lambda[-\pi, \pi]$, wenn $f(x) \in L^p[-\pi, \pi]$ $(1 \leqq p \leqq +\infty)$ und

$$\|f(x+h) - f(x)\|_{L^p[-\pi, \pi]} = O\left[\lambda\left(\frac{1}{|h|}\right)\right]$$

---

* Fortsetzung meiner im Acta Math. Acad. Sci. Hung. **20** (1969), 361—373. unter einem ähnlichen Titel erschienenen Arbeit.

[1] Die 2π-periodische Funktion $f(x) \in L^p[-\pi, \pi]$ gehört zur Klasse $\mathrm{Lip}(1, p)$, wenn

$$\left\{ \int_{-\pi}^{\pi} |f(x+h) - f(x)|^p \, dx \right\}^{1/p} = \|f(x+h) - f(x)\|_{L^p[-\pi, \pi]} = O(|h|)$$

gilt, wo $L^\infty[-\pi, \pi] = C[-\pi, \pi]$ die Klasse der stetigen Funktionen bedeutet.

ist (vgl. [4]). Wie wir früher gezeigt haben ([4]), gehört die Funktion $f(x)$ samt ihrer konjugierten $\tilde{f}(x)$ zur Klasse $L_\lambda^p = L_\lambda^p[-\pi, \pi]$ genau dann, wenn die beiden Approximationsbeziehungen

$$\|\sigma_n(f; x) - f(x)\|_{L^p[-\pi, \pi]} = O[\lambda(n)] \quad \text{bzw.} \quad \|\tilde{\sigma}_n(f; x) - \tilde{f}(x)\|_{L^p[-\pi, \pi]} = O[\lambda(n)]$$

erfüllt sind, wo

$$\sigma_n(f; x) = \sum_{k=0}^{n} \left(1 - \frac{k}{n+1}\right) A_k(x) \quad (A_k(x) = a_k \cos kx + b_k \sin kx)$$

das n-te Fejérsche Mittel der Fourierreihe von $f(x)$ und $\tilde{\sigma}_n(f; x)$ das entsprechende konjugierte Mittel bedeutet. Wollen wir dieses Ergebnis auf Differenzierbarkeitsklassen verallgemeinern, so haben wir statt der Fejérschen die entsprechenden Rieszschen Mittel r-ter Ordnung

$$R_n^{(r)}(f; x) = \sum_{k=0}^{n} \left(1 - \frac{k^r}{(n+1)^r}\right) A_k(x)$$

bzw. die konjugierten Rieszschen Mittel $\tilde{R}_n^{(r)}(f; x)$ zu benutzen. Unser Resultat lautet dann wie folgt:

SATZ 1. *Damit für die $(r-1)$-ten Derivierten $f^{(r-1)}(x) \in L_\lambda^p$ und $\tilde{f}^{(r-1)}(x) \in L_\lambda^p$, ist das Bestehen der beiden Beziehungen*

$$\begin{aligned} \|R_n^{(r)}(f; x) - f(x)\|_{L^p[-\pi, \pi]} &= O\left[\frac{\lambda(n)}{n^{r-1}}\right] \\ \|\tilde{R}_n^{(r)}(f; x) - \tilde{f}(x)\|_{L^p[-\pi, \pi]} &= O\left[\frac{\lambda(n)}{n^{r-1}}\right] \end{aligned} \tag{1}$$

notwendig, und das Bestehen einer von diesen hinreichend.

Um unser Ergebnis auch in lokalisierter Form aussprechen zu können, schicken wir die folgende Definition voraus: Wir sagen, die 2π-periodische Funktion $\varphi(x) \in L^p[-\pi, \pi]$ $(1 \leqq p \leqq +\infty)$ gehöre zur Funktionenklasse $L_\lambda^p(a, b)$ mit $[a, b] \subset [-\pi, \pi]$, wenn für jedes, völlig im Inneren von (a, b) liegendes Subintervall $[a_1, b_1]$ $(-\pi \leqq a < a_1 < b_1 < b \leqq \pi)$ die Relation

$$\left\{\int_{a_1}^{b_1} |\varphi(x+h) - \varphi(x)|^p \, dx\right\}^{1/p} = \|\varphi(x+h) - \varphi(x)\|_{L^p[a_1, b_1]} = O\left[\lambda\left(\frac{1}{|h|}\right)\right]$$

besteht, wo die Konstante im O-Symbol von a_1 und b_1 abhängt. Mit Hilfe dieser Definition können wir unser im Satz 1 ausgesprochenes Ergebnis in folgender lokalisierter Form verschärft aussprechen:

SATZ 2. *Damit die $(r-1)$-te Derivierte $f^{(r-1)}(x) \in L^p[-\pi, \pi]$ samt ihrer konjugierten $\tilde{f}^{(r-1)}(x) \in L^p[-\pi, \pi]$ zur Klasse $L_\lambda^p(a, b)$ gehöre, ist das Erfülltsein der beiden Beziehungen*

$$\begin{aligned} \|R_n^{(r)}(f; x) - f(x)\|_{L^p[a_1, b_1]} &= O\left[\frac{\lambda(n)}{n^{r-1}}\right] \\ \|\tilde{R}_n^{(r)}(f; x) - \tilde{f}(x)\|_{L^p[a_1, b_1]} &= O\left[\frac{\lambda(n)}{n^{r-1}}\right] \end{aligned} \tag{2}$$

für jedes, völlig im Inneren von (a, b) *liegendes Subintervall* $[a_1, b_1]$ *notwendig, und das Bestehen einer von diesen hinreichend, wobei die Konstante im O-Symbol von* a_1 *und* b_1 *abhängt.*

Durch die Transformation $x = \cos\theta$ können wir zwischen den Funktionen der Klasse $L^p[0, \pi]$, sowie denen der Klasse $L^p_{\frac{1}{\sqrt{1-x^2}}}[-1, 1]$ eine umkehrbar eindeutige Zuordnung herstellen und unser im Satz 2 ausgesprochenes Ergebnis auch auf den Fall nichtperiodischer Funktionen übertragen. Dabei treten als approximierende Polynome die Rieszschen Mittel der Entwicklung der Funktion nach Tschebyscheffschen Polynomen erster bzw. zweiter Art auf. Gehört nämlich eine 2π-periodische gerade Funktion $\varphi(\theta) \in L^p[0, \pi]$ inbezug auf ein Subintervall (β, α) $(0 \leqq \beta < \alpha \leqq \pi)$ zur Funktionenklasse $L^p_\lambda(\beta, \alpha)$, so gehört die Funktion $\varphi(\arccos x) = \varphi(x) \in L^p_{\frac{1}{\sqrt{1-x^2}}}[-1, 1]$ zur Klasse $L^p_\lambda(a, b)$ mit $a = \cos\alpha$, $b = \cos\beta$. Ist weiterhin $\varphi(\theta) \sim \frac{a_0}{2} + \sum_{\nu=1}^{\infty} a_\nu \cos\nu\theta$ die Fourierreihe von $\varphi(\theta)$, so ist

$$a_\nu = \frac{2}{\pi}\int_0^\pi \varphi(\theta)\cos\nu\theta\, d\theta = \frac{2}{\pi}\int_{-1}^{1} \frac{\varphi(x)T_\nu(x)}{\sqrt{1-x^2}}\, dx,$$

wo $\cos\nu\theta = \cos\nu(\arccos x) = T_\nu(x)$ das ν-te Tschebyscheffsche Polynom erster Art bedeutet. Besteht nun für $\varphi(\theta)$ die Beziehung

$$\|\varphi(\theta) - R_n^{(r)}(\varphi; \theta)\|_{L^p[\beta_1, \alpha_1]} = O[\lambda(n)]$$

mit $0 \leqq \beta < \beta_1 < \alpha_1 < \alpha \leqq \pi$, so besteht auch die entsprechende Beziehung

$$\left\{\int_{a_1}^{b_1} |\varphi(x) - R_n^{(r)}(\varphi; T; x)|^p \frac{dx}{\sqrt{1-x^2}}\right\}^{1/p} = O[\lambda(n)]$$

mit $a_1 = \cos\alpha_1$, $b_1 = \cos\beta_1$ und

$$R_n^{(r)}(\varphi; T; x) = \sum_{\nu=0}^{n}\left(1 - \frac{\nu^r}{(n+1)^r}\right) a_\nu T_\nu(x).$$

Da das Subintervall $[a_1, b_1]$ völlig im Inneren von $[-1, 1]$ liegt, gilt auch die Beziehung

$$\left\{\int_{a_1}^{b_1} |\varphi(x) - R_n^{(r)}(\varphi; T; x)|^p dx\right\}^{1/p} = \|\varphi(x) - R_n^{(r)}(\varphi; T; x)\|_{L^p[a_1, b_1]} = O[\lambda(n)].$$

Nach dieser Vorbereitung können wir nun unser auf den nichtperiodischen Fall bezogenes Resultat folgenderweise zusammenfassen:

SATZ 3. *Damit die* $(r-1)$*-te Derivierte* $f^{(r-1)}(x)$ *der im Intervall* $[-1, 1]$ *definierten Funktion* $f(x)$ *samt der „konjugierten" Derivierten*

$$\tilde{f}^{(r-1)}(x) = \frac{d^{r-1}\left(\tilde{f}(\arccos x)\right)}{dx^{r-1}} = \frac{d^{r-1}\tilde{f}(x)}{dx^{r-1}}$$

zur Funktionenklasse $L^p_\lambda(a, b)$ *gehöre,*[2] *ist das Bestehen der beiden Beziehungen*

$$\|f(x) - R_n^{(r)}(f; T; x)\|_{L^p[a_1, b_1]} = O\left[\frac{\lambda(n)}{n^{r-1}}\right]$$

(3)

$$\left\| \frac{\tilde{f}(\arccos x)}{\sqrt{1-x^2}} - \sum_{\nu=1}^{n}\left(1 - \frac{\nu^r}{(n+1)^r}\right) a_\nu T^*_{\nu-1}(x) \right\|_{L^p[a_1, b_1]} = O\left[\frac{\lambda(n)}{n^{r-1}}\right]$$

für jedes, völlig im Inneren von (a, b) *liegendes Subintervall* $[a_1, b_1]$ *notwendig, und das Bestehen einer von diesen hinreichend, wo die Konstante im O-Symbol von* a_1 *und* b_1 *abhängt.*

2. Zwei reihentheoretische Hilfssätze

Es seien die Glieder der Reihe $\sum_{\nu=0}^{\infty} a_\nu$ Elemente eines Banachschen Raumes B mit der Norm $\|\cdot\|$, ferner betrachten wir die in der Einleitung eingeführte Funktion $\lambda(x)$.

HILFSSATZ 1. *Wenn für die Rieszschen Mittel*

$$R_n^{(r)} = \sum_{\nu=0}^{n}\left(1 - \frac{\nu^r}{(n+1)^r}\right) a_\nu \qquad (r \geqq 1)$$

der Reihe Σa_ν *die Beziehung*

$$\|R_n^{(r)}\| = O[\lambda(n)] \tag{4}$$

besteht, so gilt für die Rieszschen Mittel

$$\bar{R}_n^{(r)} = \sum_{\nu=1}^{n}\left(1 - \frac{\nu^r}{(n+1)^r}\right) \frac{a_\nu}{\nu^{r-1}}$$

der Reihe $\sum \frac{a_\nu}{\nu^{r-1}}$ *die Beziehung*

$$\|\bar{R}_m^{(r)} - \bar{R}_n^{(r)}\| = O\left[\frac{\lambda(n)}{n^{r-1}}\right] \qquad (m > n).$$

---

[2] Wegen $0 < \beta_1 < \alpha_1 < \pi$ gehören die Derivierten $f^{(r-1)}(x) = \frac{d^{r-1}f(x)}{dx^{r-1}}$ bzw. $f^{(r-1)}(\theta) = \frac{d^{(r-1)}f(\theta)}{d\theta^{r-1}}$, wie leicht ersichtlich, gleichzeitig zur Klasse $L^p_\lambda(a, b)$ bzw. $L^p_\lambda(\beta, \alpha)$. Das gleiche gilt für die entsprecheden „konjugierten" Funktionen.

BEWEIS. Nach Satz 2 von [5] gilt

$$\|\bar{R}_m^{(r)} - \bar{R}_n^{(r)}\| = O(n^{-r}) \sum_{k=1}^{n-1} k^{r+1} \Delta^2 \left(\frac{1}{k^{r-1}}\right) \|R_k^{(r)}\| +$$

$$+ O(1) \sum_{k=n}^{m-1} (k+1) \Delta^2 \left(\frac{1}{k^{r-1}}\right) \|R_k^{(r)}\| + O(n^{-r}) \sum_{k=1}^{n-1} (k+1)^r \Delta \left(\frac{1}{k^{r-1}}\right) \|R_k^{(r)}\| +$$

$$+ O(m^{-r}) \sum_{k=n}^{m-1} (k+1)^r \Delta \left(\frac{1}{k^{r-1}}\right) \|R_k^{(r)}\| + O(n^{-r}) \sum_{k=1}^{n-1} (k+1)^r \Delta \left(\frac{1}{(k+1)^{r-1}}\right) \|R_k^{(r)}\| +$$

$$+ O(1) \sum_{k=n}^{m-1} \Delta \left(\frac{1}{k^{r-1}}\right) \|R_k^{(r)}\| + O(m^{-r}) \sum_{k=n}^{m-1} (k+1)^r \Delta \left(\frac{1}{(k+1)^{r-1}}\right) \|R_k^{(r)}\| +$$

$$+ O(n^{-(r-1)}) \|R_n^{(r)}\| + O(m^{-(r-1)}) \|R_m^{(r)}\|.$$

Wegen $\Delta \left(\frac{1}{k^{r-1}}\right) = O\left(\frac{1}{k^r}\right)$ und $\Delta^2 \left(\frac{1}{k^{r-1}}\right) = O\left(\frac{1}{k^{r+1}}\right)$

haben wir auf Grund von (4)

$$\|\bar{R}_m^{(r)} - \bar{R}_n^{(r)}\| = O(n^{-r}) \sum_{k=1}^{n-1} \lambda(k) + O(1) \sum_{k=n}^{m-1} k^{-r} \lambda(k) + O(n^{-r}) \sum_{k=1}^{n-1} \lambda(k) +$$

$$+ O(m^{-r}) \sum_{k=n}^{m-1} \lambda(k) + O(n^{-r}) \sum_{k=1}^{n-1} \lambda(k) + O(1) \sum_{k=n}^{m-1} k^{-r} \lambda(k) +$$

$$+ O(m^{-r}) \sum_{k=n}^{m-1} \lambda(k) + O\left(\frac{\lambda(n)}{n^{r-1}}\right) + O\left(\frac{\lambda(m)}{m^{r-1}}\right) =$$

$$= O(n^{-r}) \sum_{k=1}^{n-1} \lambda(k) + O(1) \sum_{k=n}^{m-1} k^{-r} \lambda(k) + O(m^{-r}) \sum_{k=n}^{m-1} \lambda(k) + O\left(\frac{\lambda(n)}{n^{r-1}}\right).$$

Auf Grund der Monotonie-Eigenschaften von $\lambda(x)$ erhalten wir für die einzelnen Glieder der letzten Reihe die folgenden Abschätzungen:

$$O(n^{-r}) \sum_{k=1}^{n-1} \lambda(k) = O(n^{-r}) \sum_{k=1}^{n-1} k^{\gamma} \lambda(k) k^{-\gamma} =$$

$$= O\left(n^{-r+\gamma} \lambda(n)\right) \sum_{k=1}^{n-1} k^{-\gamma} = O\left(\frac{\lambda(n)}{n^{r-1}}\right).$$

$$O(1) \sum_{k=n}^{m-1} k^{-r} \lambda(k) = O(1) \sum_{k=n}^{m-1} k^{-r-\beta} \lambda(k) k^{\beta} = O[\lambda(n) n^{\beta}] \sum_{k=n}^{m-1} k^{-r-\beta} = O\left(\frac{\lambda(n)}{n^{r-1}}\right),$$

und endlich

$$O(m^{-r}) \sum_{k=n}^{m-1} \lambda(k) = O(\lambda(n)) \frac{m-n}{m^r} = O\left(\frac{\lambda(n)}{m^{r-1}}\right) = O\left(\frac{\lambda(n)}{n^{r-1}}\right),$$

womit wir unseren Hilfssatz 1 bewiesen haben.

HILFSSATZ 2. *Bedeuten s_ν bzw. σ_ν die ν-te Teilsumme bzw. das ν-te* (C, 1)*-Mittel der Reihe Σa_ν, so gilt für ein beliebiges Element $s \in B$ die Ungleichung:*

$$\|R_n^{(r)} - s\| \leqq \frac{K_r}{n^r} \sum_{\nu=0}^{n-1} (\nu+1)^{r-1} \|\sigma_\nu - s\| + r\|\sigma_n - s\|,$$

wo $R_n^{(r)}$ das n-te Rieszsche Mittel der Reihe Σa_ν und K_r eine nur von r abhängende Konstante ist.

Es ist nämlich

$$R_n^{(r)} - s = \sum_{\nu=0}^{n} \frac{(\nu+1)^r - \nu^r}{(n+1)^r} (s_\nu - s) = \frac{r}{(n+1)^r} \sum_{\nu=0}^{n} (\nu + \vartheta_\nu)^{r-1} (s_\nu - s)$$

mit $0 < \vartheta_\nu < 1$ und die letzte Summe schreiben wir durch eine Abel-Transformation in die Gestalt:

$$\frac{r}{(n+1)^r} \sum_{\nu=0}^{n-1} [(\nu+\vartheta_\nu)^{r-1} - (\nu+1+\vartheta_{\nu+1})^{r-1}] (\nu+1)(\sigma_\nu - s) +$$

$$+ \frac{r}{(n+1)^r} (n+1)(n+\vartheta_n)^{r-1} (\sigma_n - s).$$

Nach dem Mittelwertsatz der Differenzialrechnung ergibt sich für die Differenz in der eckigen Klammer

$$(r-1)(\vartheta_\nu - 1 - \vartheta_{\nu+1})(\nu + \tau_\nu)^{r-2} \qquad (0 < \tau_\nu < 2)$$

und $|\vartheta_\nu - 1 - \vartheta_{\nu+1}| < 2$. Wenn wir auf die Normen übergehen, erhalten wir unsere Ungleichung.

3. Beweis der Sätze 1, 2 und 3

Beweis des Satzes 1. Wenn $f^{(r-1)}(x) \in L_\lambda^p$, so haben wir nach Satz 2 von [4] zugleich $\tilde{f}^{(r-1)}(x) \in L_\lambda^p$ und nach dem Satz 3 von [4] gelten die beiden Beziehungen:

$$\|\sigma_n(f^{(r-1)}; x) - f^{(r-1)}(x)\|_{L^p[-\pi,\pi]} = O[\lambda(n)]$$

und

$$\|\tilde{\sigma}_n(f^{(r-1)}; x) - \tilde{f}^{(r-1)}(x)\|_{L^p[-\pi,\pi]} = O[\lambda(n)].$$

Aus unserem Hilfssatz 2 folgen nun für die entsprechenden Rieszschen Mittel die analogen Beziehungen:

$$\|R_n^{(r)}(f^{(r-1)}; x) - f^{(r-1)}(x)\|_{L^p[-\pi,\pi]} = O[\lambda(n)]$$

und

$$\|\tilde{R}_n^{(r)}(f^{(r-1)}; x) - \tilde{f}^{(r-1)}(x)\|_{L^p[-\pi,\pi]} = O[\lambda(n)].$$

Ist z. B. $r-1$ gerade, so haben wir

$$f^{(r-1)}(x) \sim \pm \sum_{\nu=1}^{\infty} \nu^{r-1} A_\nu(x) \quad \text{und} \quad \tilde{f}^{(r-1)}(x) \sim \pm \sum_{\nu=1}^{\infty} \nu^{r-1} B_\nu(x)$$

$(B_\nu(x) = a_\nu \sin \nu x - b_\nu \cos \nu x)$ und die $L^p[-\pi, \pi]$-Normen der n-ten Rieszschen Mittel der Reihen

$$-f^{(r-1)}(x) \pm \sum_{\nu=1}^{\infty} \nu^{r-1} A_\nu(x) \quad \text{bzw.} \quad -\tilde{f}^{(r-1)}(x) \pm \sum_{\nu=1}^{\infty} \nu^{r-1} B_\nu(x)$$

sind von der Größenordnung $O[\lambda(n)]$. Nach unserem Hilfssatz 1 gelten also die Approximationsbeziehungen (1). Völlig analog verfahren wir im Fall $r-1$ ungerade.

Nehmen wir nun umgekehrt an, daß z. B. die erste Beziehung von (1) gilt. Dann haben wir für das trigonometrische Polynom 2^{n+1}-ter Ordnung

$$U_n(x) = R_{2^{n+1}}^{(r)}(f; x) - R_{2^n}^{(r)}(f; x)$$

die Abschätzung

$$\|U_n(x)\|_{L^p[-\pi, \pi]} = O\left[\frac{\lambda(2^n)}{2^{n(r-1)}}\right].$$

Wir verfahren weiter, wie es im klassischen Bernsteinschen Beweis üblich ist. Nach $(r-1)$-maler Anwendung der Bernsteinschen Ungleichung erhalten wir nämlich

$$\|U_n^{(r-1)}(x)\|_{L^p[-\pi, \pi]} = O[\lambda(2^n)],$$

woraus die Konvergenz der Reihe

$$R_1^{(r)}(f^{(r-1)}; x) + \sum_{\nu=0}^{\infty} U_\nu^{(r-1)}(x)$$

in der $L^p[-\pi, \pi]$-Metrik folgt, es konvergiert also die Folge $\{R_{2^n}^{(r)}(f^{(r-1)}; x)\}$ in der $L^p[-\pi, \pi]$-Metrik zu der Funktion $f^{(r-1)}(x)$. Wir haben nämlich:

$$\left\| \sum_{\nu=n+1}^{m} U_\nu^{(r-1)}(x) \right\|_{L^p[-\pi, \pi]} = O(1) \sum_{\nu=n}^{m} \lambda(2^\nu) = O(1) \sum_{\nu=n}^{m} \lambda(2^\nu) 2^{\nu\beta} 2^{-\nu\beta} =$$

$$= O[\lambda(2^n) 2^{n\beta}] \sum_{\nu=n}^{m} 2^{-\nu\beta} = O[\lambda(2^n) 2^{n\beta} 2^{-n\beta}] = O[\lambda(2^n)]$$

wegen der Monotonieeigenschaften von $\lambda(x)$. Wir haben also endlich

$$\|R_{2^n}^{(r)}(f^{(r-1)}; x) - f^{(r-1)}(x)\|_{L^p[-\pi, \pi]} = O[\lambda(2^n)],$$

woraus wir sehr einfach die Beziehung

$$\left\| f^{(r-1)}\left(x + \frac{1}{n}\right) - f^{(r-1)}(x) \right\|_{L^p[-\pi, \pi]} = O[\lambda(n)],$$

d.h. $f^{(r-1)}(x) \in L_\lambda^p$ erhalten. Nach Satz 2 von [4] haben wir zugleich auch $\tilde{f}^{(r-1)}(x) \in L_\lambda^p$. Damit haben wir unseren Satz 1 vollständig bewiesen.

Beweis des Satzes 2. a) Notwendigkeit. Sei z. B. $f^{(r-1)}(x) \in L^p_\lambda(a, b)$, so gilt nach dem lokalisierten Privaloffschen Satz[3] zugleich $\tilde{f}^{(r-1)}(x) \in L^p_\lambda(a, b)$. Wählen wir durch die Teilungspunkte $a < a^* < a_1 < b_1 < b^* < b$ ein Hilfsintervall (a^*, b^*) und betrachten wir die periodische Hilfsfunktion $g(x)$, die in (a^*, b^*) mit $f(x)$ identisch ist, und außerhalb (a^*, b^*) überall in $[-\pi, \pi]$ so glatt sein soll, daß $g^{(r-1)}(x) \in L^p_\lambda[-\pi, \pi]$ gelte. Es gilt also für die Punkte $x \in (a^*, b^*)$: $f^{(r-1)}(x) - g^{(r-1)}(x) = 0$, weiterhin

$$\|\sigma_n(f^{(r-1)}; x) - f^{(r-1)}(x)\|_{L^p[a_1, b_1]} \leqq \|\sigma_n(f^{(r-1)}; x) - \sigma_n(g^{(r-1)}; x)\|_{L^p[a_1, b_1]} +$$
$$+ \|\sigma_n(g^{(r-1)}; x) - g^{(r-1)}(x)\|_{L^p[a_1, b_1]} + \|g^{(r-1)}(x) - f^{(r-1)}(x)\|_{L^p[a_1, b_1]} =$$
$$= S_1 + S_2 + S_3.$$

Auf Grund der Eigenschaften von $g(x)$ haben wir

$$S_3 = 0 \quad \text{und} \quad S_2 = O[\lambda(n)].$$

Für S_1 ergibt sich nach einer leichten Zwischenrechnung die Relation

$$S_1 = \|\sigma_n(f^{(r-1)} - g^{(r-1)}; x)\|_{L^p[a_1, b_1]} = O\left(\frac{1}{n}\right),$$

wo die Konstante im O-Symbol von a_1 und b_1 abhängt. Wegen der Eigenschaften der Funktion $\lambda(x)$ haben wir also a fortiori

$$S_1 = O[\lambda(n)],$$

und nach Hilfssatz 2 endlich

$$\|R_n^{(r)}(f^{(r-1)}; x) - f^{(r-1)}(x)\|_{L^p[a_1, b_1]} = O[\lambda(n)].$$

Ist z. B. $r-1$ gerade, so entspricht der Funktion $f^{(r-1)}(x)$ die Fourierreihe $\pm \sum_{\nu=1}^{\infty} \nu^{r-1} A_\nu(x)$, und mit denselben Überlegungen wie beim Satz 1 erhalten wir nach Hilfssatz 1 die erste Beziehung von (2). Nach analogen Überlegungen betreffs der konjugierten Funktion $\tilde{f}^{(r-1)}(x)$ gewinnen wir auch die zweite Beziehung von (2). Wörtlich so verfahren wir, wenn $r-1$ ungerade ist.

---

[3] Gilt für die Funktion $f(x) \in L^p[-\pi, \pi]$ $(1 \leqq p \leqq +\infty)$ und für ein Teilintervall $[a, b] \subset [-\pi, \pi]$ die Beziehung

$$\left\{\int_a^b |f(x+h) - f(x)|^p \, dx\right\}^{1/p} = O\left[\lambda\left(\frac{1}{|h|}\right)\right],$$

so gilt für die konjugierte Funktion $\tilde{f}(x)$ die Beziehung

$$\left\{\int_{a_1}^{b_1} |\tilde{f}(x+h) - \tilde{f}(x)|^p \, dx\right\}^{1/p} = O\left[\lambda\left(\frac{1}{|h|}\right)\right]$$

für jedes inneres Subintervall $a < a_1 < b_1 < b$, wo die Konstante im O-Symbol mit a_1 und b_1 variiert. Der Beweis verläuft wörtlich so, wie im bekannten gewöhnlichen Fall $a_1 = -\pi$, $b_1 = \pi$. (Vgl. diesbezüglich [7], S. 156—157., sowie [4], Satz 2.)

b) Hinlänglichkeit. Dieser Teil beansprucht keinen ausführlichen Beweis; unser Gedankengang ist völlig identisch mit dem beim Satz 1, wir haben nur die lokalisierte Bernsteinsche Ungleichung ([2]) anzuwenden. Damit haben wir auch unseren Satz 2 bewiesen.

Beweis des Satzes 3. Notwendigkeit. Gehört $f^{(r-1)}(x)$ zur Funktionenklasse $L_\lambda^p(a, b)$, so haben wir zugleich

$$\frac{d^{r-1} f(\cos\theta)}{d\theta^{r-1}} = f^{(r-1)}(\theta) \in L_\lambda^p(\beta, \alpha)$$

und damit gleichzeitig auch $\tilde{f}^{(r-1)}(\theta) \in L_\lambda^p(\beta, \alpha)$ (vgl. Fußnote 3). Nach Satz 2 gelten also die beiden Beziehungen (2) statt a_1, b_1 und x mit β_1, α_1 und θ. Durch die Transformation $x = \cos\theta$ erhalten wir aus der ersten Beziehung (2) sofort die entsprechende erste Beziehung (3); dividieren wir die zweite Beziehung (2) durch $\sin\theta$ und übergehen wir auf die Veränderliche x, so erhalten wir auch die zweite Beziehung (3) mit dem Tschebyscheffschen Polynom zweiter Art

$$T_{\nu-1}^*(x) = \frac{\sin(\nu \arccos x)}{\sqrt{1-x^2}}.$$

Zum Beweis der Hinlänglichkeit nehmen wir an, daß z. B. die erste Beziehung (3) gilt. Mit $x = \cos\theta$ erhalten wir die analoge Beziehung (2) für die entsprechenden Funktionen der Veränderlichen θ. Nach Satz 2 gehören also $f^{(r-1)}(\theta)$ und $\tilde{f}^{(r-1)}(\theta)$ zur Klasse $L_\lambda^p(\beta, \alpha)$, d.h. die Funktionen $f^{(r-1)}(x)$ und $\tilde{f}^{(r-1)}(\arccos\theta) = \tilde{f}^{(r-1)}(x)$ zur Klasse $L_\lambda^p(a, b)$.

LITERATUR

[1] ALEXITS, G.: Sur l'ordre de grandeur de l'approximation d'une fonction périodique par les sommes de Fejér, *Acta Math. Acad. Sci. Hung.* **3** (1952), 29—40.

[2] Н. К. Бари: Обобщение неравенств С. Н. Бернштейна и А. А. Маркова, ДАН СССР, **90** (1953), 701—703.

[3] BUTZER, P. L. und PAWELKE, S.: Ableitungen von trigonometrischen Approximationsprozessen, *Acta Sci. Math.* **28** (1967), 173—183.

[4] KRÁLIK, D.: Über die approximationstheoretische Charakterisierung gewisser Funktionenklassen, *Acta Math. Acad. Sci. Hung.* **11** (1960), 377—386.

[5] KRÁLIK, D.: Über die Charakterisierung gewisser Funktionenklassen mit Hilfe der Rieszschen Mittel von Fourierreihen, *Acta Math. Acad. Sci. Hung.* **20** (1969), 361—373.

[6] ZAMANSKY, M.: Classes de saturation des procédés de sommation des séries de Fourier et applications aux séries trigonometriques, *Ann. Sci. Ecole Norm. Sup.* **67** (1950), 161—198.

[7] ZYGMUND, A.: *Trigonometrical Series*, Warsawa—Lwow, 1935.

Technische Hochschule, Budapest

(Eingegangen: 16. Juni, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 337—351.

ON THE NORMALFORM OF ANALYTIC DIFFERENTIAL EQUATIONS IN THE NEIGHBOURHOOD OF A CRITICAL POINT

(The case of the „saddle-point")

by

I. BIHARI and Á. ELBERT

1. Introduction. Let us consider the autonomous system

$$\dot{x}_k = \lambda_k x_k + F_k(\bar{x}), \quad \bar{x} = (x_1, \ldots, x_n), \quad (k = 1, \ldots, n); \quad \left(\dot{} = \frac{d}{dt} \right) \tag{1}$$

where $x_k = x_k(t)$ are, in general, complex valued funtions of the real variable t, λ_k are complex numbers, $F_k(\bar{x})$ are holomorphic functions involving terms of at least second order of $\bar{x}$ in a neighbourhood of the origin. Let $\bar{g} = (g_1, \ldots, g_n)$ be a vector (or lattice-point) consisting of non-negative integers and $|\bar{g}| = g_1 + \cdots + g_n$.

If with

$$\varepsilon_{k,\bar{g}} = \sum_{i=1}^{n} g_i \lambda_i - \lambda_k, \qquad (k = 1, \ldots, n)$$

the condition

$$|\varepsilon_{k,\bar{g}}| > 2n|\bar{g}|^{-2\nu}, \quad (k = 1, 2, \ldots, n); \quad |\bar{g}| \geqq 1 \tag{2}$$

is satisfied, where ν is a suitable positive number, then — as SIEGEL [1] has shown — the system (1) can be turned into the normalform

$$\dot{y}_k = \lambda_k y_k, \qquad (k = 1, \ldots, n) \tag{3}$$

by a holomorphic and invertable transformation

$$x_k = y_k + \varphi_k(\bar{y}), \quad \bar{y} = (y_1, \ldots, y_n), \qquad (k = 1, \ldots, n), \tag{4}$$

where the holomorphic functions φ_k begin with at least the second power. Consequently, x_k's are holomorphic functions of the functions $y_i = c_i e^{\lambda_i t}$ $(i = 1, \ldots, n)$, provided $|y_1| + \cdots + |y_n|$ is small enough.*

The functions φ_k satisfy the first order partial differential equation system $P_k = 0$ $(k = 1, \ldots, n)$ with

$$P_k = \sum_{i=1}^{n} \lambda_i y_i \frac{\partial \varphi_k}{\partial y_i} - \lambda_k \varphi_k - F_k. \tag{5}$$

The theorem of SIEGEL is a considerable extension of a similar one of POINCARÉ [2] and asserts also the fact: the set of points $\bar{\lambda} = (\lambda_1, \ldots, \lambda_n)$ not satisfying (2) has the measure zero, i.e. the theorem is valid for almost all $\bar{\lambda}$.

---

* The holomorphy of the transformation is essential, because such a transformation conserves in some sense the topological configuration of the characteristics near the origin.

Condition (2) implies

$$\varepsilon_{k,\bar{g}} \neq 0, \ (k=1, \ldots, n); \quad |g| \geqq 1 \tag{6}$$

which enables — as it will be presently shown — the formal determination of the coefficients of φ_k.

2. The question arises: how the theorem will be formed when condition (2) — or even (6) — is not satisfied.

In the sequel only the case $n=2$ will be treated.

If the line on the complex plane connecting the points λ_1 and λ_2

1° does not traverse the origin, conditions (2) and (6) are satisfied,

2° traverses the origin and $\varrho = \dfrac{\lambda_1}{\lambda_2} > 0$, the problem was solved by LINDELÖF [3], although condition (6) is not satisfied in all such cases, since $\varepsilon_{k,\bar{g}}$ may vanish for *finitely many* vectors $\bar{g}$.

3° If the line traverses the origin and $\varrho < 0$, then for rational ϱ the numbers $\varepsilon_{k,\bar{g}}$ vanish for *infinitely many* vectors $\bar{g}$.

In the present paper the case 3° will be discussed and solved in part.

By the use of (4), we have

$$\dot{x}_k - \lambda_k x_k - F_k = P_k + \sum_{i=1}^{2} (\dot{y}_i - \lambda_i y_i) \frac{\partial x_k}{\partial y_i}, \quad (k = 1, 2). \tag{7}$$

Since $\det \left\{ \dfrac{\partial x_k}{\partial y_i} \right\} \neq 0$ near the origin, the normalform in question turns out to be (3), provided the system

$$P_k = 0 \qquad (k=1, 2) \tag{8}$$

has some holomorphic solutions φ_k $(k=1, 2)$ at all.

Suppose

$$x_k = y_k + \varphi_k = \sum_{|\bar{g}| \geqq 1} c_{k,\bar{g}} y_{\bar{g}}, \quad y_{\bar{g}} = y_1^{g_1} \cdot y_2^{g_2} \tag{9}$$

with

$$c_{k,\bar{e}_i} = \delta_{i,k}, \quad \bar{e}_1 = (1, 0), \bar{e}_2 = (0, 1), \quad (i, k = 1, 2)$$

then by (9)

$$P_k = \sum_{\bar{g}} \varepsilon_{k,\bar{g}} c_{k,\bar{g}} y_{\bar{g}} - F_k, \quad F_k(\bar{x}) = \tilde{F}_k(\bar{y}), \quad (k = 1, 2)$$

and so from system (8) the coefficients $c_{k,\bar{g}}$ can be determined recursively, provided assumption (6) holds. Namely F_k begins with terms of at least second order and the coefficient of $y_{\bar{g}}$ in $\tilde{F}_k$ is formed from coefficients $c_{k,\bar{h}}$ with $|\bar{h}| < |\bar{g}|$, moreover with $\bar{h} < \bar{g}$ (this means $h_i \leqq g_i$ $(i=1, 2)$ and $h_k < g_k$ for at least one k).

If the condition (2) is not fulfilled, a modified approach must be followed.

Let $\varrho = \dfrac{\lambda_1}{\lambda_2}$ be an arbitrary negative rational number,

$$\varrho = -\frac{p}{q}, \quad (p, q) = 1,$$

i.e. $\lambda_1 = p$, $\lambda_2 = -q$, this can be achieved by a linear transformation of the independent variable. Now

$$\varepsilon_{1,\bar{g}} = g_1 p - g_2 q - p, \quad \varepsilon_{2,\bar{g}} = g_1 p - g_2 q + q$$

and $\varepsilon_{k,\bar{g}} = 0$ involves

$$g_1 = \frac{g_2}{p} q + 1 = lq + 1, g_2 = lp$$

or $(l > 0$ integer$)$

$$g_2 = \frac{g_1}{q} p + 1 = lp + 1, g_1 = lq$$

respectively.

Let the functions

$$Q_1 = \sum_{l=1}^{\infty} a_{1,l} y_1^{lq+1} y_2^{lp}, \quad Q_2 = \sum_{l=1}^{\infty} a_{2,l} y_1^{lq} y_2^{lp+1}$$

— or by the use of the notations

$$\bar{l} = (lq, lp) = l(q, p), \quad a_{i,l} = a_{i,\bar{l}}, \quad (i = 1, 2)$$

$$Q_1 = \sum_{l=1}^{\infty} a_{1,\bar{l}} y_{\bar{l}+\bar{e}_1}, \quad Q_2 = \sum_{l=1}^{\infty} a_{2,\bar{l}} y_{\bar{l}+\bar{e}_2}$$

be introduced, and let (7) be modified as follows:

$$\dot{x}_k - \lambda_k x_k - F_k = R_k + \sum_{i=1}^{2} (\dot{y}_i - \lambda_i y_i - Q_i) \frac{\partial x_k}{\partial y_i}, \quad (k = 1, 2) \tag{10}$$

with

$$R_k = P_k + \sum_{i=1}^{2} Q_i \frac{\partial x_k}{\partial y_i} = P_k + \sum_{i=1}^{2} Q_i \left(\delta_{ki} + \frac{\partial \varphi_k}{\partial y_i} \right), \quad (k = 1, 2). \tag{11}$$

The requested normalform is now

$$\dot{y}_k = \lambda_k y_k + Q_k, \qquad (k = 1, 2), \tag{12}$$

provided the system of partial differential equations $R_k = 0$, $(k = 1, 2)$ has some holomorphic solution. If it has, then by (12) for the function $u = y_1^q y_2^p$ we have

$$\frac{\dot{u}}{u} = q \frac{\dot{y}_1}{y_1} + p \frac{\dot{y}_2}{y_2} = q \left(p + \frac{Q_1}{y_1} \right) + p \left(-q + \frac{Q_2}{y_2} \right)$$

or

$$\dot{u} = uF(u), \quad F(u) = \sum_{l=1}^{\infty} a_l u^l, \quad a_l = q a_{1,l} + p a_{2,l} \tag{13}$$

which is a single holomorphic equation concerning u, whence $u = u_0 = \text{const} \neq 0$, provided $F(u_0) = 0$. Then $u = y_1^q y_2^p = u_0$ is a characteristic (integral) of (12) and the transformation inverse to $x_k = y_k + \varphi_k$ gives a characteristic of (1). Either $F(u) \equiv 0$ (case (A)) and $y_1^q y_2^p = u_0$ are characteristics for an arbitrary u_0, provided $|u_0|$ small enough (saddle-point), or $F(u) \neq 0$ for $u \neq 0$ with sufficiently small $|u|$ (case (B));

then (13) determines u as a function of t and one of y_1 and y_2 can be eliminated and the system (12) reduces to a single equation.

THEOREM. *In the Case* (A) *the system* $R_k=0$, $(k=1, 2)$ *has some holomorphic solution* $\varphi_k(y)$ $(k=1, 2)$.

After some preparatory remarks and notations the proof will be presented in pragraph 3.
Case (B) will be commented and partially solved in paragraph 6.

The structure of Q_k $(k=1, 2)$ has been chosen to involve terms corresponding to the vanishing $\varepsilon_{k,\bar{g}}$ (i.e. $\bar{g}=\bar{l}+\bar{e}_1$ or $\bar{g}=\bar{l}+\bar{e}_2$ respectively). The coefficients $a_{i,\bar{l}}$ can be uniquely determined by means of the system $R_k=0$, $(k=1, 2)$ by the requirement that the terms of Q_k cancel the corresponding terms of $\tilde{F}_k(\bar{y})$ which are "wrong" here. Then there is no obstacle (contradiction) in the formal calculation of the rest of the $c_{k,\bar{g}}$ for which $\varepsilon_{k,\bar{g}}\neq 0$. In details

$$
\begin{aligned}
R_1 &= P_1+Q_1(\bar{y})\left(1+\frac{\partial\varphi_1}{\partial y_1}\right)+Q_2(\bar{y})\frac{\partial\varphi_1}{\partial y_2} = P_1+Q_1+S_1 \\
R_2 &= P_2+Q_1(\bar{y})\frac{\partial\varphi_2}{\partial y_1}+Q_2(\bar{y})\left(1+\frac{\partial\varphi_2}{\partial y_2}\right) = P_2+Q_2+S_2
\end{aligned} \tag{14}
$$

where

$$
S_k=\sum_{|\bar{g}|\geq p+q+2} b_{k,\bar{g}}y_{\bar{g}}, \qquad b_{k,\bar{g}}=\sum_{\bar{l}+\bar{h}=\bar{g}}(h_1a_{1,\bar{l}}+h_2a_{2,\bar{l}})c_{k,\bar{h}}; \quad \bar{h}=(h_1,h_2)
$$

S_k begin — as indicated — with terms of order $p+q+2$. Let

$$
F_k(\bar{x})=\tilde{F}_k(\bar{y})=\sum_{|\bar{g}|\geq 2} f_{k,\bar{g}}y_{\bar{g}} \quad (k=1,2).
$$

Then the system $R_k=0$ $(k=1, 2)$ implies[1]

$$
\left.\begin{aligned}
&\varepsilon_{k,\bar{g}}c_{k,\bar{g}}=f_{k,\bar{g}}, \quad |\bar{g}|<p+q+2 \\
&\varepsilon_{k,\bar{g}}c_{k,\bar{g}}=f_{k,\bar{g}}-b_{k,\bar{g}}, \quad |\bar{g}|\geq p+g+2
\end{aligned}\right\} \quad (k=1,2) \quad \varepsilon_{k,\bar{g}}\neq 0 \tag{15}
$$

In the rest of the cases $\varepsilon_{k,\bar{g}}=0$, and equations

$$
\varepsilon_{k,\bar{g}}c_{k,\bar{g}}=0, \quad \bar{g}=\bar{l}+\bar{e}_k, \quad (k=1,2,;\, l=1,2,\ldots),
$$

which result by putting $f_{k,\bar{l}+\bar{e}_k}-a_{k,\bar{l}}-b_{k,\bar{l}+\bar{e}_k}=0$, show that $c_{k,\bar{l}+\bar{e}_k}$ can be chosen as one wishes. Let their values be chosen as zero. (Later it will be shown the possibility of a different choice as well.)

The determination of the coefficients $c_{k,\bar{g}}$ and $a_{k,\bar{g}}$ proceeds as follows:

$c_{k,\bar{e}_1}$ and $c_{k,\bar{e}_2}$ determine $f_{k,\bar{g}}$ for $|\bar{g}|=2$ and these latter the $c_{k,\bar{g}}$ for $|\bar{g}|=2$. By the last ones $f_{k,\bar{g}}$ are obtained for $|\bar{g}|=3$ and the process goes further until the vector $\bar{g}=\bar{l}+\bar{e}_i$ $(\bar{l}=1)$ comes with $\varepsilon_{i,\bar{g}}=0$. Then $a_{i,\bar{l}}$ $(\bar{l}=1)$ will be equated to $f_{i,\bar{l}+\bar{e}i}$ $(\bar{l}=1)$, by means of which the first "wrong" terms of $\tilde{F}_k$ will be dropped. However, at the same time the first term of S_k enters which is of order at least $p+q+2$.

[1] Or more generally $\varepsilon_{k,\bar{g}}c_{k,\bar{g}}-f_{k,\bar{g}}+a_{k,\bar{l}}\delta_{\bar{g},\bar{l}+\bar{e}_k}+b_{k,\bar{g}}=0, \qquad k=1,2, \ |\bar{g}|\geq 1.$

After adding this to $\tilde{F}_k$, the succeeding coefficients $c_{k,\bar{g}}$ and $a_{i,l}$ will be determined analogously and the process can be continued indefinitely.

The value of $a_{i,l}$ is

$$a_{i,l} = f_{i,l+\bar{e}_i}, \quad (i = 1, 2) \quad (l = 1, 2, \ldots).$$

Namely

$$a_{i,l} = f_{i,l+\bar{e}_i} - b_{i,l+\bar{e}_i},$$

but

$$b_{i,l+\bar{e}_i} = \sum_{\bar{\lambda}+\bar{j}=l+\bar{e}_i} (j_1 a_{1,\bar{\lambda}} + j_2 a_{2,\bar{\lambda}}) c_{i,\bar{j}} = 0; \quad \bar{\lambda} = (\lambda q, \lambda p),\ \bar{j} = (j_1, j_2)$$

since

$$c_{i,\bar{j}} = c_{i,l-\bar{\lambda}+\bar{e}_i} = 0, \quad \text{viz.} \quad l - \bar{\lambda} = [(l-\lambda)q, (l-\lambda)p].$$

3. The PROOF in the case (A) of the convergence of the series $\varphi_k(\bar{y})$ $(k = 1, 2)$ necessitates to look for an estimate of the coefficients $c_{k,\bar{g}}$. Now, if $\varepsilon_{k,\bar{g}}$ does not vanish, it is not small in absolute value. It is at least 1. For the sake of simplicity let the radius of convergence of the series of $F_k(\bar{x})$ be $r_k > 1$ and suppose $|F_k(\bar{x})| < M$ for $|x_k| < 1$. Clearly $M > 1$ can be assumed. Thus by the Chauchy-formula the coefficients $f_{k,\bar{g}}$ of $\tilde{F}_k(\bar{y})$ are $\leqq M$ in absolute value and with the notation $x = x_1 + x_2$ the domination relation

$$F_k \prec \frac{Mx^2}{1-x} \quad (k = 1, 2)$$

holds. Let

$$c_{\bar{g}} = |c_{1,\bar{g}}| + |c_{2,\bar{g}}|, \quad (c_{\bar{e}_i} = 1; \quad i = 1, 2),$$

then

$$x_k = y_k + \varphi_k \prec \sum_{\bar{g}} c_{\bar{g}} y_{\bar{g}}, \quad x = \sum_{k=1}^{2} (y_k + \varphi_k) \prec \sum_{\bar{g}} c_{\bar{g}} y_{\bar{g}}$$

and

$$\tilde{F}_k \prec M \sum_{r=2}^{\infty} \Big(\sum_{\bar{g}} c_{\bar{g}} y_{\bar{g}}\Big)^r,$$

whence

$$|f_{k,\bar{g}}| \leqq M \sum_{\substack{\bar{g}_1 + \cdots + \bar{g}_r = \bar{g} \\ r \geqq 2}} c_{\bar{g}_1} \ldots c_{\bar{g}_r} = f_{\bar{g}}. \tag{16}$$

In the present case (A), $F(u) \equiv 0$, i.e. $qa_{1,l} + pa_{2,l} = 0$ $(l = 1, 2, \ldots)$, which enables the elimination of $a_{2,l}$, say, obtaining

$$b_{k,\bar{g}} = \frac{1}{p} \sum_{l+\bar{h}=\bar{g}} a_{1,l} u_{\bar{g}} c_{k,\bar{h}}, \quad u_{\bar{g}} = g_1 p - g_2 q = u_{\bar{h}} = h_1 p - h_2 q.$$

Let

$$\alpha_l = |a_{1,l}|, \quad b_{\bar{g}} = \frac{1}{p} \sum_{l+\bar{h}=\bar{g}} \alpha_l |u_{\bar{g}}| c_{\bar{h}}, \tag{17}$$

then by (15)

$$c_{\bar{g}} \leqq \varrho_{\bar{g}} (f_{\bar{g}} + b_{\bar{g}}), \quad \varrho_{\bar{g}} = \frac{1}{|\varepsilon_{1,\bar{g}}|} + \frac{1}{|\varepsilon_{2,\bar{g}}|} = \frac{1}{|u_{\bar{g}} - p|} + \frac{1}{|u_{\bar{g}} + q|} \tag{18}$$

Let the positive numbers $\tau_{\bar{g}}$ be recursively defined by

$$\tau_{\bar{g}} = \sum_{\substack{\bar{g}_1+\cdots+\bar{g}_r=\bar{g} \\ r\geqq 2}} \tau_{\bar{g}_1} \ldots \tau_{\bar{g}_r}, \quad \tau_{\bar{e}_1} = \tau_{\bar{e}_2} = 1 \tag{19}$$

then the estimate

$$c_{\bar{g}} \leqq (\lambda M)^{|\bar{g}|-1} \tau_{\bar{g}} \tag{20}$$

is asserted, where $\lambda > 1$ is a number independent of $\bar{g}$ and to be determined later.

The proof proceeds by induction. (20) is valid for $\bar{g} = \bar{e}_i$ $(i = 1, 2)$. Suppose it holds for $\bar{h} < \bar{g}$, then by (16) and (19) for $\bar{l} < \bar{g}$

$$\alpha_l = |f_{1,\, l+\bar{e}_1}| \leqq M \sum_{\substack{\bar{g}_1+\cdots+\bar{g}_r=\bar{l}+\bar{e}_1 \\ r\geqq 2}} c_{\bar{g}_1} \ldots c_{\bar{g}_r} \leqq M \sum (\lambda M)^{|\bar{g}_1|-1} \ldots (\lambda M)^{|\bar{g}_r|-1} \tau_{\bar{g}_1} \ldots \tau_{\bar{g}_r} \leqq$$

$$\leqq M^{|\bar{l}|}\, \lambda^{|\bar{l}|-1} \tau_{\bar{l}+\bar{e}_1}. \tag{21}$$

Thus by (18) and (21)

$$c_{\bar{g}} \leqq \varrho_{\bar{g}} M \sum_{\substack{\bar{g}_1+\ldots+\bar{g}_r=\bar{g} \\ r\geqq 2}} c_{\bar{g}_1} \ldots c_{\bar{g}_r} + \varrho_{\bar{g}} \sum_{\bar{l}+\bar{h}=\bar{g}} \alpha_l |u_{\bar{g}}|\, c_{\bar{h}} \leqq$$

$$\leqq \varrho_{\bar{g}} M \sum_{\substack{\bar{g}_1+\ldots+\bar{g}_r=\bar{g} \\ r\geqq 2}} (\lambda M)^{|\bar{g}_1|-1} \ldots (\lambda M)^{|\bar{g}_r|-1} \tau_{\bar{g}_1} \ldots \tau_{\bar{g}_r} +$$

$$+ \varrho_{\bar{g}} \sum_{\bar{l}+\bar{h}=\bar{g}} |u_{\bar{g}}|\, \tau_{\bar{l}+\bar{e}_1} \tau_{\bar{h}} M^{|\bar{l}|}\, \lambda^{|\bar{l}|-1} (\lambda M)^{|\bar{h}|-1} \leqq$$

$$\leqq M^{|\bar{g}|-1}\, \lambda^{|\bar{g}|-2} (\varrho_{\bar{g}} \tau_{\bar{g}} + \varrho_{\bar{g}} \sum_{\bar{l}+\bar{h}=\bar{g}} |u_{\bar{g}}|\, \tau_{\bar{l}+\bar{e}_1} \tau_{\bar{h}}) \leqq$$

$$\leqq M^{|\bar{g}|-1}\, \lambda^{|\bar{g}|-2} (2\tau_{\bar{g}} + K \sum_{\bar{l}+\bar{h}=\bar{g}} \tau_{\bar{l}+\bar{e}_1} \tau_{\bar{h}}) \qquad (\varrho_{\bar{g}} \leqq 2\,!)$$

since

$$\varrho_{\bar{g}} |u_{\bar{g}}| = \left(\frac{1}{|u_{\bar{g}} - p|} + \frac{1}{|u_{\bar{g}} + q|} \right) |u_{\bar{g}}| < K, \quad (u_{\bar{g}} = g_1 p - g_2 q)$$

where K is a positive constant independent of $\bar{g}$.

In the next paragraph it will be shown that

$$\sum_{\bar{l}+\bar{h}=\bar{g}} \tau_{\bar{l}+\bar{e}_1} \tau_{\bar{h}} \leqq \mu \tau_{\bar{g}} \tag{22}$$

for some $\mu > 1$ independent of $\bar{g}$. Then putting $2 + K\mu = \lambda$ we have

$$c_{\bar{g}} \leqq (\lambda M)^{|\bar{g}|-1} \tau_{\bar{g}}$$

what was to be proved.

Now define $\psi(\bar{y})$ as

$$\psi = \sum_{\bar{g}} \tau_{\bar{g}} y_{\bar{g}} \tag{23}$$

then with $y = y_1 + y_2$,

$$\psi = y + \sum_{r=2}^{\infty} \psi^r = y + \frac{\psi^2}{1-\psi} < \frac{y + \psi^2}{1-\psi} \tag{24}$$

and $\psi \prec \chi$ provided the dominant $\chi(y)$ is defined by

$$\chi = \frac{y+\chi^2}{1-\chi}$$

Then

$$(1-4\chi)^2 = 1-8y, \quad 1+8\chi \prec (1-4\chi)^{-2} = (1-8y)^{-1}, \quad \chi \prec \frac{y}{1-8y}$$

Consequently

$$\psi \prec \frac{y}{1-8y}$$

and so by (20)

$$\sum_{\bar{g}} c_{\bar{g}} y_{\bar{g}} \prec \sum_{\bar{g}} (\lambda M)^{|\bar{g}|-1} \tau_{\bar{g}} y_{\bar{g}} = (\lambda M)^{-1} \psi(\lambda M y) \prec \frac{y}{1-8\lambda M y}$$

i.e. the series of φ_k are (absolutely and uniformly) convergent provided

$$|y_1| + |y_2| < \frac{1}{8\lambda M}.$$

4. Proof of inequality (22). Relation (24) defines ψ as a holomorphic function of y

$$\psi = \sum_{n=1}^{\infty} c_n y^n. \tag{25}$$

By (23)

$$\tau_{\bar{g}} = c_{\bar{g}} \binom{g}{g_1}, \quad \bar{g} = (g_1, g_2), \quad g = |\bar{g}| = g_1 + g_2$$

and

$$\frac{\tau_{l+\bar{e}_1}}{\tau_l} = \frac{c_{l(q+p)+1} \binom{l(q+p)+1}{lq+1}}{c_{l(q+p)} \binom{l(q+p)}{lq}} = \frac{c_{l(q+p)+1}}{c_{l(q+p)}} \frac{l(q+p)+1}{lq+1}$$

The ratio $\frac{\tau_{l+\bar{e}_1}}{\tau_l} > 0$ remains bounded as $l \to \infty$ if and only if $\frac{c_{n+1}}{c_n}$ has the same property as $n \to \infty$. But it has, as it will be inferred presently. By (24) ψ satisfies $2\psi^2 - (1+y)\psi + y = 0$, whence

$$\psi = \frac{1+y \pm \sqrt{(1+y)^2 - 8y}}{4}.$$

On account of $\psi(0) = 0$ the sign — is valid here, giving

$$\psi = \frac{1}{4} + \frac{y}{4} - \frac{1}{4}(y^2 - 6y + 1)^{1/2}, \tag{26}$$

but

$$y^2 - 6y + 1 = (y-\sigma)\left(y - \frac{1}{\sigma}\right) = \left(1 - \frac{y}{\sigma}\right)(1-\sigma y), \quad \sigma = 3 + 2\sqrt{2}$$

and so

$$(y^2-6y+1)^{1/2}=\left(1-\frac{y}{\sigma}\right)^{1/2}(1-\sigma y)^{1/2}=\sum_{i=0}^{\infty}\binom{\frac{1}{2}}{i}\left(-\frac{y}{\sigma}\right)^i\cdot\sum_{k=0}^{\infty}\binom{\frac{1}{2}}{k}(-\sigma y)^k=$$

$$=\sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\binom{\frac{1}{2}}{i}\binom{\frac{1}{2}}{k}\left(-\frac{y}{\sigma}\right)^i(-\sigma y)^k=\sum_{n=0}^{\infty}y^n\sum_{i=0}^{n}\binom{\frac{1}{2}}{i}\binom{\frac{1}{2}}{n-i}(-1)^{n-2i}\sigma^{n-2i}=$$

$$(i+k=n)$$

$$=\sum_{n=0}^{\infty}(-1)^n\sigma^n y^n\sum_{i=0}^{n}\binom{\frac{1}{2}}{i}\binom{\frac{1}{2}}{n-i}\sigma^{-2i},$$

whence with regard to (25)—(26)

$$c_n=(-1)^{n-1}\frac{\sigma^n}{4}\sum_{i=0}^{n}\binom{\frac{1}{2}}{i}\binom{\frac{1}{2}}{n-i}\sigma^{-2i}.$$

Consider now the following expression

$$\sum_{i=0}^{n}\binom{\frac{1}{2}}{i}\binom{\frac{1}{2}}{n-i}\sigma^{-2i}=\binom{\frac{1}{2}}{n}\sum_{i=0}^{n}\delta(i,n)\sigma^{-2i}=\binom{\frac{1}{2}}{n}(d_{n1}+d_{n2}+d_{n3}),$$

where

$$\delta(i,n)=\frac{\binom{\frac{1}{2}}{i}\binom{\frac{1}{2}}{n-i}}{\binom{\frac{1}{2}}{n}}$$

and

$$d_{n1}=\sum_{i=0}^{[\sqrt{n}]}\delta(i,n)\sigma^{-2i},\quad d_{n2}=\sum_{i=[\sqrt{n}]+1}^{n-[\sqrt{n}]}\delta(i,n)\sigma^{-2i},\quad d_{n3}=\sum_{i=n-[\sqrt{n}]+1}^{n}\delta(i,n)\sigma^{-2i}.$$

It will be shown that

$$d_{ni}\to 0,\quad n\to\infty\quad (i=2,3)$$

while

$$d_{n1}\to\frac{\sqrt{\varrho^2-1}}{\varrho},\quad n\to\infty.$$

Here

$$\binom{\frac{1}{2}}{i}=(-1)^{i-1}\frac{(2i-3)!!}{i!\,2^i}=(-1)^{i-1}\frac{(2i-2)!}{i!(i-1)!\,2^{2i-1}}.$$

If i is a large number, the Stirling-formula $(n!\sim\sqrt{2\pi}n^{n+1/2}e^{-n})$ gives

$$\binom{\frac{1}{2}}{i}\sim(-1)^{i-1}\frac{\sqrt{2\pi}\,(2i-2)^{2i-2+1/2}e^{-(2i-2)}}{(\sqrt{2\pi})^2 i^{i+1/2}(i-1)^{i-1+1/2}e^{-(2i-1)}2^{2i-1}}=$$

$$=(-1)^{i-1}\frac{e}{\sqrt{2\pi}}\underbrace{\left(1-\frac{1}{i}\right)^i}_{\sim e^{-1}}\frac{1}{i^{1/2}(i-1)}\sim(-1)^{i-1}\frac{1}{2\sqrt{\pi}}i^{-3/2}.$$

1° In the interval $\sqrt{n} \leqq i \leqq n-\sqrt{n}$ the numbers i, n, $n-i$ are all large if so is n. Therefore

$$\delta(i,n) \sim \frac{(-1)^{i-1}\frac{1}{2\sqrt{\pi}}\, i^{-3/2}(-1)^{n-i-1}\frac{1}{2\sqrt{\pi}}\,(n-i)^{-3/2}}{(-1)^{n-1}\frac{1}{2\sqrt{\pi}}\, n^{-3/2}} = -\frac{1}{2\sqrt{\pi}}\left[\frac{n}{i(n-i)}\right]^{3/2}.$$

The minimum of $i(n-i)$ in this interval is $\sqrt{n}\,(n-\sqrt{n})$ and

$$\frac{n}{\sqrt{n}\,(n-\sqrt{n})} = \frac{1}{\sqrt{n}\,(1-n^{-1/2})} \to 0, \quad n \to \infty.$$

Thus to every $\varepsilon > 0$ one can choose $n = n_0$ such that

$$|\delta(i,n)| < \frac{\varepsilon}{c}, \quad n > n_0, \quad \text{where} \quad c = \sum_{i=0}^{\infty} \sigma^{-2i},$$

therefore

$$|d_{n2}| < \frac{\varepsilon}{c} \sum_{i=[\sqrt{n}]+1}^{n-[\sqrt{n}]} \sigma^{-2i} < \frac{\varepsilon}{c} \sum_{i=0}^{\infty} \sigma^{-2i} = \varepsilon$$

i.e. $d_{n2} \to 0$ as $n \to \infty$.

2° In the interval $n-\sqrt{n} \leqq i \leqq n$ n and i are large. So

$$\binom{\frac{1}{2}}{i}:\binom{\frac{1}{2}}{n} \sim \frac{(-1)^{i-1}\frac{1}{2\sqrt{\pi}}\, i^{-3/2}}{(-1)^{n-1}\frac{1}{2\sqrt{\pi}}\, n^{-3/2}} = (-1)^{n-i}\left(\frac{n}{i}\right)^{3/2}.$$

The identity

$$\binom{\frac{1}{2}}{k} = \frac{(-1)^{k-1}(2k-1)!}{k!(k-1)!\,2^{2k-1}(2k-1)} = (-1)^{k-1}\frac{\binom{2k-1}{k}}{2^{2k-1}}\cdot\frac{1}{2k-1}$$

implies

$$\left|\binom{\frac{1}{2}}{k}\right| < \frac{1}{2k-1}.$$

By the use of this inequality

$$|\delta(i,n)| \sim \left(\frac{n}{i}\right)^{3/2}\left|\binom{\frac{1}{2}}{n-i}\right| < \left(\frac{n}{i}\right)^{3/2}\frac{1}{2(n-i)-1} < K = \text{const.}$$

So given $\varepsilon > 0$,

$$|d_{n3}| < K \sum_{i=n-[\sqrt{n}]+1}^{n} \sigma^{-2i} < \varepsilon$$

if n is large enough.

3° In the interval $0 \leqq i \leqq \sqrt{n}$ n and $n-i$ are large, thus

$$\varepsilon(i, n) = \frac{\binom{\frac{1}{2}}{n-i}}{\binom{\frac{1}{2}}{n}} \sim (-1)^i \left(\frac{n}{n-i}\right)^{3/2}.$$

Here

$$1 \leqq \frac{n}{n-i} < \frac{n}{n-\sqrt{n}} = \frac{1}{1-n^{-1/2}}$$

so

$$\varepsilon(i, n) \to (-1)^i, \quad \delta(i, n) \to (-1)^i \binom{\frac{1}{2}}{i}, \quad (n \to \infty)$$

and

$$d_{n1} \to \sum_{i=0}^{\infty} (-1)^i \binom{\frac{1}{2}}{i} \sigma^{-2i} = (1-\sigma^{-2})^{1/2} = \frac{\sqrt{\sigma^2-1}}{\sigma}, \quad n \to \infty.$$

Finally

$$c_n \sim (-1)^{n-1} \frac{1}{4} \sigma^n (-1)^{n-1} \frac{1}{2\sqrt{\pi}} n^{-3/2} \frac{\sqrt{\sigma^2-1}}{\sigma} = \frac{1}{8\sqrt{\pi}} \sigma^{n-1} \sqrt{\sigma^2-1}\, n^{-3/2}$$

with the consequence

$$\frac{c_{n+1}}{c_n} \sim \sigma, \quad n \to \infty,$$

what was to be proved.

5. As earlier mentioned, the normalform (12) can be replaced by a simpler one containing only two numbers $a_{i,l}$. Namely for $\bar{g} = \bar{l} + \bar{e}_k$ the coefficient $c_{k,\bar{g}}$ was undetermined and chosen to be 0. Now an alternative choice will be made. First suppose $a_{i,l} = f_{i,l+\bar{e}_i} \neq 0$ for $i = 1, 2$; $l = 1$ and put $a_i = a_{i,l}$ $(i = 1, 2;\ l = 1)$. Let $b_{k,\bar{g}}$ be defined as

$$b_{k,\bar{g}} = [(g_1 - q)a_1 + (g_2 - p)a_2]c_{k,\bar{g}-\bar{e}}, \quad \bar{e} = (q, p), \quad (k = 1, 2) \tag{27}$$

Then by the second of the equations (15) the first two undetermined coefficients $c_{k,\bar{e}+\bar{e}_k}$ $(k = 1, 2)$ can be determined by the requirement

$$f_{k,\bar{g}} - b_{k,\bar{g}} = 0 \quad (k = 1, 2) \tag{28}$$

for $\bar{g} = 2\bar{e} + \bar{e}_k$ $(k = 1, 2)$. Then $c_{k,2\bar{e}+\bar{e}_k}$ $(k = 1, 2)$ are yet undetermined and will be determined from (28) applied to $\bar{g} = 3\bar{e} + \bar{e}_k$ $(k = 1, 2)$. The process can be continued indefinitely without introducing further numbers $a_{i,l}$.[2] The corresponding normalform reads now as

$$\begin{aligned} \dot{y}_1 &= y_1(p + a_1 v) \\ \dot{y}_2 &= y_2(-q + a_2 v) \end{aligned} \qquad (v = y_1^q y_2^p) \tag{29}$$

[2] If $(g_1 - q)a_1 + (g_2 - p)a_2 = 0$, the value of $c_{k,\bar{g}-\bar{e}}$ can be chosen as 0.

If $f_{i,l+\bar{e}_i}=0$ $(i=1,2;\ l=1)$ or one of them is zero, then the role played by a_1, a_2 will be taken over by higher coefficients $a_{.,l}$ ($a_{i,\bar{e}}$ with larger l than 1). The proof of the convergence of the series of φ_i is now more simple than earlier. It does not necessitate the argument of paragraph 4. — System (29) can be integrated in closed form, what does not imply the same concerning to the original system (1)[3].

6. Comments on case (B). It will be shown here that there are cases where the transformation is holomorphic and that also cases exist where it is not so.

1° As a rather general example take the system

$$\begin{aligned}\dot{x}_1 &= x_1\left(1+uF_1(u)\right)\\ \dot{x}_2 &= x_2\left(-1+uF_2(u)\right)\end{aligned}\qquad (u=x_1x_2),\quad F_i(u)=\sum_{n=0}^{\infty}A_n^{(i)}u^n;\quad (i=1,2). \tag{30}$$

When can this system be transformed into (29) (with $p=q=1$) by a holomorphic transformation? The following assertion holds: To the existence of such a transformation it is necessary and sufficient that

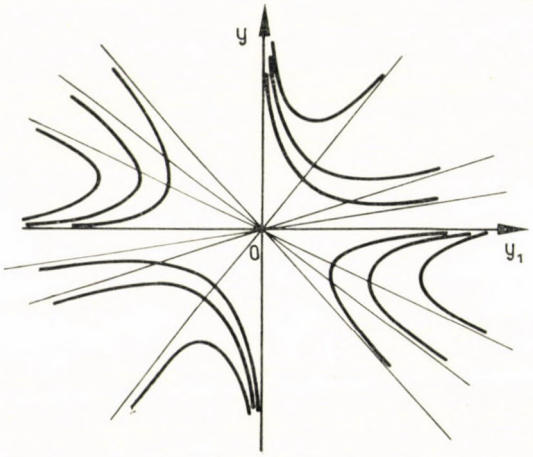
$$A_1^{(1)}=-A_1^{(2)}$$

holds.

Proof. Let the inverse of the transformation be assumed as

$$y_i=x_if_i(u),\quad f_i(u)=1+\sum_{n=1}^{\infty}B_n^{(i)}u^n,\quad (i=1,2) \tag{31}$$

[3] E.g. for $p=q=1$, $a_1=a_2=a$ we have in polarcoordinates $r=\sqrt{\frac{2}{a}}\left[\sin 2\varphi\left(c+\log|\operatorname{tg}\varphi|\right)\right]^{-\frac{1}{2}}$



the phase-plane picture of which is shown on the Figure 1. Note that one asymptote is an arbitrary straight line through 0 (a different type of saddle-point).

Then (29) transformed by (31) looks like (30) with

$$(32)\qquad \left.\begin{aligned} F_1 &= f\frac{a_1 f+ug}{(uf)'}\\ F_2 &= f\frac{a_2 f-ug}{(uf)'}\end{aligned}\right\} f=f_1 f_2,\quad g=a_1 f_2' f_1-a_2 f_1' f_2 \quad \left(' = \frac{d}{du}\right),$$

whence

$$(33)\qquad \begin{aligned} \frac{af^2}{(uf)'} &= F,\quad a=\frac{a_1+a_2}{2},\quad F=\frac{F_1+F_2}{2}\\ \frac{aufg}{(uf)'} &= G,\quad G=\frac{a_2F_1-a_1F_2}{2}.\end{aligned}$$

Now (in case (B)) $a\neq 0$. Hence for $uf=h$

$$\frac{dh}{h^2}=\frac{a}{u^2F}\,du,\quad h=-\frac{1}{\int\frac{a}{u^2F}\,du+C},\quad C=\text{const}$$

$$f=-\frac{1}{u\int\frac{a}{u^2F}\,du+Cu}.$$

But

$$F=\sum_{n=0}^{\infty}C_n u^n,\quad C_n=\frac{A_n^{(1)}+A_n^{(2)}}{2}.$$

If $C_0=\dfrac{A_0^{(1)}+A_0^{(2)}}{2}\neq 0$, then $\dfrac{1}{F}=\sum_{n=0}^{\infty}\gamma_n u^n$, $\gamma_0=\dfrac{1}{C_0}$, $\gamma_1=-\dfrac{C_1}{C_0^2},\ldots$

$$\frac{1}{u^2F}=\frac{\gamma_0}{u^2}+\frac{\gamma_1}{u}+\gamma_2+\gamma_3 u+\ldots$$

$$u\int\frac{du}{u^2F}=-\gamma_0+\gamma_1 u\log u+\gamma_2 u^2+\ldots$$

which is holomorphic together with f if and only if $\gamma_1=C_1=0$, i.e.

$$A_1^{(1)}=-A_1^{(2)}.$$

Then

$$g=\frac{1}{a}\frac{(uf)'}{uf}G.$$

But assuming f in the form

$$f=1+\sum_{n=1}^{\infty}D_n u^n$$

we have

$$\frac{(uf)'}{uf} = \frac{1+2D_1u+\dots}{u(1+D_1u+\dots)} = \frac{1+E_1u+E_2u^2+\dots}{u} = \frac{1}{u}+E_1+E_2u+\dots$$

and

$$g = \frac{1}{a}\left(\frac{1}{u}+E_1+E_2u+\dots\right)\times$$

$$\times\left(\frac{a_2A_0^{(1)}-a_1A_0^{(2)}}{2}+\frac{a_2A_1^{(1)}-a_1A_1^{(2)}}{2}u+\frac{a_2A_2^{(1)}-a_1A_2^{(2)}}{2}u^2+\dots\right) =$$

$$= \frac{1}{a}\frac{a_2A_0^{(1)}-a_1A_0^{(2)}}{2}\frac{1}{u}+\sum_{n=0}^{\infty}\alpha_nu^n$$

which is holomorphic since $a_i = A_0^{(i)}$ $(i=1, 2)$ (it can be shown easily). We have now

$$\left.\begin{aligned} f' &= f_2'f_1+f_1'f_2 \\ g &= a_1f_2'f_1-a_2f_1'f_2 \end{aligned}\right\} \quad \text{(holomorphic)}$$

whence by adding, subtracting and dividing by f

$$2a\frac{f_2'}{f_2} = a_2\frac{f'}{f}+\frac{g}{f}$$

$$2a\frac{f_1'}{f_1} = a_1\frac{f'}{f}-\frac{g}{f},$$

hence

$$f_2^{2a} = c_1f^{a_2}e^{\int\frac{g}{f}du}$$

$$f_1^{2a} = c_2f^{a_1}e^{-\int\frac{g}{f}du}$$

which shows the analyticity of f_1 and f_2.

Example. If $a_1 = a_2 = a$ and $f_1 = f_2 = \dfrac{1}{1-u}$, then $F_1 = F_2 = \dfrac{1}{1-u^2}$. The inverse transformation reads as

$$x_1 = \frac{-1+\sqrt{1+4v}}{2y_2}, \quad x_2 = \frac{-1+\sqrt{1+4v}}{2y_1} \quad (v = y_1y_2)$$

2° On the other hand the system

$$\left.\begin{aligned} \dot{x}_1 &= x_1(1+\varepsilon x_2) \\ \dot{x}_2 &= x_2(-1+x_1x_2) \end{aligned}\right\} \quad (\varepsilon \neq 0) \tag{34}$$

cannot be turned into the system

$$\dot{y}_1 = y_1, \quad \dot{y}_2 = y_2(-1+y_1y_2) \tag{35}$$

in a holomorphic way, which is the unique possible normalform of (34).

Let the inverse transformation of the form

$$y_i = x_i + \psi_i(\bar{x}) \quad (i = 1, 2)$$

Then y_1 can be written as

$$y_1 = x_1 + \psi_1(\bar{x}) = \sum_{k=1}^{\infty} p_k(\bar{x}) \tag{36}$$

where $p_1 = x_1$ and

$$p_{2n} = \sum_{k=1}^{n} \alpha_{n,k}\varepsilon^{2n+1-2k}x_1^k x_2^{2n-k}, \quad p_{2n+1} = \sum_{k=1}^{n} \beta_{nk}\varepsilon^{2n+2-2k}x_1^k x_2^{2n-k+1} \quad (n \geqq 1)$$

then

$$p_2 = \varepsilon x_1 x_2, \quad p_3 = \frac{\varepsilon^2}{2} x_1 x_2^2 \quad \left(\text{i.e.} \quad \alpha_{11} = \varepsilon, \quad \beta_{11} = \frac{\varepsilon^2}{2}\right).$$

By putting (36) into (35) and using (34) we have

$$(x_1 + \varepsilon x_1 x_2)[\ldots + p'_{2n} + p'_{2n+1}] + (-1 + x_1 x_2) x_2 [\ldots + p^*_{2n-1} + p^*_{2n} + p^*_{2n+1}] = p_{2n+1}$$

where $' = \dfrac{\partial}{\partial x_1}$, $^* = \dfrac{\partial}{\partial x_2}$. This involves

$$\begin{aligned} &\varepsilon x_1 x_2 p'_{2n} + x_1 p'_{2n+1} - x_2 p^*_{2n+1} - p_{2n+1} + x_1 x_2^2 p^*_{2n-1} = 0 \\ &\varepsilon x_1 x_2 p'_{2n+1} + x_1 p'_{2n+2} - x_2 p^*_{2n+2} - p_{2n+2} + x_1 x_2^2 p^*_{2n} = 0, \end{aligned} \tag{40}$$

whence

$$\begin{aligned} \beta_{nk} &= \frac{k}{2(n+1-k)}\alpha_{nk} + \frac{2n-k}{2(n+1-k)}\beta_{n-1,k-1} \quad (k = 2, \ldots, n) \\ \beta_{n1} &= \frac{1}{2n}\alpha_{n1} \\ \alpha_{n+1,1} &= \frac{\beta_{n1}}{2n+1} \\ \alpha_{n+1,k} &= \frac{k}{2n+3-2k}\beta_{nk} + \frac{2n+1-k}{2n+3-2k}\alpha_{n,k-1} \quad (k = 2, \ldots, n). \\ \alpha_{n+1,n+1} &= n\alpha_{nn} \end{aligned} \tag{41}$$

The last formula of (41) gives $\alpha_{nn} = n!\,\varepsilon$ what reveals the divergence of the series $\sum_{k=1}^{\infty} p_k$.

7. If $\varrho = \dfrac{\lambda_1}{\lambda_2} < 0$ is an irrational number, the values $\varepsilon_{k,\bar{g}}$ do not vanish and the *formal* calculation of the coefficients $c_{k,\bar{g}}$ goes from $P_k = 0$ $(k = 1, 2)$ without any obstacle, nevertheless the series of $\varphi_k(y)$ $(k = 1, 2)$ may converge or not and the above modification of $P_k = 0$ to $R_k = 0$ by the Q_k's is meaningless. An example where convergence holds, is $\varrho = -\sqrt{2}$, i.e. $\lambda_1 = \sqrt{2}$, $\lambda_2 = -1$. Then $\varepsilon_{1,\bar{g}} = \sqrt{2}(g_1 - 1) - g_2$, $\varepsilon_{2,\bar{g}} = \sqrt{2}g_1 - (g_2 - 1)$. As well known

$$\left|\sqrt{2} - \frac{g_2}{g_1 - 1}\right| \geqq \frac{c}{(g_1 - 1)^2} \quad \text{(for some } c > 0)$$

thus $\varepsilon_{1,\bar{g}} \geqq \dfrac{c}{g_1 - 1} \geqq \dfrac{c}{(g_1 + g_2)^\nu} = c|\bar{g}|^{-\nu}$ $(\nu > 1)$

and just so $|\varepsilon_{2,\bar{g}}| \geqq c|\bar{g}|^{-\nu}$, i.e. condition (2) is fulfilled, consequently the series of the φ_k converge.

REFERENCES

[1] SIEGEL, C. L.: Über die Normalform analytischer Differentialgleichungen in der Nähe einer Gleichgewichtslösung. *Nachr. d. Akad. d. Wiss. in Göttingen* **5** (1952), 21—30.

[2] POINCARÉ, H.: Sur les propriétés des fonctions définies par les équations aux différences partielles". Thèse, Paris, 1879.

[3] LINDELÖF, E.: Sur la forme des integrales des équations différentielles au voisinage des points singuliers. *Acta Soc. Sci. Fenn.* **22** (1897), 1—26.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received June 22, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 1(970) 353—368.

О РЕШЕНИЯХ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ ТИПА СВЕРТКИ

И. ДЬЕРИ

В этой работе мы будем изучать решения уравнений

$$y(t) = f(t) + \int_0^t y(t-\tau)\,\varphi(\tau)\,d\tau \tag{1}$$

и

$$y(t) = f(t) + \int_0^t y(t-\tau)\,d\Phi(\tau) \tag{2}$$

При этих исследованиях мы будем пользоваться неравенствами Бихари [3], Беллмана [1] и Мышкиса [4]. С помощью неравенств Беллмана и Бихари оценим величины

$$|z(t)-y(t)|, \quad \int_0^T y^2(t)\,dt \quad \text{и} \quad \int_0^T [z(t)-y(t)]^2\,dt$$

где $z(t)$ является решением уравнения:

$$z(t) = g(t) + \int_0^t z(t-\tau)\,\psi(\tau)\,d\tau.$$

С помощью леммы Мышкиса мы докажем, что если $f(t)$ непрерывна и $\Phi(t)$ является функцией с ограниченным изменением на отрезке $0 \leqq t \leqq T$ кроме того $\Phi(t)$ непрерывна на отрезке $0 \leqq t \leqq \delta (\leqq T)$, тогда уравнение (2) имеет единственное непрерывное решение.

Применяя лемму Мышкиса, затем лемму Беллмана оценим величину $|z(t)-y(t)|$, где $y(t)$ есть решение уравнения (2), а $z(t)$ удовлетворяет уравнению

$$z(t) = g(t) + \int_0^t z(t-\tau)\,d\Psi(\tau)$$

1. Основные леммы

Следующие результаты мы будем использовать различными способами. Первая и вторая из лемм играют большую роль при исследовании устойчивости решений дифференциальных уравнений и их зависимости от начальных условий и параметров. Эти леммы доказали Бихари [3] и Беллман [1].

Лемма, принадлежащая Бихари:

Лемма 1.1. *Пусть* $u(t)$ *и* $v(t)$ *положительные непрерывные функции на отрезке* $a \leqq t \leqq b$, $k > 0$, $m \geqq 0$; *функция* $g(t)$ *возрастающая, неотрицательна и непрерывна при* $t \geqq 0$.

Тогда из неравенства

$$u(t) \leqq k + m \int_a^t v(s) g(u(s))\, ds \quad (a \leqq t \leqq b) \tag{1.1}$$

следует неравенство

$$u(t) \leqq G^{-1}\left(G(k) + m \int_a^t v(s)\, ds\right) \quad (a \leqq t \leqq b' \leqq b) \tag{1.2}$$

Здесь

$$G(u) = \int_{u_0}^{u} \frac{dt}{g(t)} \quad (u \geqq 0, u_0 > 0) \tag{1.3}$$

и $G^{-1}(u)$ *есть функция, обратная функции* $G(u)$.

Лемма принадлежащая Беллману:

Лемма 1.2. *Пусть* $u(t), v(t), k, m, a, b$ *таковы же, как и выше, тогда из неравенства*

$$u(t) \leqq k + m \int_a^t u(s) v(s)\, ds \quad (a \leqq t \leqq b)$$

следует неравенство

$$u(t) \leqq k \cdot e^{m \int_a^t v(\tau)\, d\tau} \quad (a \leqq t \leqq b)$$

Очевидно, что лемма Беллмана является специальным случаем леммы Бихари.

Наконец мы формулируем принадлежащую Мышкису лемму [4]. Эта лемма играет большую роль при исследовании линейных дифференциальных уравнений с запаздывающим аргументом. Эта лемма еще не была применена при исследовании уравнения восстановления. Мы приведем несколько определений, которые будем применять к лемме Мышкиса.

Определение 1.1. Пусть на отрезке $[a, b]$ $(-\infty < a < b < \infty)$ задана функция $f(t)$ принимающая конечные значения. Тогда под полным изменением этой функции на $[a, b]$, $\bigvee_a^b (f)$, понимают $\sup \sum_{i=1}^{m} |f(t_i) - f(t_{i-1})|$ при всевозможных разложениях $[a, b]$ на конечное число отрезков

$$a = t_0 < t_1 < \ldots < t_m = b.$$

Если полное изменение конечно, то говорят, что $f(t)$ имеет на $[a, b]$ конечное изменение. Если f зависит и от других аргументов, то мы будем писать $\bigvee_{t=a}^{b} (f)$, если полное изменение берется по t.

Определение 1. 2. Пусть на отрезке $[a, b]$ дана непрерывная функция, $f(t)$. Тогда ее модулем непрерывности называют следующую функцию $\omega(\delta; f)$ определенную для $0 < \delta < \infty$:

$$\omega(\delta; f) = \max_{|t_2 - t_1| \leqq \delta} |f(t_2) - f(t_1)|.$$

Лемма 1. 3. *Пусть на отрезке* $[a, b]$ *даны непрерывная функция* $f(t)$ *и функции с конечным изменением* $g_1(t)$, $g_2(t)$. *Тогда для любого натурального* n *будет*:

$$\left| \int_a^b f(t)\, dg_2(t) - \int_a^b f(t)\, dg_1(t) \right| \leqq \omega\left(\frac{b-a}{n}; f\right)\left[\bigvee_a^b (g_1) + \bigvee_a^b (g_2)\right] +$$

$$+ |f(b)|\,|g_2(b) - g_1(b)| + |f(a)|\,|g_2(a) - g_1(a)| +$$

$$+ \frac{n}{b-a}\,\omega\left(\frac{b-a}{n}; f\right) \int_a^b |g_2(t) - g_1(t)|\, dt.$$

2. Существование и единственность

В этом параграфе мы установим простую теорему существования и единственности. Сперва мы докажем лемму:

Лемма 2. 1. *Если*

(а) $u(t)$ *непрерывна на отрезке* $[0, a]$, $u(0) = 0$;
(б) $v(t)$ *имеет ограниченную вариацию на отрезке* $[0, b]$,
(в) $v(t)$ *непрерывна на отрезке* $[0, b-a]$, *то функция*

$$U(t) = \int_{t-a}^{t} u(t-\tau)\, dv(\tau) \tag{2. 1}$$

непрерывна на отрезке $[a, b]$.

Доказательство: Из свойства интеграла Стильтеса мы получаем

$$U(t) = \int_{t-a}^{t} u(t-\tau)\, dv(\tau) = -\int_0^a u(s)\, dv(t-s),$$

и таким образом, для $a \leqq t < b$ и $0 < h$

$$|U(t+h) - U(t)| = \left| \int_0^a U(s)\, d[v(t+h-s) - v(t-s)] \right|$$

с условием $a < t+h < b$.

Так как $u(s)$ непрерывна, и $v(t)$ имеет ограниченную вариацию, то из леммы 1. 3. вытекает, что для любого натурального n будет:

$$|U(t+h)-U(t)| \leqq \omega\left(\frac{a}{n};u\right)\left[\bigvee_{s=0}^{a}\left(v(t+h-s)\right)+\bigvee_{s=0}^{a}\left(v(t-s)\right)\right]+$$

$$+|u(a)|\,|v(t+h-a)-v(t-a)|+|u(0)|\,|v(t+h)-v(t)|+ \tag{2. 2}$$

$$+\frac{n}{a}\,\omega\left(\frac{a}{n};u\right)\int_0^a |v(t+h-s)-v(t-s|\,ds.$$

Пусть $0<\varepsilon$ фиксированно и исследуем члены на правой стороне (2. 2). Так как на отрезке $0\leqq s\leqq a$ функция $u(s)$ непрерывная и $v(s)$ имеет ограниченную вариацию на отрезке $0\leqq s\leqq b$, то существует такое натуральное число N, что

$$\omega\left(\frac{a}{N};u\right)\left[\bigvee_{s=0}^{a}\left(v(t+h-s)\right)+\bigvee_{s=0}^{a}\left(v(t-s)\right)\right] \leqq \omega\left(\frac{a}{N};u\right)2\cdot\bigvee_{s=0}^{b}\left(v(s)\right)<\frac{\varepsilon}{3}. \tag{2. 3}$$

Пусть $0<\delta_1$ настолько мало, что

$$|u(a)|\,|v(t+h-a)-v(t-a)|<\frac{\varepsilon}{3}, \tag{2. 4}$$

если $h\leqq\delta_1$. Существование δ_1 следует из непрерывности функции $v(t)$ на отрезке $0\leqq t\leqq b-a$.

В неравенстве (2. 2.) третье из четырех слагаемых равно нулю, так как $u(0)=0$.

Наконец, пусть $0<\delta_2$ настолько мало, что

$$\frac{N}{a}\,\omega\left(\frac{a}{N};u\right)\int_0^a |v(t+h-s)-v(t-s)|\,ds=$$

$$=\frac{N}{a}\,\omega\left(\frac{a}{N};u\right)\int_{t-a}^{t} |v(\tau+h)-v(\tau)|\,d\tau<\frac{\varepsilon}{3} \tag{2. 5}$$

если $h\leqq\delta_2$ и $a\leqq t<b$. Существование δ_2 вытекает оттуда, что $v(s)$ имеет ограниченную вариацию на отрезке $[0, b]$, потому что по теореме Жордана

$$v(t)=v_2(t)-v_1(t)$$

где $v_1(t)$ и $v_2(t)$ монотонно возрастают. Таким образом

$$\int_{t-a}^{t} |v(\tau+h)-v(\tau)|\,d\tau \leqq \int_{t-a}^{t} |v_2(\tau+h)-v_2(\tau)|\,d\tau + \int_{t-a}^{t} |v_1(\tau+h)-v_1(\tau)|\,d\tau =$$

$$= \int_{t}^{t+h} v_2(\tau)\,d\tau - \int_{t-a}^{t+h-a} v_2(\tau)\,d\tau + \int_{t}^{t+h} v_1(\tau)\,d\tau - \int_{t-a}^{t+h-a} v_1(\tau)\,d\tau \leqq$$

$$\leqq h[v_2(t+h)-v_2(t-a)] + h[v_1(t+h)-v_1(t-a)] \leqq$$

$$\leqq h[v_2(b)-v_2(0)] + h[v_1(b)-v_1(0)] < \frac{\varepsilon}{3}\,\frac{a}{N}\,\frac{1}{\omega\left(\frac{a}{N};u\right)}$$

если h настолько мало, что $h<\delta_2$. Отсюда и из (2. 3), (2. 4) получается, что

$$|U(t+h)-U(t)| < \frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3},$$

если $h \leqq \delta = \min = (\delta_1, \delta_2)$. Отсюда вытекает, что $U(t)$ непрерывная в точках $a \leqq t < b$. Если $a < t \leqq b$ и $h<0$, то доказательство проходит аналогично. Таким образом, лемма 2. 1 доказана.

Теорема 2. 1. *Если*

(а) $f(t)$ *непрерывна на отрезке* $0 \leqq t \leqq T(<\infty)$, и $f(0)=0$;

(б) $\Phi(t)$ *имеет ограниченную вариацию на отрезке* $[0, T]$ *и непрерывна на отрезке* $[0, \delta]$, *если* $0<\delta<T$.

Тогда существует единственное непрерывное решение уравнения

$$y(t) = f(t) + \int_{0}^{t} y(t-\tau)\,d\Phi(\tau) \tag{2. 6}$$

на отрезке $0 \leqq t \leqq T$.

Доказательство.

Существование решения легко получить при помощи метода последовательных приближений.

Положим $(n=1, 2, \ldots)$

$$y_0(t) = f(t)$$

$$y_n(t) = f(t) + \int_{0}^{t} y_{n-1}(t-\tau)\,d\Phi(\tau). \tag{2. 7}$$

Пусть $h = T/N \leqq \delta$ (N натуральное число) выбран так, что

$$b = \bigvee_{0}^{h}(\Phi) < 1. \tag{2. 8}$$

Существование h вытекает оттуда, что $\Phi(t)$ непрерывная на отрезке $[0, \delta]$.

Очевидно, что все функции $y_n(t)$ являются непрерывными на участке $[0, h]$, потому что $f(t)$ и $\Phi(t)$ непрерывные функции. Покажем, что последовательность функций $y_n(t)$ $(n=1, 2, \ldots)$ равномерно сходится на отрезке $[0, h]$. Действительно, при $n=1, 2, \ldots$

$$|y_{n+1}(t)-y_n(t)| = \left|\int_0^t [y_n(t-\tau)-y_{n-1}(t-\tau)]\,d\Phi(\tau)\right| \leqq$$

(2. 9)

$$\leqq \max_{0\leqq\tau\leqq t} |y_n(t-\tau)-y_{n-1}(t-\tau)| \cdot \bigvee_{\tau=0}^{t}(\Phi(\tau)).$$

Обозначим временно

$$Y_n = \max_{0\leqq t\leqq h} |y_{n+1}(t)-y_n(t)| \qquad (n=1, 2, \ldots).$$

Тогда получается

$$Y_n \leqq b^n Y_0$$

где

$$Y_0 = \max_{0\leqq t\leqq h} |y_1(t)-y_0(t)| = \max_{0\leqq t\leqq h}\left|\int_0^t f(t-\tau)\,d\Phi(\tau)\right| \leqq b\cdot \max_{0\leqq t\leqq h}|f(t)|.$$

Отсюда

$$\sum_{n=0}^{\infty} Y_n \leqq \sum_{n=0}^{\infty} b^{n+1} \max_{0\leqq t\leqq h}|f(t)| < \infty$$

потому что $0<b<1$.

Значит, по известному признаку, ряд $\sum_{k=1}^{\infty}[y_k(t)-y_{k-1}(t)]$ а с ним и последовательность $y_n(t)$ равномерно сходится на $[0, h]$, что и утверждалось.

Положим

$$y(t) = \lim_{n\to\infty} y_n(t) \qquad (0\leqq t\leqq h).$$

Тогда эта функция будет непрерывной при $0\leqq t\leqq h$. В силу (2. 7) для $0\leqq t\leqq h$ будет:

$$\left|y(t)-f(t)-\int_0^t y(t-\tau)\,d\Phi(\tau)\right| \leqq$$

(2. 10)

$$\leqq \left|y(t)-y_{n+1}(t)+\int_0^t y_n(t-\tau)\,d\Phi(\tau)-\int_0^t y(t-\tau)\,d\Phi(\tau)\right| \leqq$$

$$\leqq |y_{n+1}(t)-y(t)| + \max_{0\leqq\tau\leqq t}|y_n(t-\tau)-y(t-\tau)| \bigvee_0^h(\Phi).$$

При фиксированном $t\in[0, h]$ и при $n\to\infty$ оба слагаемых в правой части (2. 10) стремятся к нулю. Значит, левая часть (2. 10), не зависящая от n, равна нулю, т. е. функция $y(t)$ является решением уравнения (2. 6). Итак, существование решения доказано на отрезке $[0, h]$. Доказательство единственности этого

решения проходит следующим образом. Пусть $z(t)$ и $y(t)$ два решения уравнения (2. 6). Положим $Z = \max_{0 \leqq t \leqq h} |z(t)-y(t)|$ и пусть $0<Z$. Тогда

$$|z(t)-y(t)| = \left| \int_0^t [z(t-\tau)-y(t-\tau)]\, d\Phi(\tau) \right| \leqq$$

$$\leqq \max_{0 \leqq \tau \leqq t} |z(t-\tau)-y(t-\tau)| \cdot \bigvee_0^t(\Phi) \leqq Z \cdot b,$$

то есть

$$Z \leqq b \cdot Z.$$

$Z>0$, что противоречит условию $0<b<1$. Итак единственность решения доказана на отрезке $[0, h]$.

Определив решение на отрезке $[0, h]$ мы, чтобы получить решение на отрезке $[h, 2h]$, поступим теперь следующим образом. Положим для $h \leqq t \leqq 2h$

$$y_0(t) = f(t)$$

$$y_{n+1}(t) = f(t) + \int_0^{t-h} y_n(t-\tau)\, d\Phi(\tau) + \int_{t-h}^t y(t-\tau)\, d\Phi(\tau), \tag{2. 11}$$

где $y(t)$ — функция, определенная высше следовательно

$$y_{n+1}(t) = f_1(t) + \int_0^{t-h} y_n(t-\tau)\, d\Phi(\tau)$$

где

$$f_1(t) = f(t) + \int_{t-h}^t y(t-\tau)\, d\Phi(\tau).$$

Это последовательность рекурентных соотношений точно такого же вида, как и рассмотренная высше, то есть в силу леммы 2. 1 функция $f_1(t)$ непрерывная на отрезке $[h, 2h]$. Значим, последовательность $\{y_n\}$ сходится на отрезке $h \leqq t \leqq 2h$ к решению уравнения

$$z(t) = f(t) + \int_0^{t-h} z(t-\tau)\, d\Phi(\tau) + \int_{t-h}^t y(t-\tau)\, d\Phi(\tau).$$

Если мы теперь будем рассматривать $y(t)$ и $z(t)$ так как на отрезке $[0, 2h]$, то мы получим решение уравнения (2. 6) на отрезке $[0, 2h]$. Продолжая таким же образом, мы получим решение на отрезке $[0, 3h]$ и т. д. пока не охватим весь отрезок $[0, T]$.

Если $z(t)$ и $y(t)$ — два решения уравнения (2. 6), тогда мы знаем, что $y(t)$ и $z(t)$ тождественно совпадают на отрезке $[0, h]$. Мы повторим предыдущие рассуждения для отрезка $[h, 2h]$ и т. д. Теорема 2. 1 доказана.

3. Об устойчивости решений

Теоремами устойчивости мы называем теоремы, с которыми исследуем зависимость решения уравнения

$$y(t) = f(t) + \int_0^t y(t-\tau)\,\varphi(\tau)\,d\tau \tag{3.1}$$

от функций f, φ, и решения уравнения

$$y(t) = f(t) + \int_0^t y(t-\tau)\,d\Phi(\tau) \tag{3.2}$$

от функций f и Φ.

Для уравнения (3. 1) мы можем установить следующий результат. Мы будем предполагать, что интеграл в уравнении (3. 1) является интегралом Лебега.

Теорема 3. 1. *Если* $0 < T < \infty$ *и*

(а) $f(t)$, $g(t)$ *непрерывны* $\varphi(t)$, $\psi(t)$ *интегрируемы и* $|\varphi(t)|$, $|\psi(t)| \leqq M (< \infty)$ *на отрезке* $[0, T]$;

(б) $|g(t)-f(t)| < \delta_1$; $\qquad (0 \leqq t \leqq T)$

(в) $$\int_0^T |\psi(t)-\varphi(t)\,dt < \delta_2;$$

тогда между решения $y(t)$ *уравнения* (3. 1) *и между решения* $z(t)$ *уравнения*

$$z(t) = g(t) + \int_0^t z(t-\tau)\,\psi(\tau)\,d\tau \tag{3.3}$$

имеет место неравенство

$$|z(t)-y(t)| \leqq (\delta_1 + \delta_2 e^{MT} \max_{0 \leqq \tau \leqq T} |f(\tau)|)\, e^{MT}$$

$$(0 \leqq t \leqq T).$$

Доказательство. Так как $y(t)$ и $z(t)$ решения уравнений (3. 1) и (3. 3), поэтому мы получим из (б) и (в):

$$|z(t)-y(t)| \leqq |g(t)-f(t)| + \left| \int_0^t z(t-\tau)\,\psi(\tau)\,d\tau - \int_0^t y(t-\tau)\,\varphi(\tau)\,d\tau \right| \leqq$$

$$\leqq \delta_1 + \int_0^t |z(t-\tau)-y(t-\tau)|\,|\psi(\tau)|\,d\tau + \int_0^t |y(t-\tau)|\,|\psi(\tau)-\varphi(\tau)|\,d\tau \leqq \tag{3.4}$$

$$\leqq \delta_1 + \delta_2 \max_{0 \leqq \tau \leqq T} |y(\tau)| + M \int_0^t |z(t-\tau)-y(t-\tau)|\,d\tau.$$

Функции $f(t)$, $g(t)$ непрерывные, таким образом $y(t)$ и $z(t)$ тоже непрерывные функции, поэтому при помощи леммы Беллмана (леммы 2. 2) получим неравенство:

$$|z(t)-y(t)| \leqq \left(\delta_1+\delta_2 \max_{0\leqq\tau\leqq T}|y(\tau)|\right)e^{MT} \qquad (0 \leqq t \leqq T) \tag{3. 5}$$

После этого произведем оценку для $\max_{0\leqq\tau\leqq T}|y(\tau)|$. Из уравнения (3. 1)

$$|y(t)| \leqq |f(t)| + \int_0^t |y(t-\tau)|\,|\varphi(\tau)|\,d\tau \leqq \max_{0\leqq\tau\leqq T}|f(\tau)| + M\cdot\int_0^t |y(\tau)|\,d\tau,$$

то есть из леммы 2. 2.

$$y(t) \leqq e^{MT} \max_{0\leqq\tau\leqq T}|f(\tau)| \qquad (0 \leqq t \leqq T).$$

Если последнее неравенство впишем в (3. 5), тогда

$$|z(t)-y(t)| \leqq \left(\delta_1+\delta_2 e^{MT} \max_{0\leqq\tau\leqq T}|f(\tau)|\right)e^{MT}$$

Теорема доказана.

В дальнейшем мы получим оценку для $\int_0^T [z(\tau)-y(\tau)]^2 dt$. Сначала мы установим результат, который дает оценку для $\int_0^T y^2(t)dt$.

Теорема 3. 2. *Если*

$$\Phi = \int_0^T \varphi^2(\tau)\,d\tau < \infty \tag{3. 6}$$ (а)

$$|f(t)| \leqq M(<\infty) \qquad (0 \leqq t \leqq T) \tag{3. 7}$$ (б)

и $y(t)$ квадратично интегрируемое решение уравнения (3. 1), *тогда имеет место неравенство*

$$\int_0^T y^2(\tau)\,d\tau \leqq M^2\cdot K(\Phi), \tag{3. 8}$$

где

$$K(\Phi) = \frac{4}{\Phi}\left[e^{\frac{\Phi}{2}T}\left(1+\frac{\sqrt{\Phi T}}{2}\right)-1\right]^2. \tag{3. 9}$$

Доказательство. Из уравнения (3. 1) и неравенства (3. 7) вытекает

$$y^2(t) = f^2(t)+2\cdot f(t)\int_0^t y(t-\tau)\,\varphi(\tau)\,d\tau+\left(\int_0^t y(t-\tau)\,\varphi(\tau)\,d\tau\right)^2 \leqq$$

$$\leqq M^2+2\cdot M\cdot\left|\int_0^t y(t-\tau)\,\varphi(\tau)\,d\tau\right|+\left(\int_0^t y(t-\tau)\,\varphi(\tau)\,d\tau\right)^2,$$

то есть по неравенству Буняковского—Шварца

$$y^2(t) \leqq M^2 + 2 \cdot M \sqrt{\int_0^t y^2(\tau)\,d\tau} \cdot \sqrt{\int_0^t \varphi^2(\tau)\,d\tau} + \int_0^t y^2(\tau)\,d\tau \cdot \int_0^t \varphi^2(\tau)\,d\tau.$$

При помощи интегрирования обеих частей последнего неравенства получается:

$$\int_0^t y^2(\tau)\,d\tau \leqq M^2 t + \int_0^t \left[2M\sqrt{\Phi} \sqrt{\int_0^\tau y^2(s)\,ds} + \Phi \int_0^\tau y^2(s)\,ds \right] d\tau.$$

$$(0 \leqq t \leqq T)$$

Положим

$$Y(t) = \int_0^t y^2(\tau)\,d\tau \qquad (0 \leqq t \leqq T).$$

Тогда $Y(t)$ непрерывная и

$$Y(t) \leqq M^2 T + \int_0^t \left[2M \cdot \sqrt{\Phi}\sqrt{Y(\tau)} + \Phi Y(\tau) \right] d\tau.$$

Из леммы 1. 1:

$$Y(t) \leqq G^{-1}\left(G(M^2 T) + t\right) \tag{3. 10}$$

$$0 \leqq t \leqq T$$

где

$$G(u) = \int_{u_0}^{u} \frac{1}{2M\sqrt{\Phi}\sqrt{t} + \Phi t}\,dt \qquad (u \geqq 0,\ u_0 > 0) \tag{3. 11}$$

то есть

$$G(u) = \frac{2}{\Phi} \ln \frac{2M\sqrt{\Phi} + \Phi\sqrt{u}}{2M\sqrt{\Phi} + \Phi\sqrt{u_0}} \tag{3. 12}$$

Если теперь $u_0 = 0$, тогда

$$G(u) = \frac{2}{\Phi} \ln \left(1 + \frac{\sqrt{\Phi}}{2M}\sqrt{u} \right) \qquad (u \geqq 0) \tag{3. 13}$$

откуда функция $G^{-1}(u)$ обратная функции $G(u)$:

$$G^{-1}(u) = \frac{4M^2}{\Phi}\left(e^{\frac{\Phi}{2}u} - 1\right)^2. \tag{3. 14}$$

Из соотношении (3. 13) и (3. 14) и неравенства (3. 10) получается

$$Y(t) \leqq \frac{4M^2}{\Phi} \left\{ e^{\frac{\Phi}{2} \ln \left[\left(1 + \frac{\sqrt{\Phi}}{2M}\sqrt{M^2 T} \right) + t \right]} - 1 \right\}^2$$

из которого очевидно вытекает неравенство (3. 9).

Теорема 3. 2 доказана.

Используя результат, мы установим следующее утверждение.

Теорема 3. 3. *Предположим, что функции* $f(t)$, $g(t)$, $\varphi(t)$ *и* $\psi(t)$ *определены на отрезке* $[0, T]$ *и*

(а) $|f(t)| \leqq M$ *на отрезке* $[0, T]$ (3. 15)

$$\Phi = \int_0^T \varphi^2(\tau)\, d\tau < \infty$$

(3. 16) $$\Psi = \int_0^T \psi^2(\tau)\, d\tau < \infty$$

(в) *при* $0 < \delta < \infty$ *будет*:

(3. 17) $$|g(t) - f(t)| < \delta$$

(3. 18) $$\int_0^T [\psi(\tau) - \varphi(\tau)]^2\, d\tau < \delta$$

Если при предыдущих условиях $y(t)$ *и* $z(t)$ *квадратично интегрируемые решения уравнений*

(3. 19) $$y(t) = f(t) + \int_0^t y(t-\tau)\, \varphi(\tau)\, d\tau$$

и

(3. 20) $$z(t) = g(t) + \int_0^t z(t-\tau)\, \psi(\tau)\, d\tau,$$

тогда имеет место неравенство

(3. 21) $$\int_0^T [z(t) - y(t)]^2\, dt \leqq \left(\delta + M\sqrt{\delta}\,\sqrt{K(\Phi)}\right)^2 K(\Psi)$$

где

(3. 22) $$K(u) = \frac{4}{u}\left[e^{\frac{u}{2}T}\left(1 + \frac{\sqrt{uT}}{2}\right) - 1\right]^2 \quad (u > 0).$$

Доказательство. Из уравнений (3. 19) и (3. 20) вытекает

(3. 23) $$z(t) - y(t) = f_1(t) + \int_0^t [z(t-\tau) - y(t-\tau)]\, \psi(\tau)\, d\tau,$$

где

(3. 24) $$f_1(t) = g(t) - f(t) + \int_0^t y(t-\tau)\, [\psi(\tau) - \varphi(\tau)]\, d\tau.$$

Покажем, что $f_1(t)$ ограниченная функция. Из (3. 23) и условия (в) при помощи неравенства Буняковского—Шварца получим:

$$|f_1(t)| \leqq |g(t)-f(t)|+\sqrt{\int_0^t y^2(\tau)\,d\tau \cdot \int_0^t [\psi(\tau)-\varphi(\tau)]^2\,d\tau} \leqq$$

(3. 25)

$$\leqq \delta+\sqrt{\delta}\cdot\sqrt{\int_0^T y^2(\tau)\,d\tau} \qquad (0 \leqq t \leqq T).$$

Далее по теореме 3. 2 и из (3. 25) следует

$$\int_0^T [z(t)-y(t)]^2\,dt \leqq M_1^2 K(\Psi)$$

где $K(\Psi)$ выписано в (3. 22) и

$$M_1 = \delta+\sqrt{\delta}\sqrt{\int_0^T y^2(\tau)\,d\tau}.$$

Снова применяя теорему 3. 2, получается, что

$$\int_0^T y^2(\tau)\,d\tau \leqq M^2 K(\Phi)$$

потому что $y(t)$ квадратично интегрируемое решение уравнения (3. 19), и

$$\int_0^T [z(\tau)-y(\tau)]^2\,d\tau \leqq \left(\delta+\sqrt{\delta}\,M\sqrt{K(\Phi)}\right)^2 K(\Psi).$$

Теорема 3. 3 доказана.

Докажем следующую теорему, в которой мы исследуем зависимость решения уравнения

$$y(t) = f(t)+\int_0^t y(t-\tau)\,d\Phi(\tau) \tag{3. 27}$$

от функций f и Φ. Тут видна возможность совместного применения лемм Мышкиса и Беллмана.

Теорема 3. 4. *Допустим, что*
(а) $f(t)$, $g(t)$ *непрерывны на отрезке* $[0, T]$;
(б) $\Phi(t)$ *имеет ограниченную вариацию на отрезке* $[0, T]$,
и $\Psi(t)$ *удовлетворяет условию Липшица*

$$|\Psi(t_2)-\Psi(t_1)| \leqq K|t_2-t_1|$$

для любых точек $0 \leqq t_1, t_2 \leqq T$;
(в)

$$|\Psi(0)| < \frac{1}{2}.$$

Тогда для любого $\varepsilon > 0$ *найдется такое* $\delta > 0$, *что если*

(г) $$|g(t) - f(t)| \leqq \delta; \qquad (0 \leqq t \leqq T)$$

(д) $$|\Phi(t) - \Psi(t)| \leqq \delta; \qquad (0 \leqq t \leqq T)$$

и $y(t)$ *решение уравнения* (3. 27), *а* $z(t)$ *решение уравнения*

$$z(t) = g(t) + \int_0^t z(t-\tau)\, d\Psi(\tau) \tag{3.28}$$

где решения непрерывны и имеют ограниченную вариацию, то будет:

$$|z(t) - y(t)| < \varepsilon.$$

Доказательство. Пусть t фиксировано $0 < t \leqq T$. Тогда из (3. 27) и (3. 28) при помощи условия (г) будет:

$$\begin{aligned} |z(t) - y(t)| \leqq \delta + \left| \int_0^t y(t-\tau)\, d\Psi(\tau) - \int_0^t y(t-\tau)\, d\Phi(\tau) \right| + \\ + \left| \int_0^t [z(t-\tau) - y(t-\tau)]\, d\Psi(\tau) \right|. \end{aligned} \tag{3.29}$$

Теперь мы даем оценку для интегралов в правой части (3. 29). Исследуем второй интеграл в правой части. Функция $y(t)$ непрерывная, таким образом из леммы 2. 3 получится, что для любого фиксированного натурального числа будет:

$$\left| \int_0^t y(t-\tau)\, d\Psi(\tau) - \int_0^t y(t-\tau)\, d\Phi(\tau) \right| \leqq$$

$$\leqq \omega\left(\frac{t}{n}; y(t-\tau)\right) \left[\bigvee_0^t (\Psi) + \bigvee_0^t (\Phi) \right] + |y(0)|\, |\Psi(t) - \Phi(t)| +$$

$$+ |y(t)| \cdot |\Psi(0) - \Phi(0)| + \frac{n}{t}\, \omega\left(\frac{t}{n}; y(t-\tau)\right) \int_0^t |\Psi(\tau) - \Phi(\tau)|\, d\tau.$$

Принимаем во внимание условия (д) и $y(0) = f(0)$. Тогда

$$\left| \int_0^t y(t-\tau)\, d\Psi(\tau) - \int_0^t y(t-\tau)\, d\Phi(\tau) \right| \leqq$$

$$\begin{aligned} \leqq \omega\left(\frac{t}{n}; y(t-\tau)\right) \left[\bigvee_0^t (\Psi) + \bigvee_0^t (\Phi) \right] + |f(0)|\, \delta + \delta \cdot \max_{0 \leqq \tau \leqq T} |y(t)| + \\ + \delta \cdot n \cdot \omega\left(\frac{t}{n}; y(t-\tau)\right) \leqq \omega\left(\frac{t}{n}; y(t-\tau)\right) \left[\bigvee_0^T (\Psi) + \bigvee_0^T (\Phi) \right] + \\ + \delta \left[|f(0)| + \max_{0 \leqq \tau \leqq T} |y(t)| + n \cdot \omega\left(\frac{t}{n}; y(t-\tau)\right) \right]. \end{aligned} \tag{3.30}$$

Последний член в правой части (3. 29):

$$\left|\int_0^t [z(t-\tau)-y(t-\tau)]\,d\Psi(\tau)\right| \leqq |z(0)-y(0)||\Psi(t)|+|z(t)-y(t)||\Psi(0)|+$$

$$+\left|\int_0^t \Psi(t)\,d[z(\tau)-y(\tau)]\right|.$$

По условию $\Psi(t)$ непрерывная, $y(t)$ и $z(t)$ имеют ограниченную вариацию, поэтому из леммы 2. 3 получится, что для любого фиксированного натурального числа m будет

$$\left|\int_0^t [z(t-\tau)-y(t-\tau)]\,d\Psi(\tau)\right| \leqq |z(0)-y(0)|\,|\Psi(t)|+|z(t)-y(t)|\,|\Psi(0)|+$$

$$+\omega\left(\frac{t}{m};\Psi\right)\left[\bigvee_{\tau=0}^{t}(y(t-\tau))+\bigvee_{0}^{t}(z(t-\tau))\right]+|z(0)-y(0)|\cdot|\Psi(t)|+$$

$$+|z(t)-y(t)|\cdot|\Psi(0)|+\frac{m}{t}\,\omega\left(\frac{t}{m};\Psi\right)\cdot\int_0^t|z(\tau)-y(\tau)|\,d\tau.$$

Отсюда и из (б) и (г) получится

$$\left|\int_0^t [z(t-\tau)-y(t-\tau)]\,d\Psi(\tau)\right| \leqq 2\cdot\delta\max_{0\leqq\tau\leqq T}|\Psi(t)|+2\cdot|\Psi(0)|\,|z(t)-y(t)|+$$

$$+\omega\left(\frac{t}{m};\Psi\right)\left[\bigvee_{\tau=0}^{t}(y(t-\tau))+\bigvee_{\tau=0}^{t}(z(t-\tau))\right]+K\int_0^t|z(\tau)-y(\tau)|\,d\tau$$

где K определена в (б). Если $m\to\infty$, то

$$\left|\int_0^t [z(t-\tau)-y(t-\tau)]\,d\Psi(\tau)\right| \leqq 2\cdot\delta\max_{0\leqq\tau\leqq T}|\Psi(\tau)|+2\,|\Psi(0)|\,|z(t)-y(t)|+$$

$$+K\int_0^t|z(\tau)-y(\tau)|\,d\tau.$$

Подставляя в правую часть (3. 29), последнее неравенство и (3. 30) получим

$$|z(t)-y(t)| \leqq \delta\left\{1+|f(0)|+\max_{0\leqq t\leqq T}|y(t)|+2\cdot\max_{0\leqq t\leqq T}|\Psi(t)|+n\cdot\omega\left(\frac{t}{n};y(t-\tau)\right)\right\}+$$

(3. 31)

$$+\omega\left(\frac{t}{n};y(t-\tau)\right)\left[\bigvee_{0}^{T}(\Psi)+\bigvee_{0}^{T}(\Phi)\right]+2\,|\Psi(0)|\,|z(t)-y(t)|+K\cdot\int_0^t|z(\tau)-y(\tau)|\,d\tau$$

$$(0<t\leqq T).$$

Очевидно, что это неравенство выполняется для $t=0$. Если ввести обозначения

$$F(n)=\frac{1}{1-2|\Psi(0)|}\left\{1+|f(0)|+\max_{0\leqq t\leqq T}|y(t)|+2\cdot\max_{0\leqq t\leqq T}|\Psi(t)|+n\max_{0\leqq t\leqq T}\omega\left(\frac{t}{n};y\right)\right\}$$

и

$$M=\frac{1}{1-2|\Psi(0)|}\left[\bigvee_0^T(\Phi)+\bigvee_0^T(\Psi)\right],\quad L=\frac{K}{1-2|\Phi(0)|}$$

то из оценки (3. 31)

$$|z(t)-y(t)|\leqq\delta\cdot F(n)+M\cdot\max_{0\leqq t\leqq T}\omega\left(\frac{t}{n};y(t-\tau)\right)+L\int_0^t|z(\tau)-y(\tau)|\,d\tau$$

$$(0\leqq t\leqq T)$$

Отсюда в силу леммы Беллмана

$$|z(t)-y(t)|\leqq\left\{\delta\cdot F(n)+M\cdot\max_{0\leqq t\leqq T}\omega\left(\frac{t}{n};y(t-\tau)\right)\right\}e^{Lt} \tag{3.32}$$

$$(0\leqq t\leqq T)$$

Тогда можно выбрать N столь большим, что

$$M\cdot\max_{0\leqq t\leqq T}\omega\left(\frac{t}{N};y(t-\tau)\right)e^{LT}<\frac{\varepsilon}{2}. \tag{3.33}$$

После фиксирования N за счет уменьшения δ можно сделать меньше $\varepsilon/2$ остальный член в правой части неравенства (3. 32). Отсюда и следует утверждение.

Замечание. В том частном случае, когда

$$\Phi(t)=\int_0^t\varphi(\tau)\,d\tau$$

и

$$\Psi(t)=\int_0^t\psi(\tau)\,d\tau$$

то условие (д) предыдущей теоремы будет

$$\left|\int_0^t\varphi(\tau)\,d\tau-\int_0^t\psi(\tau)\,d\tau\right|\leqq\delta$$

то есть

$$\left|\int_0^t\varphi(\tau)-\psi(\tau)]\,d\tau\right|\leqq\delta.$$

БИБЛИОГРАФИЯ

[1] BELLMAN, R.: The stability of solutions of linear differential equations, *Duke Math. Journal*, **10** (1943), 643—647.
[2] BELLMAN, R., COOKE, K. L.: *Differential-difference equations*, New York (1963), p. 238—257.
[3] BIHARI, I.: A generalization of a lemma of Bellman and its application to uniqueness problems of differential equations, *Acta Math. Acad. Sci. Hungar.* **7** (1956), 81—94.
[4] А. Д. Мышкис: *Линейные дифференциальные уравнения с запаздывающим аргументом*, Москва (1951), 233—235.

Szeged, SZOTE Központi Kutató Laboratorium

(Поступила 25-ого июня 1969. г.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 369—394.

GENERALIZATION OF McMILLAN'S THEOREM TO RANDOM SET FUNCTIONS

by

J. FRITZ

In the mathematical theory of statistical thermodynamics the physical quantities (extensive parameters) are random variables defined on the microstates, although their values are well determined for a real macroscopic system, at the least to a high degree of accuracy. Of course, this does not mean any contradiction, namely we know from fluctuation theory well, that in case of a macroscopic system the value of each physical quantity — expressed in macroscopic units — will be close to its mean value for the greatest part of the microstates. More precisely, the probability of the set of these "ordinary" microstates will be arbitrarily close to one if the system is large enough. As the extensive quantities are additive, this assertion follows also from the ergodic theorem under very general conditions.

A quite similar statement is true for the entropy, although — in general — it is not additive. In view of BOLTZMANN's principle, we interpret the entropy of a microstate as the negative logarithm of its probability (probability density). Then, taking into account the concrete form of the equilibrium distribution (canonical or grand canonical distibution), we see that the entropy of each ordinary microstate is close to the thermodynamical entropy of the system, which is defined as the expectation of the above "microscopic" entropy. The importance of this assertion consists in associating BOLTZMANN's principle with the entropy-maximum principle of T. E. JAYNES. We remark that a similar argument is developed in paper [7] of I. VINCZE. Here we will give a general formulation of the problem.

In spite of the obvious relation of this equipartition property to the McMILLAN theorem (ergodic theorem of information theory), it does not follow from this theorem, namely the McMILLAN theorem — in its usual formulation — can not be applied to a three-dimensional system.

The main goal of this paper is to prove a more-dimensional version of the McMILLAN theorem; moreover the final section treates its physical consequences. The first and second sections deal with the basic concepts and tools needed in our investigations. The most important among them are the r-dimensional ergodic theorem and a submartingale convergence theorem.

Finally, I should like to express may thanks to IMRE CSISZÁR for his valuable advices.

1. The r-dimensional ergodic theorem

A one-to-one mapping T of Ω onto itself will be called an automorphism of the probability space $(\Omega, \mathcal{A}, \mathsf{P})$ if both T and T^{-1} are measurable with respect to $\mathcal{A}$ and T preserves the probability measure P. The usual form of the ergodic

theorem is associated with a cyclic group of automorphisms. Here we shall investigate the more general situation, when a finitely generated Abelian group $\mathcal{G}$ of automorphisms of the probability space $(\Omega, \mathcal{A}, \mathsf{P})$ is given. If the system $\{T_1, T_2, \ldots, T_r\}$ generates $\mathcal{G}$, then the elements of $\mathcal{G}$ have the — not necessarily unique — form:

$$T_u = \prod_{i=1}^{r} T_i^{n_i}, \tag{1.1}$$

where $u = (n_1, n_2, \ldots n_r)$ is an arbitrary r-tuple of integers. Let $\{e_1, e_2, \ldots, e_r\}$ be a — not necessarily orthogonal — base of the r-dimensional Euclidean space R, then the elements of R have the unique form

$$x = (x_1, x_2, \ldots x_r) = \sum_{i=1}^{r} x_i e_i, \tag{1.2}$$

where the co-ordinates $x_1, x_2, \ldots x_r$ are real numbers. The elements of R with integer co-ordinates form an additive subgroup R_0 of R. In view of (1.1) and (1.2) to each $u = (n_1, n_2, \ldots n_r) \in R_0$ there corresponds an element T_u of $\mathcal{G}$ such that

$$T_{u+v} = T_u T_v \quad \text{if} \quad u, v \in R_0, \tag{1.3}$$

that is the mapping $u \to T_u$ is a homomorphism of the additive group R_0 onto $\mathcal{G}$, defined by

$$T_{e_i} = T_i, \qquad i = 1, 2, \ldots r. \tag{1.4}$$

On the other hand, each $u \in R_0$ defines a translation of R, where the image of the r-dimensional Borel set $G \in \mathcal{R}$ will be denoted by

$$G + u = \{x;\ x \in R, x - u \in G\}. \tag{1.5}$$

This transformation preserves the r-dimensional Lebesgue measure λ, and the length $\|x\|$ of an r-dimensional vector $x \in R$. We introduce some further notations: The class of bounded open convex subsets of R will be denoted by $\mathcal{B}$. For a $G \in \mathcal{B}$ we set

$$\varrho(G) = \sup \{\varrho;\ S(x, \varrho) \subset G, x \in R\} \tag{1.6}$$

where $S(x, \varrho)$ denotes the r-dimensional sphere with centre x and radius ϱ. We shall investigate convergence of certain set functions X_G if $\varrho(G) \to +\infty$. For example, if $X(\omega : G)$ is a random set function (i.e. $X(\omega : G)$ is measurable for each $G \in \mathcal{R}$), then the $L_1(\mathsf{P})$-convergence relation

$$X(\omega : G) \xrightarrow{1} Y(\omega) \quad \text{if} \quad G \in \mathcal{B}, \quad \varrho(G) \to +\infty \tag{1.7}$$

means that for every $\varepsilon > 0$ there exists a ϱ_ε such that

$$\mathsf{E}\,|X(\omega : G) - Y(\omega)| < \varepsilon \quad \text{if} \quad G \in \mathcal{B} \quad \text{and} \quad \varrho(G) \geqq \varrho_\varepsilon. \tag{1.8}$$

Now we formulate the r-dimensional ergodic theorem. We prove only $L_1(\mathsf{P})$-convergence on the base of BIRKHOFF's theorem.

THEOREM 1. *Let $X(\omega)\in L_1(\mathsf{P})$ be a random variable on the probability space $(\Omega, \mathscr{A}, \mathsf{P})$ having the above group $\mathscr{G}=\{T_u;\ u\in R_0\}$ of automorphisms. Then*

$$\frac{1}{\lambda(G)}\sum_{u\in G\cap R_0} X(T_u\omega)\xrightarrow{1} Y(\omega) \quad \text{if} \quad G\in\mathscr{B},\quad \varrho(G)\to+\infty \tag{1.9}$$

where $Y\in L_1(\mathsf{P})$ is invariant under each automorphism $T_u\in\mathscr{G}$.

PROOF: First we prove the validity of (1. 9) for the following sequence of bounded convex sets:

$$F_n=\left\{x;\ x=(x_1,\ldots,x_r)\in R,\ -\frac{2n+1}{2}\leqq x_i<\frac{2n+1}{2}\quad \text{for}\quad i=1,\ldots r\right\}. \tag{1.10}$$

Let us introduce the notation:

$$S_n^{(i)}Z(\omega)=\frac{1}{2n+1}\sum_{k=-n}^{+n} Z(T_i^k\omega) \tag{1.11}$$

for the transformation $T_i=T_{e_i}$, $i=1, 2, \ldots, r$. Using the ergodic theorem of BIRKHOFF for the automorphism T_1, we obtain that

$$S_n^{(1)}X\xrightarrow{1} Y_1, \tag{1.12}$$

where $Y_1\in L_1(\mathsf{P})$ is invariant under T_1. Therefore for each $\varepsilon>0$ there exists an n_1 such that

$$\mathsf{E}\,|S_n^{(1)}X-Y_1|<\varepsilon \quad \text{if} \quad n\geqq n_1. \tag{1.13}$$

Observe that each $T_u\in\mathscr{G}$ preserves a relation of type $\mathsf{E}|Z-Z'|<\varepsilon$, that is then also $\mathsf{E}|Z(T_u\omega)-Z'(T_u\omega)|<\varepsilon$ holds as T_u preserves P. Consequently, (1. 13) implies that

$$\mathsf{E}\,|S_n^{(1)}X(T_2^k\omega)-Y_1(T_2^k\omega)|<\varepsilon \tag{1.14}$$

holds for each integer k if $n\geqq n_1$. On the other hand.

$$S_n^{(2)}Y_1(\omega)\xrightarrow{1} Y_2(\omega), \tag{1.15}$$

where $Y_2\in L_1(\mathsf{P})$ is invariant under both T_1 and T_2, further

$$|S_n^{(2)}S_n^{(1)}X(\omega)-Y_2(\omega)|\leqq|S_n^{(2)}Y_1(\omega)-Y_2(\omega)|+\frac{1}{2n+1}\sum_{k=-n}^{n}|S_n^{(1)}X(T_2^k\omega)-Y_1(T_2^k\omega)|, \tag{1.16}$$

so that (1. 14) and (1. 15) imply the existence of such an $n_2\geqq n_1$ that

$$\mathsf{E}\,|S_n^{(2)}S_n^{(1)}X-Y_2|<2\varepsilon \quad \text{if} \quad n\geqq n_2. \tag{1.17}$$

As T_3 preserves the inequality (1. 17), the above procedure can be continued without any change of the method. Having carried it $r-1$-times out, we obtain an invariant (under $\mathscr{G}$) function $Y_r\in L_1(\mathsf{P})$, defined by

$$S_n^{(r)}Y_{r-1}\xrightarrow{1} Y_r. \tag{1.18}$$

Further, if we choose the integers $n_r \geqq n_{r-1} \geqq \cdots \geqq n_1$ such that

$$\mathsf{E}\,|S_n^{(r)} Y_{r-1} - Y_r| < \varepsilon \quad \text{if} \quad n \geqq n_r, \tag{1.19}$$

then it follows by induction that

$$\mathsf{E}\,|S_n^{(r)} S_n^{(r-1)} \dots S_n^{(1)} X(\omega) - Y_r(\omega)| < r\varepsilon \quad \text{if} \quad n \geqq n_r, \tag{1.20}$$

in the same way as (1. 17) has been proved.

With the notations

$$\hat{X}(\omega : G) = \sum_{u \in G \cap R_0} X(T_u \omega) \tag{1.21}$$

and

$$Y(\omega) = \frac{1}{\lambda(F_0)} Y_r(\omega) \tag{1.22}$$

the inequality (1. 20) implies the convergence

$$\frac{1}{\lambda(F_n)} X(\omega : F_n) \xrightarrow{1} Y(\omega), \tag{1.23}$$

as

$$\lambda(F_n) = (2n+1)^r \lambda(F_0).$$

Observe that the sets $F_m(v) = F_m + (2m+1)v$ form a partition of R into disjoint sets if v ranges over R_0 while m is fixed. Taking into account the invariance of the relation (1. 20) and Y we obtain that

$$\frac{1}{\lambda(F_m(v))} \hat{X}(\omega : F_m(v)) \xrightarrow{1} Y(\omega) \tag{1.24}$$

uniformly in $v \in R_0$, which proves (1. 9) not only for the sequence F_n, but implies the statement of the theorem by the following simple elementary lemma.

LEMMA 1. *Let $G \in \mathscr{B}$ be an r-dimensional bounded convex set and set*

$$\bar{G}^m = \cup \{F_m(v);\, F_m(v) \cap G \neq \emptyset,\, v \in R_0\}, \tag{1.25}$$

$$\underline{G}^m = \cup \{F_m(v);\, F_m(v) \subset G,\, v \in R_0\}. \tag{1.26}$$

If

$$\varrho(G) > d_m = \sup \{\|x - x'\|;\; x, x' \in F_m\}, \tag{1.27}$$

then we have the inequality

$$c^{-r} \lambda(\bar{G}^m) \leqq \lambda(G) \leqq c^r \lambda(\underline{G}^m), \quad \textit{where} \tag{1.28}$$

$$c = \frac{\varrho(G)}{\varrho(G) - d_m}. \tag{1.29}$$

Thus $\varrho(G) \to +\infty$ implies

$$\lim \frac{\lambda(\underline{G}^m)}{\lambda(G)} = 1 \textit{ and } \lim \frac{\lambda(\bar{G}^m - \underline{G}^m)}{\lambda(G)} = 0 \textit{ for each } m. \tag{1.30}$$

PROOF of the lemma: On account of assumption (1. 27) we see that the convex set G includes an r-dimensional sphere $S(y, a)$ for each a with $\varrho(G) > a > d_m$. To prove the lemma, it is sufficient to show that the transformation of similitude S with centre y and ratio $c' = \dfrac{a}{a-d_m}$; i.e.

$$Sx = c'(x-y)+y \quad \text{if} \quad x \in R, \tag{1.31}$$

is such that

$$S\underline{G}^m \supset G \quad \text{and} \quad SG \supset \overline{G}^m, \tag{1.32}$$

namely the ratio of the corresponding volumes is c'^r, and $\lim c' = c$ if $a \to \varrho(G)$.

Let us choose the points x and x' from any $F_m(v) \subset \overline{G}^m - \underline{G}^m$ such that $x \in F_m(v) \cap G$ and $x' \in F_m(v) - G$. Denoting by z and by z' the only point of the boundary of the convex set G on the half-lines $\overrightarrow{yx}$ and $\overrightarrow{yx'}$ resp., we have to show that

$$\|z-y\| \leqq c'\|x-y\| \quad \text{and} \quad \|x'-y\| \leqq c'\|z'-y\|. \tag{1.33}$$

We may assume that the half-line $\overrightarrow{yx}$ does not pass through x', namely in the opposite case $z = z'$ and (1. 33) follows from $\|z-y\| \geqq a$ and from $\|x-x'\| \leqq d_m$. Then the points y, x, x' determine a plane which contains also the point $y' = y + a\dfrac{x'-x}{\|x'-x\|} \in G$, and the straight line passing through x and y does not

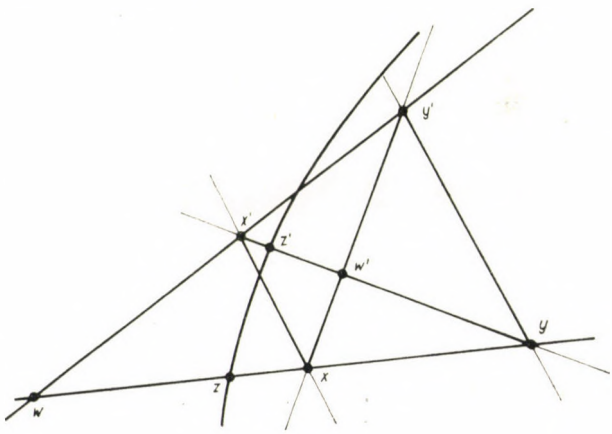


Fig. 1

separate the points x' and y'. Let us denote by w and by w' the points of intersection of the straight lines $\overline{xy}$ and $\overline{x'y'}$, and of $\overline{xy}$ and $\overline{x'y}$ respectively. (See the Fig.)

In view of their definitions, $x \in [z, y]$ and $z' \in [x', y]$, where $[t, t']$ denotes the section with endpoints t and t'. Further, as $\|x-x'\| \leqq d_m < \|y-y'\| = a$; $x \in [w, y]$,

$x' \in [w, y']$ and $w' \in [x', y]$, $w' \in [x, y']$. Since $x \in G$, $y' \in G$ and $x' \notin G$, the convexity of G implies that $w' \in G$ and $w \notin G$. Thus $z \in [w, x]$ and $z' \in [w', x']$, so that

$$\frac{\|z-y\|}{\|x-y\|} \leqq \frac{\|w-y\|}{\|x-y\|} \quad \text{and} \quad \frac{\|x'-y\|}{\|z'-y\|} \leqq \frac{\|x'-y\|}{\|w'-y\|}. \tag{1.34}$$

On the other hand, the similarity of the triangles $wxx'\Delta$ and $wyy'\Delta$ and of $w'xx'\Delta$ and $w'y'y\Delta$ implies that

$$\frac{\|w-y\|}{\|x-y\|} = \frac{\|y-y'\|}{\|y-y'\| - \|x-x'\|} \leqq \frac{a}{a-d_m}, \tag{1.35}$$

further

$$\frac{\|x'-y\|}{\|w'-y\|} = \frac{\|y'-y\| + \|x'-x\|}{\|y'-y\|} \leqq \frac{a+d_m}{a} \leqq \frac{a}{a-d_m}, \tag{1.36}$$

from which (1. 33) follows; thus the proof of the lemma is complete.

To prove the statement of the theorem, we have to show that for each $\varepsilon > 0$ there exists a ϱ_ε such that

$$\mathsf{E}\left|\frac{1}{\lambda(G)}\hat{X}(\omega:G) - Y(\omega)\right| < \varepsilon \quad \text{if} \quad G \in \mathscr{B} \quad \text{and} \quad \varrho(G) \geqq \varrho_\varepsilon. \tag{1.37}$$

On account of (1. 24) we have an m_ε such that

$$\mathsf{E}\left|\frac{1}{\lambda(F_m(v))} X(\omega:F_m(v)) - Y(\omega)\right| < \frac{\varepsilon}{2} \tag{1.38}$$

holds for each $v \in R_0$ if $m \geqq m_\varepsilon$. Let now $m \geqq m_\varepsilon$ be fixed. As the random set-function $\hat{X}(\omega:G)$, defined by (1. 21) is additive, we obtain the following inequalities:

$$\begin{aligned} \mathsf{E}\left|\frac{1}{\lambda(G)}\hat{X}(\omega:G) - Y(\omega)\right| &= \frac{1}{\lambda(G)}\mathsf{E}\,|\hat{X}(\omega:G) - \lambda(G)\,Y(\omega)| \leqq \\ \leqq \frac{1}{\lambda(G)}\mathsf{E}\,|\hat{X}(\omega:\underline{G}^m) - \lambda(\underline{G}^m)\,Y(\omega)| &+ \frac{1}{\lambda(G)}\lambda(G-\underline{G}^m)\,\mathsf{E}\,|Y| + \\ &+ \frac{1}{\lambda(G)}\mathsf{E}\,|\hat{X}(\omega:G-\underline{G}^m)| \end{aligned} \tag{1.39}$$

Since the number of points of R_0 in $G - \underline{G}^m \subset \bar{G}^m - \underline{G}^m$ is less than $\dfrac{\lambda(\bar{G}^m - \underline{G}^m)}{\lambda(F_0)}$, we have

$$\mathsf{E}\,|\hat{X}(\omega:G-\underline{G}^m)| \leqq \frac{\lambda(\bar{G}^m - \underline{G}^m)}{\lambda(F_0)}\mathsf{E}\,|X|. \tag{1.40}$$

Further, as $\underline{G}^m = \cup\{F_m(v);\ F_m(v) \subset G\}$, it follows that

$$\begin{aligned} \mathsf{E}\,|\hat{X}(\omega:\underline{G}^m) - \lambda(\underline{G}^m)\,Y(\omega)| &\leqq \sum_{F_m(v)\subset G} \mathsf{E}\,|X(\omega:F_m(v)) - \lambda(F_m(v))\,Y(\omega)| \leqq \\ \leqq \sum_{F_m(v)\subset G} \lambda(F_m(v))\,\mathsf{E}\left|\frac{1}{\lambda(F_m(v))} X(\omega:F_m(v)) - Y(\omega)\right| &\leqq \lambda(\underline{G}^m)\frac{\varepsilon}{2}. \end{aligned} \tag{1.41}$$

The estimates (1. 39), (1. 40) and (1. 41) yield

$$(1.42)\quad \mathsf{E}\left|\frac{1}{\lambda(G)}\hat{X}(\omega:G)-Y(\omega)\right| \leqq \frac{\varepsilon}{2}+\frac{\lambda(\overline{G}^m-\underline{G}^m)}{\lambda(G)}\left(\mathsf{E}|Y|+\frac{1}{\lambda(F_0)}\mathsf{E}|X|\right),$$

which proves (1. 37) by Lemma 1. The proof of Theorem 1. is complete.

REMARK: If the group $\mathcal{G}$ of automorphisms of $(\Omega, \mathcal{A}, \mathsf{P})$ is ergodic, that is the probability of each (under $\mathcal{G}$) invariant event is either 0 or 1, then $Y(\omega)=\dfrac{E(X)}{\lambda(F_0)}$ almost surely.

It is easy to observe that Theorem 1. can be extended to an important class of random set-functions.

DEFINITION 1. The random set-function $X(\omega:G)$, $G\in\mathcal{R}$ is additive if $G_1\cap G_2=\emptyset$ implies that

$$(1.43)\qquad X(\omega:G_1\cup G_2) = X(\omega:G_1)+X(\omega:G_2) \quad \text{a. s.}$$

DEFINITION 2. The random set-function $X(\omega:G)$, $G\in\mathcal{R}$ will be called a covariant one, if

$$(1.44)\qquad X(\omega:G) = X(T_u\omega:G+u) \quad \text{a. s.}$$

holds for each $u\in R_0$ and $G\in\mathcal{R}$.

For example, the random set-function $\hat{X}(\omega:G)$ defined by (1. 21) is additive and also covariant.

THEOREM 2. *Let the random set-function $X(\omega:G)$, $G\in\mathcal{R}$ be additive and covariant. Then*

$$(\mathrm{i})\qquad \sup\{\mathsf{E}|X(\omega:G')|;\ G'\subset F_0,\ G'\in\mathcal{R}\} = K < +\infty$$

implies that

$$(1.45)\qquad \frac{1}{\lambda(G)}X(\omega:G)\overset{1}{\to} Y(\omega) \quad \text{if} \quad G\in\mathcal{R} \quad \text{and} \quad \varrho(G)\to+\infty,$$

where $Y\in L_1(\mathsf{P})$ is invariant under $\mathcal{G}$.

PROOF: Let us introduce the notations $-G=\{x;\ -x\in G\}$ and $Z(\omega)=$ $=X(\omega:F_0)$. As $X(\omega:F_0(v))=X(T_{-v}\omega:F_0)=Z(T_{-v}\omega)$ a. s. for each $v\in R_0$, we have

$$(1.46)\qquad \mathsf{E}|X(\omega:G)-Z(\omega:-\underline{G})| \leqq \frac{1}{\lambda(F_0)}\lambda(\overline{G}^0-\underline{G}^0)K.$$

Consequently, the statement of the theorem follows from Theorem 1. by Lemma 1.

Based on this theorem, the ergodic properties of the additive physical quantities mentioned in the introduction can be proved. We shall see that the "microscopic" entropy is covariant, but it is not additive in general, therefore we need some further results.

2. Preliminaries on generalized entropy

This section summarizes briefly some basic definitions and theorems on generalized entropy. (See I. CSISZÁR [1]) Let $(\Omega, \mathscr{A}, \mathsf{P})$ be a probability space and let Λ be a σ-finite measure on $\mathscr{A}$. The restriction of P resp. of Λ to any σ-algebra $\mathscr{C} \subset \mathscr{A}$ will be denoted by $\mathsf{P}_{\mathscr{C}}$ resp. by $\Lambda_{\mathscr{C}}$. If $\mathsf{P}_{\mathscr{C}} \ll \Lambda_{\mathscr{C}}$ then we set $f_{\mathscr{C}}(\omega) = \dfrac{\mathsf{P}_{\mathscr{C}}(d\omega)}{\Lambda_{\mathscr{C}}(d\omega)}$.

DEFINITION 3. If $\Lambda_{\mathscr{C}}$ is σ-finite on the σ-algebra $\mathscr{C} \subset \mathscr{A}$, then the generalized entropy (Λ-entropy) of P on $\mathscr{C}$ is defined as

$$\mathsf{H}_{\Lambda}(\mathsf{P}:\mathscr{C}) = -\infty \quad \text{if} \quad \mathsf{P}_{\mathscr{C}} \not\ll \Lambda_{\mathscr{C}}, \quad \text{and} \tag{2.1}$$

$$\mathsf{H}_{\Lambda}(\mathsf{P}:\mathscr{C}) = -\int_{\Omega} f_{\mathscr{C}}(\omega) \log f_{\mathscr{C}}(\omega)\, \Lambda(d\omega) = \mathsf{E}(-\log f_{\mathscr{C}}) \tag{2.2}$$

if $\mathsf{P}_{\mathscr{C}} \ll \Lambda_{\mathscr{C}}$. If $\Lambda_{\mathscr{C}}$ is not σ-finite or the integral (2.2) does not exist, then $\mathsf{H}_{\Lambda}(\mathsf{P}:\mathscr{C})$ is not defined.

Let us remark that $\mathsf{H}_{\Lambda}(\mathsf{P}:\mathscr{C}) < +\infty$ implies the existence of $\mathsf{H}_{\Lambda}(\mathsf{P}:\mathscr{C}')$ for each σ-algebra $\mathscr{C}' \supset \mathscr{C}$, and then $\mathsf{H}_{\Lambda}(\mathsf{P}:\mathscr{C}') \leqq \mathsf{H}_{\Lambda}(\mathsf{P}:\mathscr{C})$. This statement holds obviously if $\mathsf{P}_{\mathscr{C}'} \not\ll \Lambda_{\mathscr{C}'}$. In the opposite case we have the well known inequality

$$\int_C f_{\mathscr{C}'}(\omega) \log \frac{f_{\mathscr{C}}(\omega)}{f_{\mathscr{C}'}(\omega)}\, \Lambda(d\omega) = \int_C \log \frac{f_{\mathscr{C}}(\omega)}{f_{\mathscr{C}'}(\omega)}\, \mathsf{P}(d\omega) \leqq 0 \tag{2.3}$$

for each $C \in \mathscr{C}$ as then $\mathsf{P}_{\mathscr{C}}(C) = \mathsf{P}_{\mathscr{C}'}(C)$. Thus for each $C \in \mathscr{C}$ we obtain

$$-\int_C \log f_{\mathscr{C}'}(\omega)\, \mathsf{P}(d\omega) \leqq -\int_C \log f_{\mathscr{C}}(\omega)\, \mathsf{P}(d\omega), \tag{2.4}$$

further

$$\mathsf{E}\left(-\log f_{\mathscr{C}'}(\omega)\right) \leqq \mathsf{E}\left(-\log f_{\mathscr{C}}(\omega)\right) \quad \text{if} \quad \mathscr{C}' \supset \mathscr{C}, \tag{2.5}$$

where equality holds if, and only if, $f_{\mathscr{C}} = f_{\mathscr{C}'}$ a. s. Similarly,

$$\mathsf{E}(-\log^+ f_{\mathscr{C}'}) \leqq \mathsf{E}(-\log^+ f_{\mathscr{C}}) \quad \text{if} \quad \mathscr{C}' \subset \mathscr{C}, \tag{2.6}$$

where $\log^+ x$ denotes the positive part of the function $\log x$; i.e. $\log^+ x = \log x$ if $x \geqq 1$, and $\log^+ x = 0$ if $x < 1$. Namely, if $C = \{\omega;\, f_{\mathscr{C}} \geqq 1\}$, and $C' = \{\omega;\, f_{\mathscr{C}'} \geqq 1\}$, then

$$\begin{aligned} \mathsf{E}(-\log^+ f_{\mathscr{C}'}) = -\int_{C'} \log f_{\mathscr{C}'}\, \mathsf{P}(d\omega) \leqq -\int_C \log f_{\mathscr{C}'}\, \mathsf{P}(d\omega) \leqq -\int_C \log f_{\mathscr{C}}\, \mathsf{P}(d\omega) = \\ = \mathsf{E}(-\log^+ f_{\mathscr{C}}). \end{aligned} \tag{2.7}$$

Definition 3. is motivated by the following two theorems. Theorem 3. has been proved by A. PEREZ [3] in the case when also Λ is a probability measure. His proof is based on martingale convergence theorems. Here we give an elementary proof of the general statement. First we prove a lemma, which implies a result of K. L. CHUNG [4], too.

LEMMA 2. *$(\Omega, \mathscr{A}, \mathsf{P})$ denotes a probability space, and Λ is a σ-finite measure on $\mathscr{A}$. Let $\mathscr{A}_1 \subset \mathscr{A}_2 \subset \ldots \mathscr{A}_n \subset \ldots \mathscr{A}$ be an increasing sequence of σ-algebras such that Λ is σ-finite on $\mathscr{A}_1$ and $\mathsf{P}_{\mathscr{A}_n} \ll \Lambda_{\mathscr{A}_n}$ for every n. If $f_n(\omega) = f_{\mathscr{A}_n}(\omega)$ and $g(\omega) = \inf_n f_n(\omega)$, then the existence of $\mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A}_1)$ implies*

$$\mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A}_1) \leqq \mathsf{E}(-\log g) \leqq 2 \cdot \mathsf{E}\left(\log^+ \frac{1}{f_1}\right) + c + 2, \tag{2.8}$$

where $c = \sum_{k=0}^{\infty} e^{-k} \sum_{i \leqq \frac{k}{2}} e^i < +\infty$.

PROOF: As $g \leqq f_1$, $\mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A}_1) \leqq \mathsf{E}(-\log g)$ is obviously true. The proof of the other inequality is based on the estimate

$$\mathsf{E}(-\log g) \leqq \sum_{k=0}^{\infty} \mathsf{P}(A_k) = \sum_{k=0}^{\infty} \sum_{i=-\infty}^{+\infty} \mathsf{P}(A_k C_i), \tag{2.9}$$

where $A_k = \{\omega; -\log g > k\}$ and $C_i = \{\omega; e^{-i} \leqq f_1 < e^{-i+1}\}$.

Set $A_k^n = \{\omega; f_n < e^{-k}, \inf_{1 \leqq j < n} f_j \geqq e^{-k}\}$. Since $A_k^n C_i \in \mathscr{A}_n$, we have

$$\mathsf{P}(A_k^n C_i) \leqq e^{-k} \Lambda(A_k^n C_i), \tag{2.10}$$

whence

$$\mathsf{P}(A_k C_i) \leqq e^{-k} \Lambda(A_k C_i) \tag{2.11}$$

follows as $A_k^n A_k^m = \emptyset$ if $n \neq m$, and $\sum_{n=1}^{\infty} A_k^n = A_k$. On the other hand, we see that

$$\Lambda(C_i) \leqq e^i \mathsf{P}(C_i) \leqq e^i, \tag{2.12}$$

further

$$\sum_{i=0}^{\infty} i \mathsf{P}(C_i) \leqq 1 + \sum_{i=1}^{\infty} (i-1) \mathsf{P}(C_i) \leqq 1 + \mathsf{E}\left(\log^+ \frac{1}{f_1}\right). \tag{2.13}$$

Using $\mathsf{P}(A_k C_i) \leqq e^{-k} \Lambda(C_i) \leqq e^{-k} e^i$, if $i \leqq \frac{k}{2}$ and $\mathsf{P}(A_k C_i) \leqq \mathsf{P}(C_i)$, if $i < \frac{k}{2}$, it follows that

$$\begin{aligned} \mathsf{E}(-\log g) &\leqq \sum_{k=0}^{\infty} \sum_{i=-\infty}^{+\infty} \mathsf{P}(A_k C_i) = \sum_{k=0}^{\infty} \sum_{i \leqq \frac{k}{2}} \mathsf{P}(A_k C_i) + \sum_{k=0}^{\infty} \sum_{i > \frac{k}{2}} \mathsf{P}(A_k C_i) \leqq \\ &\leqq \sum_{k=0}^{\infty} \sum_{i \leqq \frac{k}{2}} e^{-k} e^i + \sum_{k=0}^{\infty} \sum_{i > \frac{k}{2}} \mathsf{P}(C_i) \leqq c + 2 \sum_{i=1}^{\infty} i \mathsf{P}(C_i) \leqq c + 2 + 2 \mathsf{E}\left(\log^+ \frac{1}{f_1}\right), \end{aligned} \tag{2.14}$$

which proves the lemma.

THEOREM 3. *Assume that $(\Omega, \mathscr{A}, \Lambda)$ is a σ-finite measure space, where the σ-algebra $\mathscr{A}$ is generated by the increasing sequence $\mathscr{A}_1 \subset \mathscr{A}_2 \subset \ldots \mathscr{A}_n \subset \ldots \mathscr{A}$ of σ-algebras, and Λ is σ-finite on $\mathscr{A}_1$. Let P_n be a probability measure on $\mathscr{A}_n$ such that $\mathsf{P}_n = \mathsf{P}_m$ on $\mathscr{A}_n$ if $m > n$. Then the assumptions*

$$\text{(i)} \qquad \mathsf{H}_\Lambda(\mathsf{P}_1 : \mathscr{A}_1) < +\infty,$$

$$\text{(ii)} \qquad \inf_n \mathsf{H}_\Lambda(\mathsf{P}_n : \mathscr{A}_n) > -\infty$$

imply the existence of the extension P of the probability measures P_n to $\mathscr{A}$, P is uniquely determined, $\mathsf{P} \ll \Lambda$, and

$$\log f_n(\omega) \xrightarrow[\text{a. s.}]{} \log f(\omega), \tag{2.15}$$

where $f_n(\omega) = \dfrac{\mathsf{P}_n(d\omega)}{\Lambda(d\omega)}$ *and* $f(\omega) = \dfrac{\mathsf{P}(d\omega)}{\Lambda(d\omega)}$.

PROOF: Set $g_n = \inf_{1 \leqq k \leqq n} f_k$ ($\mathsf{P}_n \ll \Lambda_{\mathscr{A}_n}$ follows from (ii)), and observe that Lemma 2. can be applied to the sequence $\mathscr{A}'_m = \mathscr{A}_m$, if $m < n$, $\mathscr{A}'_m = \mathscr{A}_n$ if $m \geqq n$, and it yields

$$\int_\Omega \log f_n \, \mathsf{P}_n(d\omega) \leqq K + \int_\Omega \log g_n \, \mathsf{P}_n(d\omega), \tag{2.16}$$

where $K = 2\mathsf{E}\left(\log^+ \dfrac{1}{f_1}\right) + c + 2 - \inf_n \mathsf{H}_\Lambda(\mathsf{P}_n : \mathscr{A}_n) < +\infty$, because of (i) and (ii). Set $A_n^b = \{\omega;\ f_n > b\}$. As $\log f_n \geqq \log g_n$, further $\log^+ f_1 \geqq 0$, and $\log^+ f_1 \geqq \log g_n$, we have

$$\begin{aligned} \log b \int_{A_n^b} f_n \Lambda(d\omega) = \log b \, \mathsf{P}_n(A_n^b) \leqq \int_{A_n^b} \log f_n \, \mathsf{P}_n(d\omega) \leqq \\ \leqq K + \int_{A_n^b} \log g_n \, \mathsf{P}_n(d\omega) \leqq K + \mathsf{E}(\log^+ f_1) < +\infty. \end{aligned} \tag{2.17}$$

This means, that the sequence f_n is uniformly integrable with respect to Λ.

Let us consider now the probability measures P_n. The relation

$$\mathsf{P}(A) = \mathsf{P}_n(A) \quad \text{if} \quad A \in \mathscr{A}_n \tag{2.18}$$

defines a nonnegative, additive set function on the algebra $\mathscr{A}^\circ = \bigcup_{n=1}^{\infty} \mathscr{A}_n$. The σ-additivity of P follows from (2.17), namely for each B and for each n and $b > 0$ we have

$$\int_B f_n \Lambda(d\omega) \leqq b\Lambda(B) + \int_{B A_n^b} f_n \Lambda(d\omega) \tag{2.19}$$

implying $\lim \mathsf{P}(B_n) = 0$ if $B_n \in \mathscr{A}^\circ$ is such a decreasing sequence that $\lim_n \Lambda(B_n) = 0$. Consequently, (2.18) defines a probability measure P on $\mathscr{A}$, which is the unique extension of the probability measures P_n.

The following step of the proof is to show that

$$\liminf f_n = \limsup f_n \quad \text{a. s.} \tag{2.20}$$

As Λ is σ-finite on $\mathscr{A}_1$, it may be assumed here that Λ is a finite measure. To prove (2. 20), it is sufficient to show that $\mathsf{P}(C_a^b)=0$ if $a<b$, where

$$C_a^b=\{\omega;\ \liminf f_n<a<b<\limsup f_n\}, \tag{2.21}$$

since the divergence set is the sum of these events, if a and b range over the set of rationals.

Let us consider the events

$$A_n=\{\omega;\ \text{there exist}\ \ k,m\ \ \text{with}\ \ n\leqq k<m\ \ \text{and}\ \ f_k<a, f_m>b\}, \tag{2.22}$$

$$B_n=\{\omega;\ \text{there exist}\ \ k,m\ \ \text{with}\ \ n\leqq k<m\ \ \text{and}\ \ f_k>b, f_m<a\}. \tag{2.23}$$

Observe, that the sequences $\{A_n\}$ and $\{B_n\}$ are decreasing, further

$$\prod_{n=1}^{\infty} A_n = \prod_{n=1}^{\infty} B_n = D \supset C_a^b. \tag{2.24}$$

We show that

$$\mathsf{P}(A_n) \geqq b\Lambda(A_n) \quad \text{and} \quad \mathsf{P}(B_n) \leqq a\Lambda(B_n), \tag{2.25}$$

whence

$$\mathsf{P}(D) \geqq b\Lambda(D) \quad \text{and} \quad \mathsf{P}(D) \leqq a\Lambda(D) \tag{2.26}$$

follows by the continuity property of the finite measures P and Λ. As $a<b$, this is possible only if $\Lambda(D)=\mathsf{P}(D)=0$; consequently (2. 25) implies $\mathsf{P}(C_a^b)=0$.

To prove (2. 25) we introduce the events

$$A_n^m=\{\omega; f_m>b,\ \inf_{n\leqq k<m} f_k<a,\ \text{and if}\ \ f_k<a, k\geqq n,\ \text{then}\ \ \sup_{k\leqq j<m} f_j \leqq b\}; \tag{2.27}$$

that is after the first $k\geqq n$ with $f_k<a$, the relation $f_j>b$ holds only for $j=m$ if $j\leqq m$. Since $f_m>b$ on $A_n^m\in\mathscr{A}_m$, we have

$$\mathsf{P}(A_n^m)=\int_{A_n^m} f_m \Lambda(d\omega) \geqq b\Lambda(A_n^m) \quad \text{for each} \quad m>n, \tag{2.28}$$

which implies the first part of (2. 25) as $A_n^i A_n^j=\emptyset$ if $i\neq j$ and $A_n=\sum_{m>n} A_n^m$. The second part of (2. 25) can be proved by a quite similar argument using the events

$$B_n^m=\{\omega; f_m<a,\ \sup_{n\leqq k<m} f_k>b,\ \text{and if}\ \ f_k>b,\ \text{then}\ \ \inf_{k\leqq j<m} f_j \geqq a\}$$

instead of A_n^m.

Thus we have a nonnegative function $f(\omega)=\lim_n f_n(\omega)$, which is measurable on $\mathscr{A}$, and the sequence f_n is uniformly integrable with respect to Λ. Observe that $\mathsf{P}(A)=\lim_n \int_A f_n \Lambda(d\omega)$ if $A\in\mathscr{A}^\circ$. As the uniform integrability of the sequence f_n

implies $f_n \to f$ in the $L_1(\Lambda)$ norm if Λ is a finite measure (see M. LOEVE [9], p. 163. L_r-Convergence Theorem), we have

$$\mathsf{P}(A) = \int_A f(\omega)\Lambda(d\omega) \tag{2.29}$$

for each $A \in \mathscr{A}^\circ$ with $\Lambda(A) < +\infty$. However, as Λ is σ-finite, these events form an algebra generating $\mathscr{A}$, consequently (2. 29) is true for each $A \in \mathscr{A}$; that is $\mathsf{P} \ll \Lambda$ and $f = \dfrac{\mathsf{P}(d\omega)}{\Lambda(d\omega)} = \lim_n f_n$ a. s.

On account of the L_r-Convergence Theorem, mentioned above, the final statement (2. 15) follows from

$$\lim_n \mathsf{E}\,|\log f_n| = \mathsf{E}\,|\log f|, \tag{2.30}$$

which can be proved by means of the Fatou-Lebesgue Theorem. Namely, we know from Lemma 2. and from (i) and (ii) that $\log g = \inf_n \log f_n$ has a finite expectation, therefore

$$\liminf_n \mathsf{E}\,(\log f_n) \geqq \mathsf{E}\,(\log f). \tag{2.31}$$

Further, in view of (2. 5), we have $\mathsf{E}(\log f_n) \leqq \mathsf{E}(\log f)$ for each n, whence we obtain

$$\lim_n \mathsf{E}\,(\log f_n) = \mathsf{E}\,(\log f). \tag{2.32}$$

On the other hand, as $\log^+ f_n \geqq 0$, we have

$$\liminf_n \mathsf{E}(\log^+ f_n) \geqq \mathsf{E}\,(\log^+ f), \tag{2.33}$$

whence

$$\lim_n \mathsf{E}(\log^+ f_n) = \mathsf{E}(\log^+ f) \tag{2.34}$$

follows by (2. 6). The comparison of (2. 32) and (2. 34) yields (2. 30), thus the proof of Theorem 3. is complete.

This theorem involves its following generalization.

THEOREM 4. *Let* $(\Omega, \mathscr{A}, \mathsf{P})$ *be a probability space and let* Λ *be a* σ*-finite measure on* $\mathscr{A}$. *If the system* $\{\mathscr{A}_\gamma;\ \gamma \in \Gamma\}$ *of* σ*-algebras* $\mathscr{A}_\gamma \subset \mathscr{A}$ *satisfies the following conditions:*

(i) *For each* $\gamma_1, \gamma_2 \in \Gamma$ *there exists a* $\gamma_3 \in \Gamma$ *such that*

$$\mathscr{A}_{\gamma_3} \supset \mathscr{A}_{\gamma_1} \cup \mathscr{A}_{\gamma_2}$$

(ii)
$$\mathscr{A} = \sigma\Big(\bigcup_{\gamma \in \Gamma} \mathscr{A}_\gamma\Big)$$

(iii)
$$-\infty < \inf_{\gamma \in \Gamma} \mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A}_\gamma) < +\infty,$$

then $\mathsf{P} \ll \Lambda$, $\mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A})$ *exists and*

$$\mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A}) = \inf_{\gamma \in \Gamma} \mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A}_\gamma), \tag{2.35}$$

further, with the notations $f_\gamma(\omega)=f_{\mathscr{A}_\gamma}(\omega)$, $f(\omega)=f_{\mathscr{A}}(\omega)$, *we have*

$$\log f_\gamma(\omega) \xrightarrow{1} \log f(\omega) \tag{2.36}$$

in the sense that for each $\varepsilon>0$ *there exists a* $\gamma_\varepsilon\in\Gamma$ *such that*

$$\mathsf{E}\,|\log f_\gamma-\log f|<\varepsilon \quad \textit{if} \quad \gamma>\gamma_\varepsilon. \tag{2.37}$$

PROOF: In view of (i) and (iii) there exists a sequence $\gamma_1, \gamma_2, \dots \gamma_n \dots$ in Γ such that $\mathscr{A}_{\gamma_1}\subset\mathscr{A}_{\gamma_2}\subset\dots\subset\mathscr{A}_{\gamma_n}\dots$, $\mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A}_{\gamma_1})<+\infty$, further

$$\inf_{\gamma\in\Gamma}\mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A}_\gamma)=h=\inf_n\mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A}_{\gamma_n}). \tag{2.38}$$

On accout of Theorem 3. we know that $\mathsf{P}\ll\Lambda$ on $\mathscr{A}^*=\sigma(\bigcup_n\mathscr{A}_{\gamma_n})$, further

$$\mathsf{E}(-\log f^*)=h, \tag{2.39}$$

if $f^*(\omega)=f_{\mathscr{A}^*}(\omega)$. Now let $\mathscr{A}_\gamma$ be an arbitrary element of our system. As $\mathscr{A}_\gamma^*=\sigma(\mathscr{A}_\gamma\cup\mathscr{A}^*)=\sigma\big(\bigcup_n\sigma(\mathscr{A}_\gamma\cup\mathscr{A}_{\gamma_n})\big)$ and the sequence $\sigma(\mathscr{A}_\gamma\cup\mathscr{A}_{\gamma_n})$ is increasing, (2. 5) and Theorem 3. imply that $\mathsf{P}\ll\Lambda$ also on $\mathscr{A}_\gamma^*$, and

$$\mathsf{E}(-\log f_\gamma^*)=h=\mathsf{E}(-\log f^*), \tag{2.40}$$

where $f_\gamma^*=f_{\mathscr{A}_\gamma^*}$. On account of $\mathscr{A}_\gamma^*\subset\mathscr{A}^*$ we see from (2. 5) that $f_\gamma^*=f^*$ a. s. This means, that

$$\mathsf{P}(A)=\int_A f^*(\omega)\Lambda(d\omega) \quad \text{if} \quad A\in\bigcup_{\gamma\in\Gamma}\mathscr{A}_\gamma^*, \tag{2.41}$$

however $\mathscr{A}=\sigma(\bigcup_{\gamma\in\Gamma}\mathscr{A}_\gamma^*)$, thus (2. 41) is true even if $A\in\mathscr{A}$; that is $\mathsf{P}\ll\Lambda$, $f^*=f$ almost surely, further

$$\mathsf{H}_\Lambda(\mathsf{P}:\mathscr{A})=\mathsf{E}(-\log f)=h. \tag{2.42}$$

From Theorem 3. we see, that

$$\log f_{\gamma_n}\xrightarrow{1}\log f \tag{2.43}$$

holds for each increasing sequence $\mathscr{A}_{\gamma_n}$, satisfying (2. 38), from which the validity of (2. 37) follows by assumption (i). The proof is complete.

REMARK: It is easy to see that Theorem 4. remains valid even if P is not defined on $\mathscr{A}$, but we have a consistent system P_γ; $\gamma\in\Gamma$ of probability measures on the corresponding σ-algebras $\mathscr{A}_\gamma$.

Finally, we define the conditional generalized entropy in the following situation. (See I. CSISZÁR [1] § 3.). $(\Omega, \mathscr{A}, \mathsf{P})$ is a probability space, let $\mathscr{A}_1$ and $\mathscr{A}_2$ be qualitatively independent σ-algebras — i.e. $A_1A_2=\emptyset$ holds with $A_1\in\mathscr{A}_1$, $A_2\in\mathscr{A}_2$ only if $A_1=\emptyset$ or $A_2=\emptyset$ — such that $\mathscr{A}=\mathscr{A}_1\times\mathscr{A}_2=\sigma\{A_1A_2;\ A_1\in\mathscr{A}_1, A_2\in\mathscr{A}_2\}$. Further, let Λ_1, and. Λ_2 be σ-finite measures on $\mathscr{A}_1$ and on $\mathscr{A}_2$, resp. Because of the qualitative independence of $\mathscr{A}_1$, and $\mathscr{A}_2$, by

$$\Lambda(A_1A_2)=\Lambda_1(A_1)\cdot\Lambda_2(A_2) \quad \text{if} \quad A_1\in\mathscr{A}_1, A_2\in\mathscr{A}_2, \tag{2.44}$$

a uniquely determined σ-finite measure is defined on $\mathscr{A} = \mathscr{A}_1 \times \mathscr{A}_2$. This product measure will be denoted by $\Lambda = \Lambda_1 \times \Lambda_2$. Let us consider the other product measure $\Lambda' = \mathsf{P}_1 \times \Lambda_2$, too, where P_1 denotes the restriction of P to $\mathscr{A}_1$. It is easy to see, that $\mathsf{P} \ll \Lambda$ if, and only if, $\mathsf{P}_1 \ll \Lambda_1$ and $\mathsf{P} \ll \Lambda'$; further, then we have

$$(2.45) \qquad \frac{\mathsf{P}(d\omega)}{\Lambda'(d\omega)} = \frac{f(\omega)}{f_1(\omega)}, \quad \text{where} \quad f(\omega) = \frac{\mathsf{P}(d\omega)}{\Lambda(d\omega)} \quad \text{and} \quad f_1(\omega) = \frac{\mathsf{P}_1(d\omega)}{\Lambda_1(d\omega)}.$$

Consequently

$$(2.46) \qquad \mathsf{H}_{\Lambda'}(\mathsf{P}:\mathscr{A}) = \mathsf{E}\left(-\log \frac{f}{f_1}\right) \quad \text{if} \quad \mathsf{P} \ll \Lambda',$$

further

$$(2.47) \qquad \mathsf{H}_{\Lambda}(\mathsf{P}:\mathscr{A}) = \mathsf{H}_{\Lambda_1}(\mathsf{P}:\mathscr{A}_1) + \mathsf{H}_{\Lambda'}(\mathsf{P}:\mathscr{A}),$$

provided that its terms are defined and the sum on the right hand side exists. These relations suggest the following definition:

DEFINITION 5. In the above situation we define the conditional Λ_2-entropy of P on $\mathscr{A}$, given the σ-algebra $\mathscr{A}_1$ as

$$(2.48) \qquad \mathsf{H}_{\Lambda_2}(\mathsf{P}:\mathscr{A}_2 \,|\, \mathscr{A}_1) = \mathsf{H}_{\Lambda'}(\mathsf{P}:\mathscr{A}),$$

provided that the generalized entropy $\mathsf{H}_{\Lambda'}(\mathsf{P}:\mathscr{A})$ exists.

Observe that $\mathsf{H}_{\Lambda_2}(\mathsf{P}:\mathscr{A}_2|\mathscr{A}_1)$ does not depend on the choice of Λ_1, so that it is defined even if P_1 only is given on $\mathscr{A}_1$; of course, then $\mathsf{P}_1 = \Lambda_1$ may be chosen, and (2. 17) is trivially true.

With this notation (2. 47) becomes

$$(2.49) \qquad \mathsf{H}_{\Lambda}(\mathsf{P}:\mathscr{A}) = \mathsf{H}_{\Lambda_1}(\mathsf{P}:\mathscr{A}_1) + \mathsf{H}_{\Lambda_2}(\mathsf{P}:\mathscr{A}_2 \,|\, \mathscr{A}_1),$$

further, we have

$$(2.50) \qquad \mathsf{H}_{\Lambda_2'}(\mathsf{P}:\mathscr{A}_2 \,|\, \mathscr{A}_1) \leqq \mathsf{H}_{\Lambda_2}(\mathsf{P}:\mathscr{A}_2)$$

with equality if, and only if, $\mathscr{A}_1$ and $\mathscr{A}_2$ are P-independent. We see that the well known properties of conditional entropy are valid in our case, too.

As the conditional entropy is a special type of generalized entropy, Theorem 4. can be formulated also for conditional entropies, without any further difficutly.

Before formulation of the theorem, we introduce some notations. Let $\mathscr{C}$ and $\mathscr{A}_0$ be qualitatively independent σ-algebras in the probability space $(\Omega, \mathscr{A}, \mathsf{P})$, such that $\mathscr{A} = \mathscr{C} \times \mathscr{A}_0$. $\{\mathscr{C}_\gamma;\ \gamma \in \Gamma\}$ is a system of sub-σ-algebras of $\mathscr{C}$, further Λ_0 is a σ-finite measure on $\mathscr{A}_0$. The restriction of P to $\mathscr{C}_\gamma$ resp. to $\mathscr{C}_\gamma \times \mathscr{A}_0$ will be denoted by $\overline{\mathsf{P}}_\gamma$ resp. by $\overline{\mathsf{P}}_{\gamma_0}$, while $\overline{\mathsf{P}}$ denotes the restriction of P to $\mathscr{C}$. If $\mathsf{P} \ll \overline{\mathsf{P}} \times \Lambda_0$, then we set $g_\gamma(\omega) = \dfrac{\overline{\mathsf{P}}_{\gamma_0}(d\omega)}{(\overline{\mathsf{P}}_\gamma \times \Lambda_0)(d\omega)}$, $g(\omega) = \dfrac{\mathsf{P}(d\omega)}{(\overline{\mathsf{P}} \times \Lambda_0)(d\omega)}$.

Observe, that $\overline{\mathsf{P}}_\gamma \times \Lambda_0$ is exactly the restriction of $\overline{\mathsf{P}} \times \Lambda_0$ to $\mathscr{C}_\gamma \times \mathscr{A}_0$, thus we have for each $\gamma \in \Gamma$

$$(2.51) \qquad \mathsf{H}_{\Lambda_0}(\mathsf{P}:\mathscr{A}_0 \,|\, \mathscr{C}) \leqq \mathsf{H}_{\Lambda_0}(\mathsf{P}:\mathscr{A}_0 \,|\, \mathscr{C}_\gamma) \leqq \mathsf{H}_{\Lambda_0}(\mathsf{P}:\mathscr{A}_0),$$

furthermore Theorem 4. implies:

THEOREM 5. *If the system* $\{\mathscr{C}_\gamma;\ \gamma\in\Gamma\}$ *satisfies the assumptions:*

(i) *For each* $\gamma_1,\gamma_2\in\Gamma$ *there exists a* $\gamma\in\Gamma$ *such that* $\mathscr{C}_\gamma\supset\mathscr{C}_{\gamma_1}\cup\mathscr{C}_{\gamma_2}$

(ii) $\mathscr{C}=\sigma\left(\bigcup_{\gamma\in\Gamma}\mathscr{C}_\gamma\right)$

(iii) $-\infty<\inf_{\gamma\in\Gamma}\mathsf{H}_{\Lambda_0}(\mathsf{P}:\mathscr{A}_0|\mathscr{C}_\gamma)<+\infty,$

then

$$\mathsf{H}_{\Lambda_0}(\mathsf{P}:\mathscr{A}_0|\mathscr{C})=\inf_{\gamma\in\Gamma}\mathsf{H}_{\Lambda_0}(\mathsf{P}:\mathscr{A}_0|\mathscr{C}_\gamma), \tag{2.52}$$

furthermore $\mathsf{P}\ll\bar{\mathsf{P}}\times\Lambda_0$ *and*

$$\log g_\gamma(\omega)\xrightarrow{1}\log g(\omega). \tag{2.53}$$

3. The r-dimensional McMillan theorem

The generalization of McMillan's theorem is formulated in terms of a probability space $(\Omega,\mathscr{A},\mathsf{P})$ having the structure, detailed in **A**, **B** and **C**. (We use the notations of the preceding sections, unless stated otherwise.)

A: The probability space $(\Omega,\mathscr{A},\mathsf{P})$ has an Abelian group $\mathscr{G}=\{T_u;\ u\in R_0\}$ of automorphisms, that is

$$T_uT_v=T_{u+v}\quad\text{if}\quad u\in R_0,\ v\in R_0, \tag{3.1}$$

$$\mathsf{P}(T_uA)=\mathsf{P}(A)\quad\text{for each}\quad A\in\mathscr{A},\ u\in R_0. \tag{3.2}$$

B: To every Borel set $G\in\mathscr{R}$ there is assigned a sub-σ-algebra $\mathscr{A}_G$ of $\mathscr{A}$ such that

$$\mathscr{A}_G=\sigma\left(\bigcup_r\mathscr{A}_{G_n}\right)\quad\text{if}\quad G=\bigcup_n G_n, \tag{3.3}$$

further $\mathscr{A}_{G_1}$ and $\mathscr{A}_{G_2}$ are qualitatively independent if $G_1\cap G_2=\emptyset$, thus

$$\mathscr{A}_{G_1\cup G_2}=\mathscr{A}_{G_1}\times\mathscr{A}_{G_2} \tag{3.4}$$

If G is a bounded Borel set, then a finite measure Λ_G is defined on $\mathscr{A}_G$ in such a way that

$$\Lambda_{G_1\cup G_2}=\Lambda_{G_1}\times\Lambda_{G_2}\quad\text{if}\quad G_1\cap G_2=\emptyset. \tag{3.5}$$

C: The connection of **A** and **B** is expressed by

$$T_u\mathscr{A}_G=\mathscr{A}_{G+u}\quad\text{for every}\quad u\in R_0,\ G\in\mathscr{R},\quad\text{and} \tag{3.6}$$

$$\Lambda_{G+u}(T_uA)=\Lambda_G(A)\quad\text{if}\quad A\in\mathscr{A}_G\quad\text{and}\quad G\quad\text{is bounded}, \tag{3.7}$$

that is each $T_u\in\mathscr{G}$ preserves the measures Λ_G, too.

Let us remark that these assumptions are satisfied — for instance — in case of an r-dimensional point process. (See J. FRITZ [2]) In general, each probability space satisfying the postulates **A**, **B**, **C**, can be represented as that of a random set-function taking values in an abstract set.

We introduce some notations, $\overline{\mathsf{P}}_G$ denotes the restriction of P to $\mathscr{A}_G$ (but Λ_G is not the restriction of any measure Λ, in general). If G, $G' \in \mathscr{R}$, $G \cap G' = \emptyset$ and G is bounded, then we set

$$\mathsf{H}(G) = \mathsf{H}_{\Lambda_G}(\mathsf{P}:\mathscr{A}_G), \tag{3.8}$$

$$\mathsf{H}(G|G') = \mathsf{H}_{\Lambda_G}(\mathsf{P}:\mathscr{A}_G|\mathscr{A}_{G'}), \tag{3.9}$$

$$f(\omega:G) = \frac{\overline{\mathsf{P}}_G(d\omega)}{\Lambda_G(d\omega)} \tag{3.10}$$

$$f(\omega:G|G') = \frac{\overline{\mathsf{P}}_{G\cap G'}(d\omega)}{(\mathsf{P}_{G'} \times \Lambda_G)(d\omega)} \tag{3.11}$$

that is

$$\mathsf{H}(G) = \mathsf{E}(-\log f(\omega:\mathrm{G})), \tag{3.12}$$

$$\mathsf{H}(G|G') = \mathsf{E}(-\log f(\omega:\mathrm{G}|G')) \tag{3.13}$$

hold in the absolutely continuous case, which will be assumed.

From (3. 4) and from (3. 5) we see that $\mathscr{A}_\emptyset = \{\Omega, \emptyset\}$ and $\Lambda_\emptyset(\Omega) = 1$, consequently $\mathsf{H}(G|\emptyset) = \mathsf{H}(G)$ and $f(\omega:G|\emptyset) = f(\omega:G)$.

Let G_1 and G_2 be bounded sets of $\mathscr{R}$, $G_1 \cap G_2 = \emptyset$ and $G = G_1 \cup G_2$. Then from (2. 49), (2. 50) and from (2. 45) we obtain

$$\mathsf{H}(G) = \mathsf{H}(G_1) + \mathsf{H}(G_2|G_1) \leqq \mathsf{H}(G_1) + \mathsf{H}(G_2), \tag{3.14}$$

provided that its terms are defined, while

$$f(\omega:G) = f(\omega:G_2|G_1)\, f(\omega:G_1) \quad \text{a. s.} \tag{3.15}$$

holds in the absolutely continuous case. These relations have a more general form, namely, if $G' \cap G = \emptyset$ with an arbitrary set $G' \in \mathscr{R}$ such that $\overline{\mathsf{P}}_{G\cup G'} \ll \Lambda_{G_2} \times \overline{\mathsf{P}}_{G'\cup G_1}$ and $\overline{\mathsf{P}}_{G'\cup G_1} \ll \overline{\mathsf{P}}_{G'} \times \Lambda_G$, then $\overline{\mathsf{P}}_{G'\cup G_1} \times \Lambda_{G_2} \ll \overline{\mathsf{P}}_{G'} \times \Lambda_{G'} = \overline{\mathsf{P}}_{G_1} \times \Lambda_{G_1} \times \Lambda_{G_2}$ and the corresponding Radon-Nikodym derivative is exactly $f(\omega:G_1|G')$, therefore $\overline{\mathsf{P}}_{G'\cup G} \ll \ll \overline{\mathsf{P}}_{G'} \times \Lambda_G$, further

$$f(\omega:G|G') = f(\omega:G_2|G' \cup G_1) f(\omega:G_1|G') \quad \text{a. s.}, \tag{3.16}$$

$$\mathsf{H}(G|G') = \mathsf{H}(G_2|G' \cup G_1) + \mathsf{H}(G_1|G'). \tag{3.17}$$

It is easy to check that (3. 17) remains valid in the general case, too.

If $G'' \supset G'$, $G'' \cap G = \emptyset$ and G is bounded, then (2. 51) becomes:

$$\mathsf{H}(G|G'') \leqq \mathsf{H}(G|G') \leqq \mathsf{H}(G). \tag{3.18}$$

As every transformation $T_u \in \mathscr{G}$ preserves both $\overline{\mathsf{P}}_G$ and Λ_G, the entropies $\mathsf{H}(G)$ and $\mathsf{H}(G|G')$ are invariant under translations of the sets G and G', further

$$f(T_u\omega:G+u) = f(\omega:G) \quad \text{a. s.}, \tag{3.19}$$

$$f(T_u\omega:G+u|G'+u) = f(\omega:G|G') \quad \text{a. s.}, \tag{3.20}$$

where G, $G' \in \mathscr{R}$, G is bounded and $G \cap G' = \emptyset$. Of course, (3. 20) is true even if G' is not bounded.

The proof of the generalized McMillan theorem is based on a decomposition of the following type: Let $\emptyset = G_0 \subset G_1 \subset \cdots \subset G_{n-1} \subset G_n = G$ be an increasing sequence of bounded sets. Then the sets $G_k - G_{k-1}$, $k = 1, 2, \ldots, n$ form a partition of G, thus (3. 15) implies

$$\log f(\omega : G) = \sum_{k=1}^{n} \log f(\omega : G_k - G_{k-1} | G_{k-1}) \text{ a. s.} \tag{3. 21}$$

If $r = 1$, then the suitable decomposition is obtained in a quite natural way. In the general case it will be constructed by means of the set F_+ defined by the recursion:

$$F_+^{(1)} = (-\infty, +\infty) \tag{3. 22}$$

$$F_+^{(2)} = \left\{(x_1, x_2);\ x_1 < -\frac{1}{2} \text{ or } -\frac{1}{2} \leqq x_1 < \frac{1}{2},\ x_2 < -\frac{1}{2}\right\} = \tag{3. 23}$$

$$= F_+^{(1)} \times (-\infty, +\infty) \cup F_0^{(1)} \times \left(-\infty, -\frac{1}{2}\right),$$

and so on, finally

$$F_+ = F_+^{(r)} = F_+^{(r-1)} \times (-\infty, +\infty) \cup F_0^{(r-1)} \times \left(-\infty, -\frac{1}{2}\right) \tag{3. 24}$$

if $F_+^{(r-1)}$ is already defined; where

$$F_0^{(k)} = \left\{(x_1, x_2, \ldots, x_k);\ -\frac{1}{2} \leqq x_i < +\frac{1}{2} \text{ for } 1 \leqq i \leqq k\right\}. \tag{3. 25}$$

For the translates $F_+(u) = F_+ + u$ of F_+ the following lemma holds:

LEMMA 3. *For every u, $v \in R_0$ we have*

$$F_+(u) \subset F_+(v) \quad \text{or} \quad F_+(u) \supset F_+(v), \tag{3. 26}$$

further $F_+(u) \subset F_+(v)$ holds in the strict sense if, and only if, $F_0(u) \subset F_+(v)$. Consequently, the relation $F_+(u) \prec F_+(v)$ if $F_+(u) \subset F_+(v)$, but $u \neq v$, orders the sets $F_+(u)$, $u \in R_0$ in such a way, that exactly $F_+(v + e_r) = F_+(v) \cup F_0(v)$ is the first one among them, which contains the set $F_+(v)$.

PROOF: It is sufficient to consider the case $v = 0$. In the one-dimensional case the statement is obviously true, let us assume its validity also for $k = r - 1$. Let $u = n_1 e_1 + n_2 e_2 + \cdots + n_r e_r$ and set $w = u - n_r e_r$. In view of the definition of F_+ we have

$$F_+(u) = (F_+^{(r-1)} + w) \times (-\infty, +\infty) \cup (F_0^{(r-1)} + w) \times \left(-\infty, +\frac{1}{2} - n_r\right). \tag{3. 27}$$

Now, if $w = 0$, then we see that $F_+ \subset F_+(u)$ and $F_0 \subset F_+(u)$ if $n_r > 0$, while $F_+(u) \subset$

$\subset F_+$, $F_0(u) \subset F_+$ if $n_r < 0$. If $w \neq 0$, then, in view of our assumption, for example $F_+^{(r-1)} \subset F_+ + w$ and $F_0^{(r-1)} \subset F_+^{(r-1)} + w$; therefore (3. 24) and (3. 27) imply that

$$F_+ \subset (F_+^{(r-1)} + w) \times (-\infty, +\infty) \subset F_+(u), \tag{3.28}$$

$$F_0 = F_0^{(r-1)} x \left[-\frac{1}{2}, \frac{1}{2}\right] \subset (F_+^{(r-1)} + w) \times (-\infty, +\infty) \subset F_+(u), \tag{3.29}$$

which prove the lemma.

We are now in a position to prove the main result of this paper: the generalization of McMillan's theorem to r-dimensional random set-functions.

THEOREM 6. *If the probability space* $(\Omega, \mathscr{A}, \mathsf{P})$ *satisfies the postulates* **A**, **B**, **C**, *further*

(i) $\sup \{\mathsf{H}(G);\ G \subset F_0,\ G \in \mathscr{R}\} < +\infty$,

(ii) $\inf_n \dfrac{\mathsf{H}(F_n)}{\lambda(F_n)} = H > -\infty$,

then there exists an invariant function $h(\omega)$ *such that*

$$\mathsf{E}(-h(\omega)) = \inf \left\{ \frac{\mathsf{H}(G)}{\lambda(G)};\ G \in \mathscr{B} \right\} = H, \tag{3.30}$$

and

$$\frac{1}{\lambda(G)} \log f(\omega : G) \xrightarrow{1} h(\omega) \quad \textit{if} \quad G \in \mathscr{B} \quad \textit{and} \quad \varrho(G) \to +\infty. \tag{3.31}$$

PROOF: First we treate some simple consequences of the assumptions of the theorem. Because of (i), there exists a finite number K_1 such that

$$\mathsf{H}(G|G') \leqq \mathsf{H}(G) \leqq K_1 \quad \text{if} \quad G \subset F_0 \quad \text{and} \quad G \cap G' = \emptyset, \tag{3.32}$$

thus (3. 14) and the invariance of the entropy imply

$$\mathsf{H}(G) \leqq \sum_{F_0(u) \cap G \neq \emptyset} \mathsf{H}(G \cap F_0(u)) \leqq \frac{\lambda(\bar{G}^\circ)}{\lambda(F_0)} K_1 < +\infty \tag{3.33}$$

for every bounded $G \in \mathscr{R}$, namely $\dfrac{\lambda(\bar{G}^\circ)}{\lambda(F_0)}$ is exactly the number of terms of the sum in (3. 33). We shall show that $\mathsf{H}(G) > -\infty$ if G is bounded, that is then $\overline{\mathsf{P}}_G \ll \Lambda_G$. On account of Lemma 3. the sets $\{G \cap F_+(u);\ F_0(u) \cap G \neq \emptyset\}$ form a finite increasing sequence, if G is bounded, such that

$$G \cap F_+(u) = \cup \{G \cap F_0(v);\ F_+(v) \prec F_+(u)\}, \tag{3.34}$$

consequently we have the decomposition

$$\mathsf{H}(G) = \sum_{F_0(u) \cap G \neq \emptyset} \mathsf{H}(G \cap F_0(u) \,|\, G \cap F_+(u)), \tag{3.35}$$

which implies the finiteness of $\mathsf{H}(G)$ by the following relation:

$$\sup\{|\mathsf{H}(G|G')|;\, G\subset F_0,\, G'\subset F_+\} = K < +\infty. \tag{3.36}$$

To prove (3. 36), let us consider the sequence $\mathsf{H}(F_0|F_+\cap F_m)$. Putting $G=F_n$ in (3. 55), we obtain

$$\mathsf{H}(F_n) = \sum_{F_0(u)\subset F_n} \mathsf{H}\big(F_0(u)|F_n\cap F_+(u)\big). \tag{3.37}$$

If $n>m$ and $F_n\supset F_m+u$, then (3. 18) implies $\mathsf{H}\big(F_0(u)|F_n\cap F_+(u)\big)\leqq$ $\leqq \mathsf{H}\big(F_0(u)|F_m+u\cap F_+(u)\big) = \mathsf{H}(F_0|F_m\cap F_+)$, while $\mathsf{H}\big(F_0(u)|F_n\cap F_+(u)\big)\leqq K_1$ otherwise. Therefore

$$\begin{aligned} H \leqq \frac{\mathsf{H}(F_n)}{\lambda(F_n)} \leqq \frac{1}{\lambda(F_n)}\big((2n-2m+1)^r\,\mathsf{H}(F_0|F_m\cap F_+)\big) + \\ + \frac{1}{\lambda(F_n)}\big((2n+1)^r-(2n-2m+1)^r\big)K_1, \end{aligned} \tag{3.38}$$

from which

$$\lambda(F_0)\,H \leqq \inf_n \mathsf{H}(F_0|F_+\cap F_m) \tag{3.39}$$

follows by letting $n\to+\infty$, since $\lambda(F_n)=\lambda(F_0)(2n+1)^r$. From (3. 39) by Theorem 5. we obtain that

$$\mathsf{H}(F_0|F_+)\geqq\lambda(F_0)H. \tag{3.40}$$

However, (3. 17), (3. 18) and (3. 32) imply $\mathsf{H}(F_0|F_+) = \mathsf{H}(F_0-G|F_+\cup G)+$ $+\mathsf{H}(G|F_+)\leqq K_1+\mathsf{H}(G|F_+)\leqq K_1+\mathsf{H}(G|G')$ if $G\subset F_0$ and $G'\subset F_+$, which proves (3. 36).

A further consequence of Theorem 5. and of (3. 39) is the statement, that for each $\varepsilon>0$ there exists a natural integer m_ε such that

$$\mathsf{E}\,|\log f(\omega:F_0|G'\cap F_+)-\log f(\omega:F_0|F_+)|<\varepsilon \tag{3.41}$$

holds for each $G'\in\mathscr{R}$ if $G'\cap F_+\supset F_m\cap F_+$ and $m\geqq m_\varepsilon$.

To prove the theorem, let us consider the following decomposition of $\log f(\omega:G)$

$$\log f(\omega:G) = \sum_{F_0(u)\cap G\neq\emptyset} \log f\big(\omega:G\cap F_0(u)|\,G\cap F_+(u)\big), \tag{3.42}$$

the validity of which follows from Lemma 3. in the same way as (3. 35) has been proved. In view of (3. 20), we obtain from (3. 41) that

$$\mathsf{E}\,\big|\log f\big(\omega:G\cap F_0(u)|\,G\cap F_+(u)\big)-\log f\,(T_{-u}\omega:F_0|F_+)\big|<\varepsilon \tag{3.43}$$

if $F_m+u\subset G$ and $m\geqq m_\varepsilon$, while

$$\mathsf{E}\,\big|\log f\big(\omega:G\cap F_0(u)|\,G\cap F_+(u)\big)-\log f(T_{-u}\omega:F_0|F_+)\big|\leqq K'+\big|\mathsf{H}(F_0|F_+)\big| \tag{3.44}$$

holds otherwise, as follows from (3. 36), where $K'=K+2\Lambda_{F_0}(\Omega)$.

On the other hand, on account of Theorem 1. we have an invariant function $h(\omega)$ such that

$$\mathsf{E}\left|\frac{1}{\lambda(G)}\sum_{u\in G\cap R_0}\log f(T_{-u}\omega:F_0\,|\,F_+)-h(\omega)\right|<\varepsilon \tag{3.45}$$

holds for each bounded convex set G, for which $\varrho(G)$ is large enough. The comparison of the relations (3. 42) and (3. 41), (3. 43), (3. 44), (3. 45) yields the inequality

(3. 46)

$$\mathsf{E}\left|\frac{1}{\lambda(G)}\log f(\omega:G)-h(\omega)\right|\leqq\frac{\lambda(G^m)}{\lambda(G)\lambda(F_0)}\varepsilon+2\frac{\lambda(\bar{G}^m-\underline{G}^m)}{\lambda(G)\lambda(F_0)}\left(K+|\mathsf{H}(F_0\,|\,F_+)|\right)+\varepsilon$$

where $m\geqq m_\varepsilon$, $G\in\mathscr{B}$, $\varrho(G)$ is so large that (3. 45) is valid. The statement of the theorem follows from (3. 46) by Lemma 1.

4. Some examples

In this section we give three examples, where the postulates **A**, **B**, **C** are satisfied; that is Theorem 6. can be applied.

EXAMPLE 1. Let $(X, \mathscr{X}, \mu)$ be a σ-finite measure space. We consider the following product space: Ω is the set of all functions $\omega(u)$ defined on R_0 and taking values in X, further

$$\mathscr{A}_G=\bigtimes_{u\in G\cap R_0}\mathscr{X}_u,\quad \mathscr{A}=\mathscr{A}_R \tag{4.1}$$

where the σ-algebra $\mathscr{X}_u$, $u\in R_0$ is generated by the cylinder sets $\{\omega;\omega(u)\in U\}$ if U ranges over $\mathscr{X}$. The transformation T_u, $u\in R_0$ is the shift, defined by

$$T_u\omega=\omega' \quad \text{if} \quad \omega'(u+v)=\omega(v) \quad \text{for each} \quad v\in R_0. \tag{4.2}$$

As $T_u\mathscr{X}_0=\mathscr{X}_u$, we have a measure $\mu_u=T_u\mu_0$ on every $\mathscr{X}_u$, further

$$\Lambda_G=\bigtimes_{u\in G\cap R_0}\mu_u \tag{4.3}$$

defines a σ-finite measure on $\mathscr{A}_G$ if G is bounded.

It is easy to verify that the assumptions **A**, **B**, **C** are satisfied if P is a probability measure on $\mathscr{A}$, and P is invariant under the shifts T_u, $u\in R_0$. We see that assumption (i) of Theorem 6. becomes

$$\mathsf{H}(F_0)=\mathsf{H}_u(\mathsf{P}:\mathscr{X}_0)<+\infty. \tag{4.4}$$

In the case when X is a countable set and μ denotes the counting meausre on its subsets, then Theorem 6. reduces to the usual (L_1-convergence) form of McMillan's theorem if $r=1$. In this special case Lemma 2. implies also the a. s. convergence by the argument of L. BREIMAN [5]. See K. L. CHUNG [4], too.

The most typical example is that of an r-dimensional point process:

EXAMPLE 2. Let $(\Omega, \mathscr{A}, \mathsf{P})$ be the probability space of an r-dimensional point process, that is Ω denotes the set of all σ-finite integral valued measures $\omega(F)$, defined on $\mathscr{R}$. The σ-algebra $\mathscr{A}_G$ is generated by the events

$$A(F_1, \dots F_j; k_1, \dots, k_j) = \{\omega; \omega(F_i) = k_i \text{ for } 1 \leqq i \leqq j\}, \tag{4.5}$$

if $F_1, \dots, F_j \subset G$ and $k_1, \dots, k_j$ are nonnegative integers. The measures Λ_G are characterized by the relation

$$\Lambda_G(A(G,k)) = \frac{1}{k!}\lambda(G)^k, \tag{4.6}$$

where $\frac{1}{k!} = 1$ if $k = 0$, $\frac{1}{k!} = 0$ if $k = +\infty$, and $\lambda(G)^\circ = 1$ even if $\lambda(G) = +\infty$. The uniqueness of Λ_G follows from the requirement (3. 5), namely, for every partition $\{F_1, F_2, \dots, F_j\}$ of G into disjoint sets of $\mathscr{R}$ we obtain

$$A(F_1, \dots, F_j; k_1, \dots, k_j) = \prod_{i=1}^{j} \frac{1}{k_i!}\lambda(F_i)^{k_i}, \tag{4.7}$$

which determines Λ_G on $\mathscr{A}_G$. The existence of Λ_G and the validity of (3. 5) are proved in the paper [2] of the author. In the following example and in the physical interpretation it will be important to know that Λ_G is defined and it is σ-finite even if G is not bounded.

The transformations T_u are defined by

$$T_u\omega = \omega' \quad \text{if} \quad \omega'(F+u) = \omega(F) \text{ for each } F \in \mathscr{R}. \tag{4.8}$$

An r-dimensional point process is defined by a probability measure P on $\mathscr{A} = \mathscr{A}_R$, we assume that P is invariant under each shift T_u, then the postulates **A**, **B**, **C** are satisfied.

We remark, that in this case the assumption (i) of Theorem 6. can be replaced by

$$\mathsf{E}(\omega(F_0)) < +\infty. \tag{4.9}$$

Namely, from (2. 5) it follows

$$\mathsf{H}(G) \leqq \sum_{k=0}^{\infty} p_k \log \frac{\lambda(G)^k}{k!\, p_k}, \tag{4.10}$$

where $p_k = \mathsf{P}(A(G,k))$, however $-p_k \log p_k \leqq ke^{-k}$ if $-\log p_k \geqq k$, thus we have

$$\mathsf{H}(G) \leqq \sum_{k=0}^{\infty} (kp_k + ke^{-k}) + \log\lambda(G) \sum_{k=0}^{\infty} kp_k, \tag{4.11}$$

which proves the statement as $\sum_{k=1}^{\infty} kp_k = \mathsf{E}(\omega(G)) \leqq \mathsf{E}(\omega(F_0))$ if $F_0 \supset G$.

The above results can be extended to more general types of point processes. We treate the so called signed point processes, which can be applied in statistical physics.

EXAMPLE 3. Let us consider the measure space $(\hat{R}, \hat{\mathcal{R}}, \hat{\lambda}) = (R \times S, \mathcal{R} \times \mathcal{S}, \lambda \times \mu)$, where $(S, \mathcal{S}, \mu)$ is an arbitrary separable and σ-finite measure space. $(\Omega, \mathcal{A}, \mathsf{P})$ is the probability space of a point process in $(\hat{R}, \hat{\mathcal{R}})$, that is Ω is the set of σ-finite, integral valued measures $\omega(\tilde{F})$, defined on $\hat{\mathcal{R}}$ and

$$\mathcal{A} = \sigma\{A(F \times U; k);\ F \in \mathcal{R},\ U \in \mathcal{S},\ k \geqq 0 \text{ integer}\}, \tag{4.12}$$

further P is a probability measure on $\mathcal{A}$.

This point process can be interpreted as an r-dimensional point process, where to every point of the process there is assigned an element of S. Such a process will be called an r-dimensional signed point process. The following definitions are suggested by this interpretation.

The points of $\hat{R}$ have the form $(x, y) = z$, where $x \in R$, $y \in S$; thus each $u \in R_0$ represent a translation of R onto itself by

$$(x, y) + u = (x + u, y). \tag{4.13}$$

Consequently,

$$T_u \omega = \omega' \quad \text{if} \quad \omega'(F + u \times U) = \omega(F \times U) \quad \text{for each} \quad F \in \mathcal{R}, U \in \mathcal{S}, \tag{4.14}$$

defines an automorphism of Ω onto itself so that

$$T_{u+v} = T_u T_v \quad \text{if} \quad u, v, \in R_0, \tag{4.15}$$

and each T_u is measurable on $\mathcal{A}$. In view of **A** we assume that the shifs T_u preserve P.

If $G \in \mathcal{R}$ is an r-dimensional Borel set. then we set

$$\mathcal{A}_G = \sigma\{A(F \times U; k);\ F \subset G,\ F \in \mathcal{R},\ U \in \mathcal{S},\ k \geqq 0 \text{ integer}\}. \tag{4.16}$$

The construction of the measures Λ_G is similar to the case of Example 2., they are determined by

$$\Lambda_G\left(\prod_{i=1}^{j} A(G \times U_i; k_i)\right) = \prod_{i=1}^{j} \frac{\lambda(G)^s}{k_i!} \mu(U_i)^{k_i}, \tag{4.17}$$

where $\{U_1, \ldots, U_j\}$ is a partition of S into disjoint sets of $\mathcal{S}$, and $s = \sum_{i=1}^{j} k_i$, s may be also infinite. It can be proved that Λ_G is defined for each $G \in \mathcal{R}$, it is σ-finite and satisfies the product property (3. 5). (See J. FRITZ [2]). We see that Λ_G is not finite not even if G is bounded. The further requirements of **A**, **B**, **C** can be easily verified.

From the point of view of physical applications that special case will be important, when the sign-space S is a subset of a linear normed space. Then, one can define the additive random set-function $Y(\omega : G)$ — called the "amount of the sign" in G — by

$$Y(\omega : G) = \Sigma y_i, \tag{4.18}$$

where the sum is over those points $z_i = (x_i, y_i)$, for which $\omega(\{z_i\}) > 0$ and $x_i \in G$; that is $Y(\omega : G)$ is the sum of signs of the points of the realization ω in G.

Observe that such a signed point process can be interpreted as an additive random set-function taking values in a linear normed space. This observation will be the starting point of the following section.

5. The physical interpretation

Let us consider a one-phase, homogeneous thermodynamical system γ, the macroscopic behaviour of which is described by the extensive quantities $X_1, X_2, \ldots, X_n$. We assume that γ is in a total equilibrium. In the mathematical model the requirement that γ is a large system will be expressed by the assumption that it is infinite, thus each three-dimensional Borel set $G \in \mathscr{R}$ is associated with a subsystem γ_G of γ consisting of the particles of γ in G. The extensive quantities are defined for the subsystems of γ, $X_i(G)$ is called the amount of the extensive quantity X_i in G. In the following investigations $(R, \mathscr{R})$ will denote the three-dimensional space with the σ-algebra of its Borel subsets.

The starting point of statistical physics is the assumption that a thermodynamical system can be described by a probability space $(\Omega, \mathscr{A}, \mathsf{P})$ in the following sense: The possible realizations of the system γ — called its microstates — are represented by the elementary events $\omega \in \Omega$, while the extensive quantities are random variables, defined on the microstates: $X_i(\omega : G)$ denotes the amount of the quantity X_i in G for the microstate ω. The "extensive" property means, that $X_i(\omega : G)$ is an additive set-function if ω is fixed; i.e. it is an additive random set-function. The σ-algebra $\mathscr{A}$ is the minimal σ-algebra with respect to which every $X_i(\omega : G)$, $G \in \mathscr{R}$ is measurable. The measurable space $(\Omega, \mathscr{A})$ characterizes the structure of the system γ, while its thermodynamical state is described by a probability measure P on $\mathscr{A}$. Of course, P is not arbitrary, it is determined by certain physical conditions characterizing γ, but we do not investigate this problem here. Any subsystem γ_G of γ is described by the probability space $(\Omega, \mathscr{A}_G, \overline{\mathsf{P}}_G)$, where the σ-algebra $\mathscr{A}_G$ is generated by the variables $X_i(\omega : F)$, $F \subset G$, $F \in \mathscr{R}$, and $\overline{\mathsf{P}}_G$ denotes the restriction of P to $\mathscr{A}_G$. The macroscopic value $X_i(G)$ of the extensive quantity X_i in G is given by the expectation

$$X_i(G) = \mathsf{E}(X_i(\omega : G)). \tag{5.1}$$

It may be assumed that $X_i(G)$ is finite if G is bounded.

Taking into account the atomic structure of matter, we obtain further informations on the construction of the probability space $(\Omega, \mathscr{A}, \mathsf{P})$. In statistical physics we always have a finite system $Y_1, Y_2 \ldots Y_s$ of discrete or continuous additive quantities describing the physical state of the single particles of γ, by means of which the relevant extensive quantities $X_1, X_2, \ldots, X_n$ can be expressed. We shall denote by S_j the additive semigroup of real numbers, generated by the possible values of Y_j; $\mathscr{S}_j$ denotes the σ-algebra of Borel subsets of S_j, further, let μ_j be a σ-finite measure on $\mathscr{S}_j$. If Y_j is continuous, then S_j is the set of nonnegative numbers or the set of all real numbers and μ_j is the Lebesque-measure on $\mathscr{S}_j$. In case of a discrete quantity Y_j, the set S_j is countable, then the most convenient choice of μ_j is the counting measure; however the use of other measures is also justified, this depends on the concrete problem.

Continuous parameters are, for instance, the co-ordinates of momenta and of angular momenta, the kinetic energy, rotational energy, polarisation and other classical quantities. The electric charge, the oscillation energy and other quantum mechanical quantities are discrete. Further discrete parameters occur if γ consists of several kinds of particles. Then each type is associated with a parameter, the value of which is 1 for a particle of the corresponding type and it is 0 otherwise.

The above reasonings suggest to represent the probability space $(\Omega, \mathscr{A}, \mathsf{P})$ describing the system γ, as that of a three-dimensional signed point process with sign space $(S, \mathscr{S}, \mu) = \bigtimes_{j=1}^{s} (S_j, \mathscr{S}_j, \mu_j)$. (See Example 3.) Then the particles of the system are associated with the signed points of the process, so that any point represents the location of the corresponding particle, while its sign specifies its physical state.

The value of an extensive parameter $X_i(\omega : G)$ will be the amount of a suitable quantity in G, which quantity need not be one of the quantities $Y_1, \ldots, Y_s$, it is a function of them and of the locations of the particles, in general. Each extensive quantity is covariant.

The choice of the lattice R_0 depends on the system γ. If γ is a liquid or a gas, then R_0 can be choosen arbitrarily, while R_0 consists of the lattice points of the chrystal lattice if γ is a solid.

To see an example for this construction, we treate the simple case of a one-component, classical, monoatomic gas with an intermolecular potential energy given by the sum of pair interactions, in detail. Then the extensive parameters are the total internal energy U and the particle number N. The state of a particle is described by its momenta (Y_1, Y_2, Y_3), consequently our gas system will be described by a three-dimensional signed point process, where the sign space $(S, \mathscr{S}, \mu)$ is the three-dimensional momenta-space, μ is the Lebesgue measure. Let us consider a realisation ω of this points of process, the points of which are $x_1, x_2, \ldots, x_k, \ldots$ and the sign of the point x_k is $(Y_1(x_k), Y_2(x_k), Y_3(x_k))$. The kinetic energy $K(\omega : G)$ will be

$$K(\omega : G) = \sum_{x_k \in G} \frac{1}{2m} (Y_1(x_k)^2 + Y_2(x_k)^2 + Y_3(x_k)^2), \tag{5.2}$$

where m denotes the mass of a particle. If the interaction of the particles is described by the two-body potential $\varphi(\|x_1 - x_2\|)$, then we define the potential energy $V(\omega : G)$ by

$$V(\omega : G) = \frac{1}{2} \sum_{x_i \in G} \varphi(\|x_i - x_j\|). \tag{5.3}$$

With this definition the total internal energy $U = K + V$ will be additive. The particle number $N(\omega : G)$ is given by

$$N(\omega : G) = \omega(G \times S). \tag{5.4}$$

More complicated examples can be described similarily.

Let us consider now the general case. Since our system γ is in equilibrium, its state is invariant under the translations $u \in R_0$ therefore the shift operators T_u have to preserve the probability measure P, representing the state of γ. Of course, this is only a necessary condition of the equilibrium.

The physical requirement that the interaction of the subsystems of γ is weak, will be expressed in the mathematical theory by the assumption that the group $\mathscr{G}$ of shift operators is ergodic; then the invariant limit functions in Theorems 2. and 6. will be constant with probability one. For example, if the interaction potential has a finite range, then the subsystems of γ are P-independent if their distance is large enough, which implies the ergodicity of $\mathscr{G}$.

As the microscopic extensive parameters $X_i(\omega:G)$ are additive and covariant, and assumption (i) of Theorem 2. is a quite natural one from the point of view of statistical physics, Theorem 2. yields

$$\frac{1}{\lambda(G)} X_i(\omega:G) \xrightarrow{1} \overline{X}_i \quad \text{if} \quad G \in \mathscr{B} \quad \text{and} \quad \varrho(G) \to +\infty, \tag{5. 5}$$

where

$$\overline{X}_i = \lim \frac{X_i(G)}{\lambda(F_0)} = \frac{X_i(F_0)}{\lambda(F_0)}. \tag{5. 6}$$

As the volume $\lambda(G)$ and the extensive quantities are measured in macroscopic units, the system γ is "large" if $\lambda(G) \sim 1$; thus (5. 5) means that the value of every extensive quantity $X_i(\omega:G)$ is close to its mean value $\overline{X}_i \lambda(G)$ for the greatest part of microstates.

Finally, we treate the case of the entropy. The microscopic entropy, i.e. the entropy of a microstate ω for a subsystem γ_G is defined by

$$h(\omega:G) = -\log f(\omega:G), \tag{5. 7}$$

where $f(\omega:G)$ denotes the Radon—Nikodym derivative of $\overline{\mathsf{P}}_G$ with respect to Λ_G. This definition depends on the choice of the dominating measures Λ_G, which is not determined by the postulates (3. 5) and (3. 7), but our Λ_G-s are the most convenient among all the measures satisfying them. The macroscopic entropy will be defined as

$$\mathsf{H}(G) = \mathsf{E}\,(h(\omega:G)). \tag{5. 8}$$

From the point of view of thermodynamics, the assumptions (i) and (ii) of Theorem 6. are quite natural, thus we have

$$\frac{1}{\lambda(G)} h(\omega:G) \xrightarrow{1} \overline{H} \quad \text{if} \quad G \in \mathscr{B} \quad \text{and} \quad \varrho(G) \to \infty, \tag{5. 9}$$

where

$$\overline{H} = \frac{\mathsf{H}(F_0|F_+)}{\lambda(F_0)} = \lim \frac{\mathsf{H}(G)}{\lambda(G)}. \tag{5. 10}$$

On account of Boltzmann's principle, this result motivates the interpretation of the thermodynamical entropy as the information-theoretical entropy of a point process.

Further, more concrete applications will be treated in a forthcoming paper.

REFERENCES

[1] CSISZÁR, I.: On generalized entropy, *Studia Math. Sci. Hung.* **4** (1969), 401—415.
[2] FRITZ, J.: Entropy of point processes, *Studia Math. Sci. Hung.* **4** (1969), 389—399.
[3] PEREZ, A.: Notions generalisées d'incertitude, d'entropie et d'information du point de vue de la théorie des martingales, *Trans. First Prague Conf.* (1956), 209—243.
[4] CHUNG, K. L.: A note on the ergodic theorem of information theory, *Ann. Math. Stat.* **32** 612—614.
[5] BREIMAN, L.: The individual ergodic theorem of information theory, *Ann. Math. Stat.* **28** 809—811.
[6] FRITZ, J.: Information theory and thermodynamics of gas systems, *Proc. Coll. Information Theory, Debrecen* (1967), 167—175.
[7] VINCZE, I.: On some distribution laws of statistical physics, *MTA III. Oszt. Közl.* (1969), (in Hungarian), Read on the Europa Meeting 1968 on Statistics, Econometrics and Management Science. Amsterdam, Sept. 1968.
[8] JAYNES, T. E.: Information theory and stat. mech. *Phys. Rev.* **106.**
[9] LOÈVE, M.: *Probability theory*, Van Nostrand, 1955.
[10] HILL, T. L.: *Statistical mechanics*, McGraw-Hill, 1956.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received June 26, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 395—399.

HERMITE—FEJÉR INTERPOLATION BASED ON THE ROOTS OF JACOBI POLYNOMIALS

by

P. O. H. VÉRTESI

1. For a continuous function on the interval $[-1, 1]$ we define the uniquely determined Hermite—Fejér interpolating polynomials of degree $\leqq 2n-1$ as follows.

$$(1.1)\qquad \begin{cases} H_n^{(\alpha,\beta)}(f;x) = \sum_{k=1}^{n} f(x_{kn}^{(\alpha,\beta)})\, h_{kn}^{(\alpha,\beta)}(x) & \text{or}^* \\ \overline{H}_n^{(\alpha,\beta)}(f;x) = H_n^{(\alpha,\beta)}(f;x) + \sum_{n=1}^{n} \beta_{kn}\, \mathfrak{h}_{kn}^{(\alpha,\beta)}(x), \end{cases}$$

where $P_n^{(\alpha,\beta)}(x)$ is the n-th Jacobi polynomial defined by the well known relation

$$(1.2)\qquad (1-x)^{\alpha}(1+x)^{\beta} P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n}[(1-x)^{n+\alpha}(1+x)^{n+\beta}],$$

the roots of $P_n^{(\alpha,\beta)}(x)$ are

$$(1.3)\qquad -1 < x_{nn}^{(\alpha,\beta)} < x_{n-1,n}^{(\alpha,\beta)} < \ldots < x_{1n}^{(\alpha,\beta)} < 1,$$

$$(1.4)\qquad l_{kn}^{(\alpha,\beta)}(x) = \frac{P_n^{(\alpha,\beta)}(x)}{P_n'^{(\alpha,\beta)}(x_{kn}^{(\alpha,\beta)})(x - x_{kn}^{(\alpha,\beta)})},$$

$$(1.5)\qquad h_{kn}^{(\alpha,\beta)}(x) = \left[1 - \frac{P_n''^{(\alpha,\beta)}(x_{kn}^{(\alpha,\beta)})}{P_n'^{(\alpha,\beta)}(x_{kn}^{(\alpha,\beta)})}(x - x_{kn}^{(\alpha,\beta)})\right][l_{kn}^{(\alpha,\beta)}(x)]^2 = v_{kn}^{(\alpha,\beta)}(x)[l_{kn}^{(\alpha,\beta)}(x)]^2,$$

$$(1.6)\qquad \mathfrak{h}_{kn}^{(\alpha,\beta)}(x) = (x - x_{kn}^{(\alpha,\beta)})[l_{kn}^{(\alpha,\beta)}(x)]^2,$$

β_{kn} are prescribed (s. [1], 14. 1.).

It is well known (s. [1], 14. 1.) that

$$(1.7)\qquad \sum_{k=1}^{n} h_{kn}^{(\alpha,\beta)}(x) = 1 \qquad (n = 1, 2, \ldots; x \in [-1, 1]),$$

$$(1.8)\qquad H_n^{(\alpha,\beta)}(f; x_{kn}^{(\alpha,\beta)}) = \overline{H}_n^{(\alpha,\beta)}(f; x_{kn}^{(\alpha,\beta)}) = f(x_{kn}^{(\alpha,\beta)}) \qquad (k = 1, 2, \ldots, n),$$

$$(1.9)\qquad H_n'^{(\alpha,\beta)}(f; x_{kn}^{(\alpha,\beta)}) = 0, \quad \overline{H}'^{(\alpha,\beta)}_n(f; x_{kn}^{(\alpha,\beta)}) = \beta_{kn} \qquad (k = 1, 2, \ldots, n).$$

* Throughout this paper let $\alpha, \beta > -1$.

The following theorem is known (s. [1], Theorem 1. 4. 6.).

THEOREM 1. 1. *Let $f(x)$ be continuous on $[-1, 1]$. The sequence $H_n^{(\alpha,\beta)}(f; x)$ uniformly converges to $f(x)$ in $[-1+\varepsilon, 1-\varepsilon]$. If $\alpha<0$ then the uniform convergence holds in $[-1+\varepsilon, 1]$. Further, if $\alpha\geqq 0$, then there exists a continuous function such that*

$$\overline{\lim_{n\to\infty}}\,|H_n^{(\alpha,\beta)}(f; 1)-f(1)|>0. \tag{1. 10}$$

2. The purpose of this paper is to give some further estimations for the difference $H_n^{(\alpha,\beta)}(f; x)-f(x)$.

We can prove the following theorem.

THEOREM 2. 1. *Let $f(x)$ be a continuous function on $[-1, 1]$ and $\omega(f; t)=$ $=O[\omega_1(t)]$ where $\omega(f; t)$ is the modulus of continuity of $f(x)$, $\omega_1(t)$ is a modulus of continuity. Then we have for an arbitrary subinterval $[a, b]\subset(-1, 1)$ that*

$$|f(x)-H_n^{(\alpha,\beta)}(f; x)| = O(1)\sum_{i=1}^{n}\omega_1\left(\frac{i}{n}\right)\frac{1}{i^2} \qquad (x\in[a,b]\subset(-1,1)) \tag{2. 1}$$

for arbitrary α and β. (Here the sign O depends on α, β, a and b.)

2. 1. We prove the convergence from (2. 1). Indeed, we have

$$\begin{aligned}|f(x)-H_n^{(\alpha,\beta)}(f; x)|) &= O(1)\,\omega_1\left(\frac{\log n}{n}\right)\sum_{i=1}^{n}\left(\frac{i}{\log n}+1\right)\frac{1}{i^2} = \\ &= O(1)\,\omega_1\left(\frac{\log n}{n}\right)(x\in[a,b]).\end{aligned} \tag{2. 2}$$

Further, if $\omega_1(t)=t^\varrho$ $(0<\varrho\leqq 1)$, we have

$$|f(x)-H_n^{(\alpha,\beta)}(f; x)| = \begin{cases} O(n^{-\varrho}) & \text{for} \quad x\in[a,b], \quad 0<\varrho<1, \\ O\left(\dfrac{\log n}{n}\right) & \text{for} \quad x\in[a,b], \quad \varrho=1. \end{cases} \tag{2. 3}$$

These results are similar to the estimations in [2], where we investigated the case $\alpha=\beta=-\frac{1}{2}$, but there we were able to prove the estimations for the whole interval $[-1, 1]$.

2. 2. PROOF of Theorem 2. 1. We shall use in this and the following parts some formulae of [1]. We have

$$\vartheta_{kn}^{(\alpha,\beta)} = \frac{1}{n}[k\pi+O(1)] \qquad (k=1,2,\ldots,n; \quad n=1,2,\ldots)^* \tag{2. 4}$$

$$P_n'^{(\alpha,\beta)}(\cos\vartheta_k)\sim k^{-\alpha-3/2}n^{\alpha+2} \quad \left(0<\vartheta_k\leqq\frac{\pi}{2}, \quad n=1,2,\ldots\right)^{**}, \tag{2. 5}$$

---

* $x_{kn}^{(\alpha,\beta)}=\cos\vartheta_{kn}^{(\alpha,\beta)}$, $x=\cos\vartheta$; sometimes we denote $x_{kn}^{(\alpha,\beta)}$ by x_k, etc. Here O depends only on α and β.

** The notation $Z_n\sim W_n$ means that $c_1\leqq\dfrac{|Z_n|}{|W_n|}\leqq c_2$ $(n\geqq N)$ where $0<c_1\leqq c_2<\infty$, $W_n\neq 0$. In (2. 5) c_1 and c_2 depend only on α and β.

$$\max_{0\leqq x\leqq 1}|P_n^{(\alpha,\beta)}(x)| = \begin{cases} \binom{n+\alpha}{\alpha} = O(n^\alpha) & \text{for} \quad \alpha \geqq -\frac{1}{2}, \\ O(n^{-1/2}) & \text{for} \quad \alpha \leqq -\frac{1}{2}, \end{cases} \tag{2.6}$$

$$P_n^{(\alpha,\beta)}(x) = (-1)^n P_n^{(\beta,\alpha)}(-x) \tag{2.7}$$

(s. (14. 5. 3), (8. 9. 1), (8. 9. 2), (7. 32. 2) and (4. 1. 3) in [1]).

Let us denote by $x_{jn}^{(\alpha,\beta)}$ the nearest roots to x. Obviously, $j=j(n)$. Using (2. 4), as in the paper [4], we can prove that

$$|f(x)-f(x_k)| = \begin{cases} O(1)\left[\omega_1\left(\frac{\sin\vartheta}{n}\right)+\omega_1\left(\frac{1}{n^2}\right)\right] & \text{if} \quad k=j \\ O(1)\left[\omega_1\left(\frac{i\sin\vartheta}{n}\right)+\omega_1\left(\frac{i^2}{n^2}\right)\right] & \text{if} \quad j<k=j+i\leqq n \\ & \text{or} \quad 1\leqq k=j-i<j. \end{cases} \tag{2.8}$$

We have

$$v_{kn}^{(\alpha,\beta)}(x) = \frac{1-x[\alpha-\beta+(\alpha+\beta+2)x_{kn}]+(\alpha-\beta)x_{kn}+(\alpha+\beta+1)x_{kn}^2}{1-x_{kn}^2} \tag{2.9}$$

(s. [1], (14. 5. 2)).

By (2. 9) we can see that for the interval $[a, b]\subset(-1, 1)$ there exist the positive numbers $\xi(\alpha, \beta)$, $m(\alpha, \beta)$ and $M(\alpha, \beta)$ such that

(2. 10)

$$0<m(\alpha,\beta)\leqq v_{kn}^{(\alpha,\beta)}(x)\leqq M(\alpha,\beta) \quad \text{for} \quad |x-x_{kn}|\leqq\xi(\alpha,\beta),\ x\in[a,b]\subset(-1,1).$$

Further, by (2. 4) we have

$$1-x_{kn}^2 \geqq \begin{cases} c_1(\alpha,\beta)\dfrac{k^2}{n^2} & (0\leqq x_{kn}<1), \\ c_2(\alpha,\beta)\dfrac{(n-k)^2}{n^2} & (-1<x_{kn}\leqq 0). \end{cases} \tag{2.11}$$

Now we prove our statement. We can suppose that $-1<a<0\leqq x\leqq b<1$ and $b>0$. By (1. 1) and (1. 7) we have

$$|f(x)-H_n(f;x)| = \left|\sum_{k=1}^{n}[f(x)-f(x_k)]h_k(x)\right| = O(1)\Big[|f(x)-f(x_j)|h_j(x)+$$

$$+\sum_{\substack{|x-x_k|\leqq\xi(\alpha,\beta)\\ k\neq j}}|f(x)-f(x_k)|h_k(x)+\sum_{x-x_k>\xi(\alpha,\beta)}|f(x)-f(x_k)\|h_k(x)|+$$

$$+\sum_{x_k-x>\xi(\alpha,\beta)}|f(x)-f(x_k)\|h_k(x)|\Big] \equiv I_1+\sum_{k\in I_2}+\sum_{k\in I_3}+\sum_{k\in I_4}.$$

Let us estimate the parts.

If $x=x_j$ then $H_n(f;x_j)=f(x_j)$, i.e., by all means

$$I_1 = O(1)\omega_1\left(\frac{1}{n}\right). \tag{2.12}$$

Otherwise we have to use the formulae

$$\frac{d}{dx}\{P_n^{(\alpha,\beta)}(x)\} = \frac{1}{2}(n+\alpha+\beta+1)P_{n-1}^{(\alpha+1,\beta+1)}(x), \tag{2.13}$$

$$|P_n^{(\alpha,\beta)}(x)| = O(n^{-1/2}) \quad \text{for} \quad x\in[a,b]\subset(-1,1) \tag{2.14}$$

(s. [1], (4. 21. 7) and (7. 32. 5)).

We have

$$l_{jn}^{(\alpha,\beta)}(x) = \frac{P_n^{(\alpha,\beta)}(x)-P_n^{(\alpha,\beta)}(x_j)}{P_n'^{(\alpha,\beta)}(x_j)(x-x_j)} = \frac{P_n'^{(\alpha,\beta)}(x^*)}{P_n'^{(\alpha,\beta)}(x_j)} \quad (x\overset{<}{(>)}x^*\overset{<}{(>)}x_j). \tag{2.15}$$

By (1. 5), (2. 10), (2. 15), (2. 13), (2. 14) and (2. 5)

$$I_1 = O(1)\omega_1\left(\frac{1}{n}\right)l_j^2(x) = O(1)\omega_1\left(\frac{1}{n}\right)(nn^{-1/2}n^{\alpha+3/2}n^{-\alpha-2})^2 = O(1)\omega_1\left(\frac{1}{n}\right)^*,$$

as above.

(We used that now $j\sim n$.)

Further, by (2. 5) and (2. 7) we have

$$P_n'(\cos\vartheta_k)\sim n^{1/2} \quad \text{if} \quad \begin{cases} k\sim n & \text{for} \quad x_k\geqq 0, \\ n-k\sim n & \text{for} \quad x_k\geqq 0. \end{cases} \tag{2.16}$$

We have the elementary estimation

$$(\cos\vartheta-\cos\vartheta_k)^{-2} = \left(2\sin\frac{\vartheta+\vartheta_k}{2}\sin\frac{\vartheta-\vartheta_k}{2}\right)^{-2} = O\left(\frac{n^2}{i^2}\right); \quad k=j\pm i,\ k\neq j. \tag{2.17}$$

(s. (2. 4)).

So by (2. 8), (2. 10), (2. 16), (2. 14) and (2. 17) we get

$$\sum_{k\in I_2}|f(x)-f(x_k)|h_k(x) = O(1)\sum_{k\in I_2}\omega_1\left(\frac{i}{n}\right)n^{-1}n^{-1}(n^2i^{-2}) = O(1)\sum_{i=1}^{n}\omega_1\left(\frac{i}{n}\right)\frac{1}{i^2}. \tag{2.18}$$

Further, by (2. 8), (2. 11), (2. 14) and (2. 5) we obtain

$$\sum_{k\in I_3}|f(x)-f(x_k)||h_k(x)| = O(1)\sum_{k\in I_3}\omega_1\left(\frac{i}{n}\right)(n^2k^{-2})n^{-1}(k^{2\alpha+3}n^{-2\alpha-4})\xi^{-2} =$$

$$= O(n^{-2\alpha-3})\sum_{k\in I_3}\omega_1\left(\frac{i}{n}\right)k^{2\alpha+1} = O(n^{-2\alpha-3})\sum_{k=1}^{n}k^{2\alpha+1} = O\left(\frac{1}{n}\right). \tag{2.19}$$

* If $x<0$, ve replace α by β.

Similarly, applying (2. 7) we have

$$\sum_{k \in I_4} |f(x)-f(x_k)|\,|h_k(x)| = O\left(\frac{1}{n}\right). \tag{2.20}$$

By (2. 12) and (2. 18)—(2. 20) we get

$$|f(x)-H_n(f;x)| = O(1)\left[\sum_{i=1}^{n} \omega_1\left(\frac{i}{n}\right)\frac{1}{i^2}+\frac{1}{n}\right] = O(1)\sum_{i=1}^{n} \omega_1\left(\frac{i}{n}\right)\frac{1}{i^2}.$$

Q.e.d.

3. Finally we mention the following theorem.

THEOREM 3. 1. *If $\omega_1(t)$ is a modulus of continuity such that $\lim_{t=0} \omega_1(t)t^{-1} = +\infty$ further $\alpha > 0$ then there exists a continuous function $f(x)$ for which $\omega(f;t) = O[\omega_1(t)]$ and*

$$|f(1)-H_n^{(\alpha,\beta)}(f;1)| > n^{2\alpha}\omega_1\left(\frac{1}{n^2}\right) \quad (n = n_1, n_2, \ldots; \alpha > 0).$$

Indeed, we have

$$\sum_{k=1}^{n} |h_{kn}^{(\alpha,\beta)}(1)| \geqq c(\alpha,\beta)n^{2\alpha}$$

(see [1], (14. 6. 15)). Applying Theorem 3. 1 of [3], we obtain our statement.

Note. We can obtain similar results for $\bar{H}_n^{(\alpha,\beta)}(f;x)$ as well.

REFERENCES

[1] SZEGŐ, G.: *Orthogonal polynomials,* Amer. Math. Soc. New York, 1959.
[2] VÉRTESI, P. O. H.: On the convergence of Hermite—Fejér interpolation, *Acta Math. Acad. Sci. Hung.*, **22** (1971),
[3] KIS, O., VÉRTESI, P. O. H.: On certain linear operators, ibid, **22** (1971),
[4] О. Киш (O. Kis): Замечания о порядке сходимости Лагранжева интерполирования, *Annales Univ. Sci. Budapest,* **11** (1968), 27—40.

Eötvös L. University, Budapest

(Received July 12, 1969, in revised form October 2, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 401—410.

LOWER ESTIMATIONS FOR SOME INTERPOLATING PROCESSES

by

P. O. H. VÉRTESI

1. Before proceeding to the consideration of our object we wish to have a short survey of our tools.

At first we give the following definitions (see [1]).

Let us denote by $\omega_m(t)$ a function with the following properties.

(i) $\omega_m(t)>0$ for $t>0$, $\omega_m(0)=0$,

$\omega_m(T)\geqq\omega_m(t)$ if $T\geqq t$, $\omega_m(t)$ is continuous function for $t\geqq 0$,

(ii) $\dfrac{t^m}{\omega_m(t)}$ is monotone increasing function for $t>0$,

(iii) $\lim\limits_{t=+0}\dfrac{t^m}{\omega_m(t)}=0$ $\quad(m\geqq 1$ is a fixed integer number)*.

Let $[a, b]$ be an arbitrary finite interval,

$$(1.1)\qquad \begin{cases} b\geqq x_{0n}>x_{1n}>\ldots>x_{pn}\geqq a \quad (n=1,2,3,\ldots), \text{ where} \\ p=p(n) \quad \text{and} \quad \overline{\lim\limits_{n=\infty}}\, p(n)=\infty. \end{cases}$$

Let us denote by $C^{[a,b]}(\omega_m)$ the class of all continuous functions on $[a, b]$ for which

$$(1.2)\qquad \omega_m(f;t)\leqq a_m(f)\omega_m(t)$$

holds. Here $\omega_m(f;t)$ is the modulus of smoothness of order m of $f(x)$, $a_m(f)$ depends only on $f(x)$, $\omega_m(t)$ is defined by (i), (ii) and (iii).

Let

$$(1.3)\qquad l_{kn}(x) \qquad (n=1,2,\ldots;\ k=0,1,\ldots,p)$$

be continuous functions on $[a, b]$,

$$(1.4)\qquad L_n(f;x)=\sum_{k=0}^{p} f(x_{kn})\,l_{kn}(x),$$

$$(1.5)\qquad \lambda_n(x)=\sum_{k=0}^{p}|l_{kn}(x)|;\quad \lambda_n=\max_{a\leqq x\leqq b}\lambda_n(x),$$

$$(1.6)\qquad d_n=\min_{0\leqq k\leqq p-1}(x_{kn}-x_{k+1,n}).$$

* E. g., if $\omega_m(t)=t^\alpha(0<\alpha<m)$, (i), (ii) and (iii) are fulfilled.

We have proved (see [1], Theorem 3. 1.)

THEOREM 1. 1. *Let x^* be an arbitrary point in the interval $[a, b]$, further let $0 < n^{(1)} < n^{(2)} < n^{(3)} < \dots$ be integer numbers. If $\overline{\lim}_{n=n^{(1)}, n^{(2)}, \dots} \lambda_n(x^*) > 1$ or $\underline{\lim}_{n=n^{(1)}, n^{(2)}, \dots} \lambda_n(x^*) < 1$ then there exists an $f(x) \in C^{[a,b]}(\omega_m)$ such that*

(1. 7.)

$$|L_n(f; x^*) - f(x^*)| > \lambda_n(x^*)\,\omega_m(d_n) \quad (n = n_1, n_2, n_3, \dots, \text{ where } \{n_i\}_{i=1}^{\infty} \subset \{n^{(j)}\}_{j=1}^{\infty}).$$

We shall use a stronger statement in this paper (see [1], Theorem 3. 2).

Let us denote by $I_n = \{x_{r(n),n}; x_{s(n),n}\}$ $(p(n) \geqq r(n) > s(n) \geqq 0)$ general intervals, further let $[I_n] = [x_{rn}, x_{sn}]$ and $(I_n) = (x_{rn}, x_{sn})$, $I_{n+1} \subseteq I_n$.

Let

$$d_n(I_n) = \min_{x_{kn}, x_{k+1,n} \in [I_n]} (x_{kn} - x_{k+1,n}) \tag{1. 8}$$

such that

$$\lim_{n \to \infty} d_n(I_n) = 0 \tag{1. 9}$$

and

$$\lambda_n(I_n; x) = \sum_{x_{kn} \in (I_n)} |l_{kn}(x)|. \tag{1. 10}$$

We have proved

THEOREM 1. 2. *If $x^* \in [a, b]$ and I_n are such that*

$$\overline{\lim}_{n=n^{(1)}, n^{(2)}, \dots} \lambda_n(I_n; x^*) > 1 \quad \text{or} \quad \underline{\lim}_{n=n^{(1)}, n^{(2)}, \dots} \lambda_n(I_n; x^*) < 1 \tag{1. 11}$$

then there exists an $f(x) \in C^{[a,b]}(\omega_m)$ for which

$$|f(x^*) - L_n(f; x^*)| > \lambda_n(I_n; x^*)\,\omega_m\big(d_n(I_n)\big) \quad \big(n = n_1, n_2, n_3, \dots, \text{ where } \{n_i\}_{i=1}^{\infty} \subset \{n^{(j)}\}_{j=1}^{\infty}\big). \tag{1. 12}$$

Our aim is to give some applications of these theorems for some interpolating processes.

2. At first let us consider the nodes

$$x_{kn} = \cos \vartheta_{kn} = \cos \frac{k\pi}{n} \quad (k = 0, 1, \dots, n;\ n = 1, 2, \dots) \tag{2. 1}$$

on the interval $[-1, 1]$.

Further let

$$l_{kn}(x),\ L_n(f; x) = \sum_{k=0}^{n} l_{kn}(x) f(x_{kn}),\ \lambda_n(x) = \sum_{k=0}^{n} |l_{kn}(x)|,\ \lambda_n = \max_{-1 \leqq x \leqq 1} \lambda_n(x) \tag{2. 2}$$

be the fundamental polynomials of degree n of the Lagrange interpolation based on the nodes (2. 1), the Lagrange interpolating polynomials for $f(x)$, the Lebesgue functions and the Lebesgue constants of the interpolation, respectively.

By the definitions (2. 1) and (2. 2) the following theorem holds.

THEOREM 2. 1. *If x^* is an arbitrary point of $(-1, 1)$, then there exist an $f(x)\in C^{[-1,1]}(\omega_m)$ and a sequence $0<n_1<n_2<\dots$ such that*

$$|L_n(f; x^*)-f(x^*)| > \log n\,\omega_m\left(\frac{1}{n}\right) \qquad (n=n_1, n_3, \dots; x^*\neq\pm 1). \tag{2. 3}$$

3. To prove (2. 3) we need some lemmas.

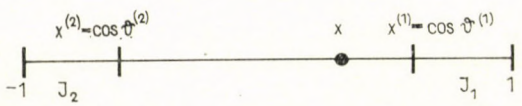
We have the following estimation (see e.g. [2], 3. 10)

$$\lambda_n(x)=\frac{2\sin\alpha_n\pi}{\pi}\log n+O(1). \tag{3. 1}$$

Here $x=\cos\vartheta$ and for $j=j(n)$

$$\vartheta_{jn}\leqq\vartheta\leqq\vartheta_{j+1,n}, \quad \vartheta=\frac{(j+\alpha_n)\pi}{n}. \tag{3. 2}$$

In connection with (3. 1) we notice the following fact. Let $x\in(-1, 1)$, then we may choose the intervals $I=(x^{(2)}, x^{(1)})$ as follows:



Here (with $J_1=[x^{(1)}, 1]$ and $J_2=[-1, x^{(2)}]$) $|J_1|=|J_2|$.* If I_n are the largest intervals** such that

$$[I_n]\subseteq[x^{(2)}, x^{(1)}], \quad [I_n]\subseteq[I_{n-1}] \qquad (n=1, 2, \dots) \tag{3. 3}$$

then

$$\lambda_n(I_n; x)=\frac{2\sin\alpha_n\pi}{\pi}\log n+o(1) \quad \text{for} \quad n=1, 2, \dots. \tag{3. 4}$$

Indeed, we have (see [2], (3. 10))

$$\lambda_n(x)=\frac{1}{2n}\sum_{k=-n}^{n}\left|\frac{\sin(2n+1)\dfrac{\vartheta-\vartheta_{kn}}{2}}{\sin\dfrac{\vartheta-\vartheta_{kn}}{2}}\right|+O(1).$$

At first let $x_{kn}\in J_1$. Then $\left(\text{supposing e.g. that } 0\leqq x<1, \text{ i.e., } 0<\vartheta\leqq\frac{\pi}{2}\right)$

$$\left.\begin{aligned} 0<\frac{\vartheta-\vartheta^{(1)}}{2}\leqq\frac{\vartheta-\vartheta_{kn}}{2}<\frac{\vartheta}{2}\leqq\frac{\pi}{4},\\ \frac{\pi}{2}>\frac{\vartheta+\vartheta^{(1)}}{2}\geqq\frac{\vartheta+\vartheta_{kn}}{2}>\frac{\vartheta}{2}\end{aligned}\right\} \text{ for } x_{kn}\in J_1.$$

* $|E|$ is the Lebesgue-measure of the set E.

** As for the definition of I_n, see Part 1.

If $x_{kn} \in J_2$, then

$$\left.\begin{aligned} &\frac{\pi}{2} > \frac{\vartheta_{kn} - \vartheta}{2} > \frac{\vartheta^{(2)} - \vartheta}{2} > 0, \\ &\pi > \vartheta^{(2)} \geqq \pi - \frac{\vartheta + \vartheta_{kn}}{2} > \frac{\pi}{2} - \frac{\vartheta}{2} > \frac{\pi}{4} \end{aligned}\right\} \text{ for } x_{kn} \in J_2.$$

By these relations we have

$$\min_{x_{kn} \in J_1 \cup J_2} \left(\sin \frac{|\vartheta - \vartheta_{kn}|}{2}, \sin \frac{|\vartheta + \vartheta_{kn}|}{2} \right) \geqq \sin \delta > 0 \qquad (n = 1, 2, \ldots).$$

I.e., we obtain

$$\lambda_n(x) = \frac{1}{2n} \sum_{x_{kn} \in J_1 \cup J_2} (\ldots) + \frac{1}{2n} \sum_{x_{kn} \in I} (\ldots) + O(1) =$$

$$= \frac{1}{2n} \sum_{k=-n}^{n} \frac{1}{\sin \delta} + \lambda_n(I_n; x) + O(1) = \lambda_n(I_n; x) + O(1),$$

as it was stated.

Here we mention the well-known important relations

$$\text{(3. 5)} \qquad \begin{cases} d_n \geqq \dfrac{c}{n^2} & (n = 1, 2, \ldots; \ c > 0) \\ d_n(I_n) \leqq \dfrac{c_1}{n} \quad \text{for} \quad n \geqq N(x) \quad \text{and} \quad c_1 = c_1(x) > 0. \end{cases}$$

Now we prove

LEMMA 3. 1. *For an arbitrary* $x^* \in (-1, 1)$ *there exists a number* p *such that*

$$\text{(3. 6)} \qquad x^* \notin \Delta = \{x_{kn}\} \qquad (k = 0, 1, \ldots, n; \ n = p, p^2, \ldots).$$

(I.e., there exists a node-system containing infinitely many nodes for which x^* is not a node.)

PROOF. Indeed, if $\dfrac{\vartheta^*}{\pi} = t^*$ is irrational, then we may choose $p = 2$. On the other hand, if $t^* = \dfrac{r}{s}$, where $(r, s) = 1$, then let p be the least prime number such that $(p, s) = 1$. It is easy to verify that this p is a suitable number for (3. 6)*. Q.e.d.

We prove still the following

LEMMA 3. 2. *For an arbitrary* $x^* \in (-1, 1)$ *there exists a sequence* $0 < n^{(1)} < n^{(2)} < \ldots$ *such that*

$$\text{(3. 7)} \qquad \sin \alpha_n^* \pi \geqq \sin \frac{\pi}{2} \frac{1}{p} \geqq \frac{1}{p} \qquad (n = n^{(1)}, n^{(2)}, \ldots),$$

where α^* *is defined by* (3. 2).

*If for an arbitrary $x_{kn} \in \Delta$, $x^* = x_{kn}$, then $\left(\text{with } \dfrac{\vartheta_{kn}}{\pi} = t_{kn} = \dfrac{l}{p^i}, (l, p^i) = 1 \right) \dfrac{r}{s} = \dfrac{l}{p^i}$ or $r \cdot p^i = l \cdot s$. But partly $p / (r \cdot p^i)$, partly $(p, l) = (p, s) = 1$, which is a contradiction.

PROOF. Let us suppose that $x_{kn} \in \Delta$ (Δ is defined in Lemma 3. 1.). Then evidently

$$0 < \alpha_n^* < 1 \quad (n = p, p^2, p^3, \ldots). \tag{3. 8}$$

At first let $n = p$ and e.g. $0 < \alpha_p^* \leqq \frac{1}{2}$.*

If $\alpha_p^* \geqq \frac{1}{2p}$, then we may choose $n^{(1)} = p$. Otherwise, using the relation $\alpha_{p^{s+1}}^* = p\alpha_{p^s}^*$ $\left(\text{for } 0 < \alpha_{p^s}^* < \frac{1}{2p}\right)$ we can find the positive integer exponents i and $i+1$ such that $\alpha_{p^i}^* < \frac{1}{2p}$, but $\frac{1}{2} > \alpha_{p^{i+1}}^* \geqq \frac{1}{2p}$. Now we may choose $n^{(1)} = p^{i+1}$. For $n = p^{i+1}$ we have to recently find the nodes $\vartheta_{j,p^{i+1}}$ and $\vartheta_{j+1,p^{i+1}}$ such that $\vartheta_{jp^{i+1}} \leqq \vartheta^* \leqq \vartheta_{j+1,p^{i+1}}$, then we can follow our procedure in analogous way. If for an arbitrary $n^{(r)}$ $\frac{1}{2} < \alpha_{n^{(r)}}^* < 1$, then we can choose the number $n^{(r+1)}$ such that $\frac{1}{2} > 1 - \alpha_{n^{(r+1)}}^* \geqq \frac{1}{2p}$. In both cases we obtain (3. 7).

Q.e.d.

4. PROOF of Theorem 2. 1. Now we can prove our statement in a few lines. Indeed, let us choose the sequence $n^{(1)} < n^{(2)} < \ldots$ as in the Lemma 3. Without loss of generality we can suppose that $n^{(1)} \geqq N(x^*)$ (see (3. 5)).

Then by (1. 12), (3. 1), (3. 4), (3. 7) and (3. 5) there exists an $\tilde{f}(x) \in C^{[-1,1]}(\omega_m)$ such that

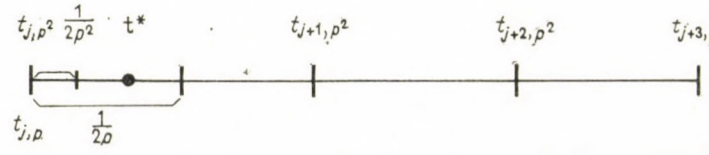
$$|L_n(\tilde{f}; x^*) - \tilde{f}(x^*)| > c_1 \log n \cdot \omega_m\left(\frac{c}{n}\right) > c_2 \log n \cdot \omega_m\left(\frac{1}{n}\right) (c_2 > 0).^{**}$$

If we consider the function $f(x) = \frac{1}{c_2}\tilde{f}(x)$, we obtain (2. 3). Q.e.d.

5. In this Part we intend to sketch a similar result for the Chebyshev-nodes

$$x_{kn} = \cos\frac{2k-1}{2n+2}\pi \qquad (k = 1, 2, \ldots, n+1; n = 1, 2, \ldots). \tag{5. 1}$$

*Let $p = 3$, $i = 1$ (i.e., $n^{(1)} = p^2$).



Here $\alpha_p^* \approx \frac{1}{9} < \frac{1}{2p} = \frac{1}{6}$ and $\alpha_{p^2}^* \approx \frac{3}{3} > \frac{1}{2p} = \frac{1}{6}$.

**If $c \geqq 1$ then by (i) $\omega_m\left(\frac{c}{n}\right) \leqq \omega_m\left(\frac{1}{n}\right)$. For $0 < c < 1$ by (ii) $\omega_m\left(\frac{c}{n}\right) \leqq c^m \omega_m\left(\frac{1}{n}\right)$.

By analogous notations to (2. 2) we have

THEOREM 5. 1. *If x^* is an arbitrary point of the interval $[-1, 1]$ then there exist an $f(x)\in C^{[-1,1]}(\omega_m)$ and a sequence $0<n_1<n_2<\dots$ (depending on x^*) such that*

$$(5.2)\qquad \begin{cases} |L_n(f;x^*)-f(x^*)| > \log n\cdot\omega_m\left(\dfrac{1}{n}\right) & (n=n_1,n_2,\dots;\ x^*\neq\pm 1),\\ |L_n(f;\pm 1)-f(\pm 1)| > \log n\cdot\omega_m\left(\dfrac{1}{n^2}\right) & (n=n_1,n_2,\dots). \end{cases}$$

PROOF. We have (see e.g. [2], 3. 8 and 3. 9)

$$(5.3)\qquad \begin{cases} \lambda_n(\pm 1)\geqq c\cdot\log n,\\ \lambda_n(x)=\dfrac{2\sin\alpha_n\pi}{\pi}\log n+O(1), \end{cases}$$

where for $j=j(n)$

$$(5.4)\qquad \vartheta_{jn}\leqq\vartheta\leqq\vartheta_{j+1,n},\quad \vartheta=\frac{2j-1+2\alpha_n}{2n+2}\pi.$$

Further, we can prove a lemma like Lemma 3. 2 choosing

$$(5.5)\qquad \begin{cases} \Delta=\{x_{kn}\},\ k=1,2,\dots,n+1;\ n=2s,2^2s,\dots \text{ for } t^*=\dfrac{r}{s},\ (r,s)=1, \text{ or}\\ \Delta=\{x_{kn}\},\ k=1,2,\dots,n+1;\ n=3,3^2,\dots \text{ for an irrational } t^*. \end{cases}$$

By (5. 3)—(5. 5), (3. 5), Theorem 1. 2. and Theorem 1. 1. we get (5. 2). Q.e.d.

6. We obtain analogous results for trigonometric interpolation. E.g., if

$$(6.1)\qquad x_{km}=\frac{2k+1}{2n+1}\pi\qquad (k=0,1,\dots,2n;\ n=1,2,\dots),$$

then we get

THEOREM 6. 1. *For an arbitrary $x^*\neq(2l+1)\pi$ ($l=0,\pm1,\pm2,\dots$) there exist a $g(x)\in\tilde{C}(\omega_m)$ and a sequence $0<n_1<n_2<\dots$ such that*

$$(6.2)\qquad |L_n(g;x^*)-g(x^*)| > \log n\cdot\omega_m\left(\frac{1}{n}\right)\qquad (n=n_1,n_2,\dots;\ x^*\neq(2l+1)\pi).$$

Here $\tilde{C}(\omega_m)=\{g(x);\ g(x)$ is 2π-periodic, continuous, and satisfies (1. 2)$\}$; $L_n(g;x)$ are the trigonometric interpolating polynomials of degree $\leqq n$ for the nodes (6. 1); Theorem 1. 1. can be applied for $\tilde{C}(\omega_m)$ supposing that $l_{kn}(x)$ are 2π-periodic as well, $\{a,b\}=[0,2\pi\}$ and $d_n=\min\limits_{0\leqq k\leqq p}(x_{kn}-x_{k+1,n})$ $(x_{p+1,n}\equiv x_{pn})$, respectively (see [1], Notes).

Further, we have

$$\lambda_n(x)=\frac{2\sin\alpha_n\pi}{\pi}\log n+O(1),$$

where for $j=j(n)$

$$x_{jn} \leqq x \leqq x_{j+1,n}, \quad x = x_{jn} + \frac{2\alpha_n \pi}{2n+1}$$

(see [2], 3. 6). Let

$$\Delta = \{x_{kn}\}, k = 0, 1, \ldots, 2n; 2n+1 = 3, 3^2, \ldots, \quad \text{if} \quad \frac{x^*}{\pi} \text{ is irrational,}$$

$$\Delta = \{x_{kn}\}, k = 0, 1, \ldots, 2n; 2n+1 = p, p^2, \ldots \text{ for } \frac{x^*}{\pi} = \frac{r}{s} \text{ and } (r,s) = (s,p) = 1,$$

then we can obtain (6. 2) for every x^*, because $d_n = \dfrac{2\pi}{2n+1}$.

7. In this part we give analogous statement for the interpolations based on the roots of the Jacobi-polynomials

$$(7.1) \quad P_n^{(\alpha,\beta)}(x) = (1-x)^{-\alpha}(1+x)^{-\beta}\frac{(-1)^n}{2^n n!}\frac{d^n}{dx^n}\{(1-x)^{\alpha+n}(1+x)^{\beta+n}\} \quad (\alpha, \beta > -1)$$

of degree n.

By powerful tools we can prove the estimations

$$(7.2) \quad \lambda_n^{(\alpha,\beta)}(I_n; x) = c_1^{(\alpha,\beta)}(x)\log n, \text{ where } -1 < x < 1, c_1(x) > 0, n = n^{(1)}, n^{(2)}, \ldots$$

for I_n defined as in Part 3. Here $0 < n^{(1)} < n^{(2)} < \ldots$ is a suitable sequence (see [3], (14. 4. 7)). It is well-known that (3. 5) is fulfilled in our case. (See [3], (8. 9. 1).) By these relations we can easily obtain the following

THEOREM 7. 1. *If* $-1 < x^* < 1$ *then there exist an* $f(x) \in C^{[-1,1]}(\omega_m)$ *and a sequence* $0 < n_1 < n_2 < \ldots$ *such that*

$$(7.3) \quad |L_n^{(\alpha,\beta)}(f; x^*) - f(x^*)| > \log n \cdot \omega_m\left(\frac{1}{n}\right) \qquad (n = n_1, n_2, \ldots; x^* \neq \pm 1).$$

Here $L_n^{(\alpha,\beta)}(f; x)$ are the nth Lagrange interpolatory polynomials based on the roots of $P_n^{(\alpha,\beta)}(x)$ $(\alpha, \beta > -1)$.

8. To obtain the estimations for $x^* = \pm 1$, we shall use two methods. At first we wish to apply Theorem 1. 2. For the sake of simplicity we suppose that $x^* = 1$. We have (see [3], (4. 1. 1) and (8. 9. 2))

$$(8.1) \qquad P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{\alpha} \geqq c_1 n^\alpha \qquad (n = 1, 2, \ldots; \quad \alpha, \beta > -1),$$

(8. 2)

$$0 < c_2(\alpha,\beta)k^{-\alpha-3/2}n^{\alpha+2} \leqq |P_n^{(\alpha,\beta)\prime}(\cos\vartheta_{kn})| \leqq c_3(\alpha,\beta)k^{-\alpha-3/2}n^{\alpha+2}\left(0 < \vartheta_k \leqq \frac{\pi}{2}\right)^*.$$

---

*$x_{kn} = \cos\vartheta_{kn}$ is the k-th root of $P_n^{(\alpha,\beta)}(x)$.

Let $I_n = \left\{x_{\left[\frac{n}{2}\right]+1,n}; x_{\left[\frac{n}{4}\right]-1,n}\right\}$. By (8. 1) and (8. 2) we get

$$\lambda_n^{(\alpha,\beta)}(I_n; 1) = \sum_{k=\left[\frac{n}{4}\right]}^{\left[\frac{n}{2}\right]} |l_{kn}(x)| = \sum_{k=\left[\frac{n}{4}\right]}^{\left[\frac{n}{2}\right]} \frac{P_n^{(\alpha,\beta)}(1)}{|P_n^{(\alpha,\beta)\prime}(x_{kn})|(1-x_{kn})} \geqq$$

$$\geqq c_4(\alpha,\beta)\frac{1}{n^2} \sum_{k=\left[\frac{n}{4}\right]}^{\left[\frac{n}{2}\right]} k^{\alpha+\frac{3}{2}} \qquad (n = 8, 9, \ldots).$$

(We used that $1-x_{kn} \geqq \delta$.)

I.e., we have

$$\lambda_n^{(\alpha,\beta)}(I_n; 1) \geqq c(\alpha,\beta) n^{\alpha+\frac{1}{2}} \qquad (n=8, 9,\ldots). \tag{8. 3}$$

Now by (8. 3), (3. 5) and Theorem 1. 2 we have

THEOREM 8. 1. *There exist an* $f(x) \in C^{[-1,1]}(\omega_m)$ *and a sequence* $0<n_1<n_2<\ldots$ *such that*

$$|L_n^{(\alpha,\beta)}(f; 1)-f(1)| > n^{\alpha+\frac{1}{2}}\,\omega_m\left(\frac{1}{n}\right) \qquad (n=n_1, n_2, \ldots; \alpha, \beta > -1). \tag{8. 4}$$

To apply Theorem 1. 1. we have to estimate $\lambda_n^{(\alpha,\beta)}(1)$. By (8. 1), (8. 2) and the relations (see [3], (8. 9. 1) and (4. 1. 3))

$$0 < c_5(\alpha,\beta)\frac{k^2}{n^2} \leqq 1-x_{kn} \leqq c_6(\alpha,\beta)\frac{k^2}{n^2} \qquad (0 < x_{kn} \leqq 1)^*, \tag{8. 5}$$

$$P_n^{(\alpha,\beta)}(x) = (-1)^n P_n^{(\beta,\alpha)}(-x) \tag{8. 6}$$

we have

$$\lambda_n^{(\alpha,\beta)}(1) = \sum_{k=1}^{n} |l_{kn}(1)| = \sum_{k=1}^{\left[\frac{n}{2}\right]} + \sum_{k=\left[\frac{n}{2}\right]+1}^{n} \geqq$$

$$\geqq c_7(\alpha,\beta)\left(\sum_{k=1}^{\left[\frac{n}{2}\right]} k^{\alpha-\frac{1}{2}} + n^{\alpha-\beta-2} \sum_{k=\left[\frac{n}{2}\right]+1}^{n} k^{\beta+\frac{3}{2}}\right) \geqq c_8(\alpha,\beta)\left(\sum_{k=1}^{\left[\frac{n}{2}\right]} k^{\alpha-\frac{1}{2}} + n^{\alpha+\frac{1}{2}}\right).$$

I.e., we have

$$\lambda_n^{(\alpha,\beta)}(1) \geqq \begin{cases} c_9(\alpha,\beta)\, n^{\alpha+\frac{1}{2}} & \alpha > -\frac{1}{2}, \\ c_9(\alpha,\beta)\log n & \alpha = -\frac{1}{2}, \\ c_9(\alpha,\beta) & -1 < \alpha < -\frac{1}{2} \end{cases} \tag{8. 7}$$

for $c_9(\alpha,\beta) > 0$.

---

* $1-\cos\vartheta_{kn} = 2\sin^2\frac{\vartheta_{kn}}{2}$. I.e., by (8. 9. 1) (see [3]) we obtain (8. 5),

Paying attention to (8.7), (3, 5) and Theorem 1. 1. we get

THEOREM 8. 2. *There exist an* $f(x) \in C^{[-1,1]}(\omega_m)$ *and a sequence* $0 < n_1 < n_2 < \ldots$ *such that*

$$(8.8) \qquad |L_n^{(\alpha,\beta)}(f;1) - f(1)| > \begin{cases} n^{\alpha+\frac{1}{2}}\,\omega_m\left(\frac{1}{n^2}\right) & \text{if} \quad \alpha > -\frac{1}{2}, \\ \log n \cdot \omega_m\left(\frac{1}{n^2}\right) & \text{if} \quad \alpha = -\frac{1}{2}, \\ \omega_m\left(\frac{1}{n^2}\right) & \text{if} \quad \alpha < -\frac{1}{2}, \quad \lim\limits_{n=\infty} \lambda_n^{(\alpha,\beta)}(1) \neq 1. \end{cases}$$

9. Notes

9. 1. We can obtain analogous estimations to (8. 4) and (8. 8) for the point -1 replacing α by β.

9. 2. We can consider the trigonometric cosine interpolation based on the nodes $\vartheta_{kn}^{(\alpha,\beta)}$ $(n = 1, 2, 3, \ldots;\ k = \pm 1, \pm 2, \ldots, \pm n)$ for 2π-periodic continuous even functions.

9. 3. We can find the case of $\alpha = \beta = -\frac{1}{2}$ of Theorem 8. 1. in [4]; similarly, we can find some special cases of Theorem 8. 2, in [5] $(\alpha, \beta > -\frac{1}{2},\ \omega_1(t) = t^a)$. The proofs run in "direct way".

9. 4. With the previous notations we have the well-known estimation

$$(9.1) \qquad |f(x) - L_n(f;x)| = O(1)\left[\lambda_n(x)\,\omega_1\left(\frac{1}{n}\right) + \omega_1\left(\frac{1}{n}\right)\right] \qquad (n = 1, 2, \ldots)^*$$

for the mentioned Lagrange interpolations. This means that the order of (2. 3) and (7. 3) is the best possible for $-1 < x < 1$.

9. 5. By (7. 2), Theorem 7. 1, (8. 3), Theorem 8. 1 and 9. 1 we have the following fact. If $\alpha, \beta > -\frac{1}{2}$, then there exists an $f(x) \in C^{[-1,1]}(\omega_1)$ such that

$$\lambda_n^{(\alpha,\beta)}(x^*)\,\omega_1\left(\frac{1}{n}\right) < |f(x^*) - L_n^{(\alpha,\beta)}(f;x^*)| = O(1)\,\lambda_n^{(\alpha,\beta)}(x^*)\,\omega_1\left(\frac{1}{n}\right)$$

$$(n = n_1, n_2, \ldots),$$

where $x^* \in [-1, 1]$, $\alpha, \beta > -\frac{1}{2}$, the sequence $0 < n_1 < n_2 < \ldots$ and the function $f(x)$ depend on the point x^* as well. This means that our estimations are the best-possible in the whole interval $[-1, 1]$ for $\alpha, \beta > -\frac{1}{2}$.

---

* Here $\omega_1(t) = t$ is admissible as well, $f(x)$ is continuous.

9. 6. Let $\alpha=\beta=-\frac{1}{2}$ and $\lim_{t=0}\frac{\omega_1(t^2)}{\omega_1(t)}>0$.* Then by (7. 2), Theorem 7. 1, (8. 7), Theorem 8. 2., 9. 1 and (9. 1) we have that there exists an $f(x)\in C^{[-1,1]}(\omega_1)$ such that

$$\log n\cdot\omega_1\left(\frac{1}{n}\right)<|f(x^*)-L_n^{\left(-\frac{1}{2},-\frac{1}{2}\right)}(f;x^*)|=O(\log n)\,\omega_1\left(\frac{1}{n}\right)$$

$$(n=n_1,n_2,\ldots),$$

where $x^*\in[-1,1]$, the sequence $0<n_1<n_2<\ldots$ and the function $f(x)$ depend on the point x^* as well. So we have proved a statement which is more exact than the corresponding special case of the general theorem of LOSINSKY (see [6]).

9. 7. By the notations Theorem 6. 1 we have

$$|g(x)-L_n(g;x)|=O(1)\left[\lambda_n(x)\,\omega_m\left(\frac{1}{n}\right)+\omega_m\left(\frac{1}{n}\right)\right].$$

(If $g(x)$ is even then $L_n(g;x)$ is a cosine-polynom.) By these we can consider the above mentioned facts for the trigonometric interpolation as well. We can use the note 9. 2 also.

REFERENCES

[1] KIS, O. and VÉRTESI, P. O. H.: On certain linear operators, *Acta Math. Acad. Sci. Hung.*, **22** (1971).

[2] А. Х. Турецкий: *Теория интерполирования в задачах*, Минск, 1968.

[3] SZEGŐ, G.: *Orthogonal polynomials*, Amer. Math. Soc. New York, 1959.

[4] О. Киш (Kis, O.): Замечания о порядке сходимости Магранжева интерполирования, *Annales Univ. Sci. Budapest*, **11** (1968), 27—40.

[5] RÓNA, G.: Lagrange-interpolation based on the nodes of Jacobi polynomials. Thesis, Budapest, 1969, (Hungarian).

[6] С. М. Лозинский: Пространства $\tilde{C}_\omega$ и $\tilde{C}^*_\omega$ и сходимость интерполяционных процессов в них, ДАН **59** (1948), (1389—1392).

Eötvös L. University, Budapest

(Received July 12, 1969.)

* E.g.

$$\omega_1(t)=\begin{cases}\dfrac{1}{|\log t|} & (0\leqq t\leqq t_1<1),\\[2ex] \dfrac{1}{|\log t_1|} & (t_1\leqq t).\end{cases}$$

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 411—416.

GEODESIC LINES ON A BOUNDED CLOSED CONVEX POLYHEDRON

by

K. A. POST

1. Introduction

Any equilateral tetrahedron has infinite non-self-intersecting geodesic lines. We shall see below how this follows from a consideration of the network of such a tetrahedron. The main result to be proved in this paper (cf. theorem 2. 1) is that the converse is also true. In other words, among all bounded closed convex polyhedra the equilateral tetrahedron is characterized by the existence of an infinite non-self-intersecting geodesic line. In the final Section 3 a necessary condition for the existence of a closed geodesic line without double-points on a closed convex polyhedron is given, and some related problems are suggested.

The local structure of a polyhedron may be isometrically described by

(a) a circular disk in the Euclidean plane, if we consider a sufficiently small neighbourhood of a point that is not a vertex, or by

(b) a cone, if we consider a sufficiently small neighbourhood of a vertex T. In this case we can assign a positive quantity γ, called *curvature*, to the vertex T, taking the missing part of the full angle when we make a network of the cone.

Euler's polyhedral relation may be stated in the form

$$\sum \gamma_i = 4\pi, \tag{1. 1}$$

the summation being taken over all vertices T_i (cf. ALEXANDROW [1] p. 68, 69). This formula is clearly a discrete version of the integral curvature theorem (cf. STRUIK [3]).

As an example, we have the regular dodecahedron with 20 vertices having curvature $\frac{1}{5}\pi$ each. Our main example is the equilateral tetrahedron (disphenoid, cf. FEJES TÓTH [2]) which has the following network (Fig. 1): The plane is regularly covered with copies of an acute-angled triangle. Identically indexed vertices have to be identified, and the triangles corresponding with one face of the tetrahedron are shaded.

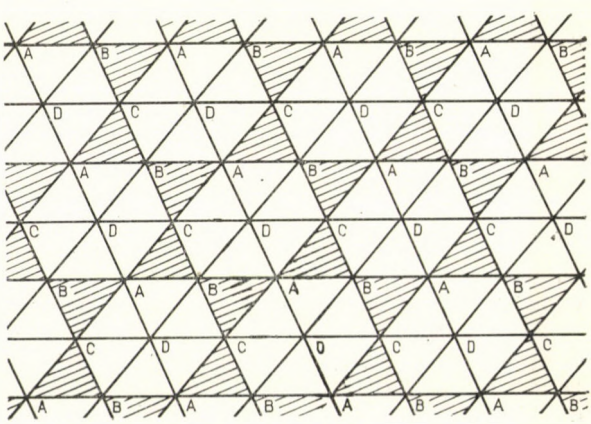


Fig. 1

LEMMA 1. 1. *A convex polyhedron is an equilateral tetrahedron iff four of its vertices have curvature* π.

PROOF. Trivial (cf. (1. 1)).

DEFINITION. A curve on a polyhedron is called a *geodesic line*, if it contains no vertex, and if it is locally isometric with a straight line. In practice one may draw a geodesic line by constructing a straight line in some network of the polyhedron. This network has to be understood in such a way, that each face may occur more than once (cf. Fig. 1). Without entering into the details of the proof we state the following theorem:

THEOREM 1. 1. *Through any point (not a vertex) of a polyhedron there may be constructed infinite geodesic lines in all but countably many directions.*

Since in the network of an equilateral tetrahedron (Fig. 1) all copies of the same face may be obtained from each other by translation, or by rotation through π radians, a geodesic line in this case cannot have double-points. Therefore we have:

THEOREM 1. 2. *On an equilateral tetrahedron there exist infinitely many infinite geodesic lines without double-points.*

2. Geodesic characterization of an equilateral tetrahedron

In this Section we shall prove the following theorem:

THEOREM 2. 1. *A bounded, closed, convex polyhedron is an equilateral tetrahedron iff it contains an infinite, non-self-intersecting geodesic line.*

By theorem 1. 2 we only have to prove the "if" part. We need some elementary results on geometry in the plane and on the cone.

LEMMA 2. 1. *Suppose* $0<\alpha<\pi$ *and let* O, A, B *and* C *be four points in the plane such that* $\sphericalangle AOB = \alpha$, $\sphericalangle OAC = \sphericalangle OBC = \frac{\pi}{2}$, $|OA| = p$ *and* $|OB| = q$, *then* $|OC| = (p^2 - 2pq\cos\alpha + q^2)^{1/2}\operatorname{cosec}\alpha$.

PROOF. By straightforward calculation.

The network of a cone $C(T, \gamma)$ with top T and curvature γ is given by a sector of $2\pi-\gamma$ radians, the bounds of which correspond with the same generator.

LEMMA 2. 2. *If* l *is a geodesic line on a cone* $C(T, \gamma)$ *with curvature* $\gamma>\pi$, *and* T *has distance* s *from* l, *then* l *has a double-point* P *satisfying* $|PT| = -s\cdot\sec\frac{1}{2}\gamma$.

PROOF. Let the bounds of the network of $C(T, \gamma)$ correspond with the generator $l^{\perp}$ that intersects l perpendicularly. The assertion follows immediately.

It should be remarked that two generators of the cone $C(T, \gamma)$ span two geodesic angles, α and β say, such that $\alpha+\beta = 2\pi-\gamma$. The mutual position of two geodesic lines l and m on $C(T, \gamma)$ is determined by their distances from T and one of the geodesic angles spanned by the generators $l^{\perp}$ and $m^{\perp}$ perpendicular to l and m, respectively.

LEMMA 2. 3. *If $\gamma<\pi$, and l and m are two geodesic lines on $C(T,\gamma)$ with distances p and q from T, respectively, such that $l^{\perp}$ and $m^{\perp}$ span a geodesic angle α satisfying $\pi \leqq \alpha < \pi+\frac{1}{2}(\pi-\gamma)$, then l and m have a common point S such that*

$$|ST| < 2(p+q)\operatorname{cosec}\frac{1}{2}(\pi-\gamma).$$

PROOF. The inequality $\pi \leqq \alpha < \pi+\frac{1}{2}(\pi-\gamma)$ implies $\frac{1}{2}(\pi-\gamma) < \beta \leqq \pi-\gamma$ so that in view of Lemma 2. 1 we obtain a point S satisfying

$$|ST| < (p+q)\operatorname{cosec}\beta < 2(p+q)\operatorname{cosec}\frac{1}{2}(\pi-\gamma).$$

The network of a cone $C(T,\pi)$ with curvature $\gamma=\pi$ is the whole plane, one point T being marked, and antipodal points with respect to T being identified. Therefore we have

LEMMA 2. 4. *A vertex of curvature π is not a singularity in the network of a polyhedron.*

PROOF of theorem 2. 1 ("if")

Suppose Π is a polyhedron with v vertices $T_1, \ldots, T_v$, which contains an infinite, non-self-intersecting geodesic line l. We define

$$\text{(2. 1)} \qquad \begin{cases} d := \operatorname{Min}\{|T_iT_j| : 1 \leqq i < j \leqq v\}, \\ \varrho := \operatorname{Min}\{\gamma_j : 1 \leqq j \leqq v,\ \gamma_j > \pi\}, \\ \sigma := \operatorname{Min}\{\operatorname{Min}\{\gamma_j, \pi-\gamma_j\} : 1 \leqq j \leqq v,\ \gamma_j < \pi\}. \end{cases}$$

Let $\{P_k\}_{k=-\infty}^{+\infty}$ be a sequence of distinct points on l such that, distances being measured along l, we have

$$\text{(2. 2)} \qquad |P_kP_{k+1}| = \frac{1}{4}d \quad \text{for all } k.$$

The indices of P_k induce a direction on l, which is called positive, say. By Bolzano—Weierstrass' theorem the set $\{P_k\}_{k=-\infty}^{+\infty}$ has an accumulation point, Q say. In view of (2. 1) and (2. 2) we may assume that $|QT_j| \geqq \frac{1}{4}d$ $(j=1, \ldots, v)$. This implies that any geodesic circular neighbourhood V of Q with radius $r < \frac{1}{4}d$ is actually isometric with a plane circular disk. We take

$$\text{(2. 3)} \qquad r < \operatorname{Min}\left\{-\frac{1}{2}d\cos\frac{1}{2}\varrho, \frac{1}{8}d\sin\frac{1}{2}\sigma\right\}$$

(cf. (2. 1)) and remark that $V_r \cap l$ consists of infinitely many disjoint directed segments. Therefore we may assume that l_1 and l_2 are two disjoint, directed segments of $V_r \cap l$, the arguments of which differ by $\delta < \frac{1}{2}\sigma$ (cf. (2. 1)). We distinguish between several cases:

(a) $\delta>0$. Then in a network of Π we take points $A_1 \in l_1$, $A_2 \in l_2$ (hence $|A_1A_2|<2r$). Let $\bar{l}_1$ and $\bar{l}_2$ be obtained by extending l_1 and l_2 up to their common point U. If the triangle A_1A_2U contains no vertex of Π, then U actually corres-

ponds with a double-point of l. Contradiction. Therefore let us assume that $A_1 A_2 U$ contains a vertex T with curvature γ.

(a_1) $\gamma > \pi$. Now $|A_1 A_2| \leqq 2r < -d \cos \frac{1}{2} \varrho \leqq -d \cos \frac{1}{2} \gamma$ (cf. (2. 1) and (2. 3)) so that T has distance less than $-d \cos \frac{1}{2} \gamma$ from $\tilde{l}$. Hence, $\tilde{l}_1$ has a double-point S such that $|ST| \leqq d \cos \frac{1}{2} \gamma \sec \frac{1}{2} \gamma = d$ (cf. lemma 2. 2.) Contradiction.

(a_2) $\gamma < \pi$. In this case the distances p and q of T from $\tilde{l}_1$ and $\tilde{l}_2$ are both less than $2r$, so that $p+q < 4r < \frac{1}{2} d \sin \frac{1}{2} \sigma \leqq \frac{1}{2} d \sin \frac{1}{2} (\pi - \gamma)$ (cf. (2. 1) and (2. 3)). The geodesic angle α between $\tilde{l}_1^{\perp}$ and $\tilde{l}_2^{\perp}$ has the value $\pi + \delta$, so that $\pi < \alpha < \pi + \frac{1}{2} \sigma \leqq \pi + \frac{1}{2} (\pi - \gamma)$. Therefore, $\tilde{l}_1$ and $\tilde{l}_2$ have a common point S satisfying $|ST| < 2 \cdot \frac{1}{2} d \sin \frac{1}{2} (\pi - \gamma) \operatorname{cosec} \frac{1}{2} (\pi - \gamma) = d$ (cf. lemma 2. 3). Contradiction.

(a_3) $\gamma = \pi$. Such a vertex is not a singularity in the network (cf. lemma 2. 4) and this case immediately reduces to one of the previous cases. Contradiction.

Hence, we may conclude $\delta = 0$.

(b) $\delta = 0$. Let us fix the parallel segments l_1 and l_2 in a network at a distance $< 2r$ from each other and extend them in both directions to obtain the lines $\tilde{l}_1$ and $\tilde{l}_2$. If the strip between $\tilde{l}_1$ and $\tilde{l}_2$ contains a vertex T with curvature $\gamma \neq \pi$, the situation is essentially the same as in (a_1) and (a_2) and we obtain a contradiction.

Conclusion: l_1 and l_2 must be equidirected and the strip between $\tilde{l}_1$ and $\tilde{l}_2$ can only contain vertices with curvature π.

Since such a vertex is not a singularity in the network of Π (lemma 2. 4), and l_1 and l_2 originate from the same geodesic line l we must conclude that on Π the strip between $\tilde{l}_1$ and $\tilde{l}_2$ has to cover all equidistant, equidirected strips in the neighbourhood of l_1 and l_2, and hence must cover the vertices of at least two faces of Π. Therefore, Π must have four vertices of curvature π, so that by lemma 1. 1 it is an equilateral tetrahedron. Q.e.d.

3. Closed geodesic lines

Let Π be a polyhedron, and let k be a closed polygon on Π without double-points. Then k divides Π into two parts Π_1 and Π_2, for each of which we have

$$f - e + v = 1, \tag{3. 1}$$

where f, e and v denote the numbers of faces, edges and vertices of $\Pi_{1(2)}$, respectively, the elements of k being included. So let us write for Π_1

$$f - e_{(i)} - e_{(k)} + v_{(i)} + v_{(k)} = 1, \tag{3. 2}$$

the suffixes (i) and (k) referring to elements in the interior of Π_1 and on k, respectively. Similarly, we have the following relation for the complex Π_1' obtained by triangulation of Π_1:

$$f' - e'_{(i)} - e'_{(k)} + v'_{(i)} + v'_{(k)} = 1. \tag{3. 3}$$

The structure of Π_1' implies

$$3f' = 2e'_{(i)} + e'_{(k)}, \tag{3. 4}$$

$$e'_{(k)} = v'_{(k)}, \tag{3. 5}$$

$$\pi f' = \Sigma_{(i)} \alpha' + \Sigma_{(k)} \alpha'. \tag{3. 6}$$

In the last expression the angles are summed up pro vertex in the interior of Π_1' and on k, respectively. Elimination of f', $e'_{(i)}$ and $e'_{(k)}$ from (3. 3), (3. 4), (3. 5) and (3. 6) yields for Π_1'

$$\Sigma_{(i)}(2\pi-\alpha')+\Sigma_{(k)}(\pi-\alpha')=2\pi,$$

and hence for Π_1

$$\Sigma_{(i)}(2\pi-\alpha)+\Sigma_{(k)}(\pi-\alpha)=2\pi; \tag{3. 7}$$

in other words,

$$\Sigma_{(i)}\gamma+\Sigma_{(k)}\lambda=2\pi, \tag{3. 8}$$

where γ denotes the curvature in the interior vertices of Π_1 and λ the geodesic curvature of k in its vertices. Evidently (3. 8) is a discrete analogue of the Gauss—Bonnet's theorem (cf. STRUIK [3], ALEXANDROW [1] p. 68, 69). As a special case of (3. 8) we have the following theorem:

THEOREM 3. 1. *A necessary condition for the existence of a closed geodesic line on Π without double-points is that the vertices of Π may be partitioned into two classes, for each of which $\Sigma\gamma_i=2\pi$.*

However, this condition is not sufficient. Indeed, a counterexample is provided by a regular double-pyramid, the tops of which have curvature ε and $2\pi-\varepsilon$, ε being a sufficiently small positive number.

Problem 1. Find sufficient conditions for the existence of closed geodesic lines without double-points on a polyhedron.

Problem 2. Find necessary and sufficient conditions for the existence of closed geodesic lines on a polyhedron.

Example. By consideration of the regular tessellation of the plane with equilateral triangles (cf. FEJES TÓTH [2], p. 24) it is observed that any polyhedron the faces of which are built up from congruent equilateral triangles admits infinitely many closed geodesic lines issuing from any given point which is not a vertex.

Any closed geodesic line k on Π has a positive distance d, say, from the set of vertices. Therefore, there exists a family of closed geodesic lines parallel to k.

Problem 3. For a fixed polyhedron determine the number of different families of closed geodesic lines.

Given a family F of mutually parallel closed geodesic lines on Π the number $N(F)$ of double-points is the same for all members of the family.

Problem 4. For a fixed polyhedron Π determine all possible values of $N(F)$.

Example. If Π is an equilateral tetrahedron, then the number of geodesic families is infinite and $N(F)=0$ for all F.

REFERENCES

[1] ALEXANDROW, A. D.: *Konvexe Polyeder,* Berlin, Akademie-Verlag (1958).
[2] FEJES TÓTH, L.: *Regular Figures*, Oxford, Pergamon Press (1964).
[3] STRUIK, D. J.: *Lectures on classical differential geometry,* Cambridge Mass. Addison—Wesley (1950).

Technological University, Eindhoven, Netherlands

(Received July 12, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 417—419.

A CHARACTERIZATION OF THE COSINE

by

PL. KANNAPPAN

Characterization of various functions and in particular the trigonometric functions with the help of functional equations had been studied extensively. In this paper, a characterization of the cosine function with the help of a functional equation in a single variable is considered. For similar characterization of the cosine and sine functions refer [1], [2]. We prove the following theorem.

THEOREM. *Let $f: R \to R$ (R, real numbers) be such that, f satisfies*

$$f(2x) = 1 - 2f\left(\frac{\pi}{2} - x\right)^2 \quad \text{for all} \quad x \in R, \tag{1}$$

and f be even, continuous in a neighborhood of the origin and $f(x) \geqq 0$ for $0 \leqq x \leqq \frac{\pi}{2}$ and $f(x) \leqq 0$ for $\frac{\pi}{2} \leqq x \leqq \pi$. Then $f(x) = \cos x$ for all x in R.

PROOF. Changing x into $\frac{\pi}{2} + x$ in (1), we get

$$f(\pi + 2x) = 1 - 2f(-x)^2. \tag{2}$$

Replacting x by $\frac{\pi}{2} - x$ in (1), we have

$$f(\pi - 2x) = 1 - 2f(x)^2. \tag{3}$$

From (2) and (3) and using f even, we obtain

$$f(\pi + x) = f(\pi - x), \quad \text{for all} \quad x \in R. \tag{4}$$

Replacing x by $\pi + x$ in (4) and using f even, we have

$$f(2\pi + x) = f(x), \quad \text{for all} \quad x \in R. \tag{5}$$

Hence f is periodic with period 2π.

Since by hypothesis

$$\begin{cases} f(x) \geqq 0, & 0 \leqq x \leqq \frac{\pi}{2} \\ f(x) \leqq 0, & \frac{\pi}{2} \leqq x \leqq \pi \end{cases} \tag{6}$$

and f even and periodic with period 2π, the sign of $f(x)$ is determined for each x in R.

We will now prove that the function f is continuous everywhere. For, from (1), we have

$$f(x) = 1 - 2f\left(\frac{\pi}{2} - \frac{x}{2}\right)^2.$$

As x varies over $(\pi-\varepsilon, \pi+\varepsilon)$, $\frac{\pi}{2} - \frac{x}{2} \in \left(\frac{-\varepsilon}{2}, \frac{\varepsilon}{2}\right)$. As, by hypothesis, f is continuous in a neighborhood of the origin, f is continuous in a neighborhood of π, say in $(\pi-\varepsilon, \pi+\varepsilon)$. When $\frac{\pi}{2} - x \in (\pi-\varepsilon, \pi+\varepsilon)$, $2x \in (-\pi-2\varepsilon, -\pi+2\varepsilon)$. So, from (1), we see that f is continuous in $(-\pi-2\varepsilon, -\pi+2\varepsilon)$. As f is even, f is continuous in $(\pi-2\varepsilon, \pi+2\varepsilon)$. Proceeding in this way, we can conclude that f is continuous at all points. Finally, we will show that f is determined uniquely at every point x in R. As f is continuous everywhere, it is enough to prove that f takes unique values at every point x of the form $x = k\pi/2^{n-1}$, k any integer, and n any integer ≥ 1. The proof is based on induction on n.

Putting $x = \frac{-\pi}{2}$ in (1) and using f even, we get $f(\pi) = \frac{1}{2}$ or -1. But from (6), we have $f(\pi) = -1$. Letting $x = \frac{\pi}{2}$ in (1), we get $f(0) = 1$, again using (6). Now, for $n = 1$, the claim is true. For, $f(k\pi) = f(0) = 1$ for k, even integer and $f(k\pi) = f(\pi) = -1$, for k, odd integer. Now, suppose that f takes unique values for all x of the form, $x = k\pi/2^{n-1}$. Then, as

$$f\left(\frac{k\pi}{2^n}\right) = \pm\sqrt{\frac{1}{2}\left\{1 - f\left(\frac{(2^{n-1}-k)\pi}{2^{n-1}}\right)\right\}},$$

by using induction hypothesis and the fact that f has fixed sign at every point, we see that f takes unique values at every point $x = \frac{k\pi}{2^n}$. Thus f takes unique values at every point. But the function $\cos x$ satisfies all the conditions of the theorem and hence $f(x) = \cos x$, for all x.

Of course the equation (1) can be reduced to

$$g(2x) = 2g(x)^2 - 1, \quad x \in R \tag{7}$$

and then the solution of (1) can be found by using [3], by the following procedure.

Now define a function $g: R \to R$, such that

$$g(x) = -f(\pi - x), \quad x \in R. \tag{8}$$

Then from (1), (8) and using f even, we have

$$2g(x)^2 = 2f(\pi - x)^2 = 1 - f(2x - \pi) = 1 + g(2x), \quad x \in R.$$

Hence g satisfied the duplication formula (7). As f is continuous in a neighborhood 0 and so in a neighborhood of π, from (8), we see that g is continuous in a neighborhood of 0.

Since f is periodic with period 2π and satisfies (6), we have from (8) that $g(x) \geqq 0$ for $\frac{-\pi}{2} \leqq x \leqq \frac{\pi}{2}$ and $g(x) \leqq 0$, for $\frac{\pi}{2} \leqq x \leqq \frac{3\pi}{2}$ and g has period 2π. Thus from [3, p. 231], we see that $g(x) = \cos x$ for all x in R. Hence from (8), we have $f(x) = \cos x$ for all x in R.

Remark 1. $f(x) = +\frac{1}{2}$ satisfies (1) and (5) and f is continuous and even. But f does not satisfy (6).

Remark 2. $f(x) = \cos(2k+1)x$, k any integer, is even, continuous and satisfies (1) but not (6). It remains open to find out whether these are the only even, continuous solutions of (1).

REFERENCE

[1] DUBIKAJTIS: Sur une caracterisation de la fonction sinus, *Ann. Polon. Math.* **16** (1964), 117—120.
[2] KUCZMA, M.: On a characterization of the cosine, *Ann. Polon. Math.* **16** (1964), 53—57.
[3] KUCZMA, M.: *Functional equations in a single variable*, PWN, Warszawa, 1968.

University of Waterloo
Waterloo

(Received July 30, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 421—426

ÜBER DIE DURCHSICHTIGKEIT GITTERFÖRMIGER KUGELPACKUNGEN

von

J. HORVÁTH

Hern Prof. Dr. P. Szász zum 70. Geburstag gewidmet

1. Eine Menge kongruenter nicht übereinandergreifender Kugeln des n-dimensionalen euklidischen Raumes wird Kugelpackung (kurz Packung) genannt. Wir nennen die Packung gitterförmig, wenn die Mittelpunkte der Kugeln ein Punktgitter bilden.

Die Packung bildet in der Richtung eines gegebenen k-dimensionalen Unterraumes eine Wolke, wenn jeder Unterraum, der zu dem gegebenen parallel liegt, mindestens eine Kugel trifft. Wenn wir ein solches Kugelsystem auf einen zu dem gegebenen Unterraum total senkrecht [2] ligenden $(n-k)$-dimensionalen Unterraum orthogonal projizieren, so überdecken die Projektionen der Kugeln den $(n-k)$-dimensionalen Unterraum vollständig. Wenn eine Kugelpackung in der Richtung jedes k-dimensionalen Unterraumes eine Wolke bildet, dann sagen wir, daß die Packung bezüglich k-dimensionaler Unterräume eine Dunkelwolke bildet.

HEPPES [7] hat gezeigt, daß keine gitterförmige Kugelpackung des 3-dimensinalen Raumes eine Dunkelwolke bildet. Er bewies nämlich, daß jede gitterförmige Kugelpackung in drei linear unabhängigen Richtungen durchsichtig ist. HORTOBÁGYI [10] verschärfte dieses Resultat, indem er zeigte, daß sogar drei Rotationszylinder vom Radius $\frac{3\sqrt{2}}{4}-1 \approx 0{,}0606\ldots$ mit linear unabhängigen Achsenrichtungen existieren, die mit den Kugeln keinen gemeinsamen inneren Punkt haben.

All diese Untersuchungen wurden durch eine Arbeit von FEJES TÓTH [4] angeregt. Vgl. noch die Aufsätze von BÖRÖCZKY [1], DANZER [3], HAJÓS [6], und HEPPES [8], [9].

In dieser Arbeit beschäftigen wir uns mit der analogen n-dimensionalen Problemen.

2. Wir bezeichnen einen l-dimensionalen Unterraum des n-dimensionalen euklidischen Raumes $\mathbf{E}_n$ mit Π_n^l oder Σ_n^l.

Wenn r eine gegebene positive Zahl ist, so verstehen wir unter einem Zylinder $\mathbf{Z}_l^{n-l}(r)$ die Gesamtheit aller in $\mathbf{E}_n$ liegenden Punkte, deren Entfernung von einem vorgegebenen Unterraum Π_n^{n-l} höchtens r ist [12]. Π_n^{n-l} wird die Achse des Zylinders $\mathbf{Z}_l^{n-l}(r)$ genannt. Der Durchschnitt von $\mathbf{Z}_l^{n-l}(r)$ mit dem auf Π_n^{n-l} total orthogonalen l-dimensionalen Unterraum ist eine l-dimensionale Kugel von Radius r. Diese Kugel wird die Grundkugel des Zylinders $\mathbf{Z}_l^{n-l}(r)$ genannt und wird mit $\mathbf{K}_n^l(r)$ bezeichnet.

SATZ 1. *Es sei im* $\mathbf{E}_4$ *eine gitterförmige Packung von Einheitskugeln vorgegeben. Dann können wir stets vier Zylinder* $\mathbf{Z}_3^1\left(\frac{\sqrt{5}}{2}-1\right)$ *angeben, die alle voneinander*

linear unabhängige Achsenstellung haben und die Packung nicht treffen. Die Konstante $\frac{\sqrt{5}}{2}-1 \approx 0{,}118\ldots$ *lässt sich durch keine grössere ersetzen.*

Der Beweis beruht auf folgendem

HILFSSATZ. *Ist die Länge jeder Kante eines Tetraeders mindestens* $\sqrt{3}$ *und der Umkreisradius jeder Fläche mindestens* $\frac{3\sqrt{2}}{4}$, *so gilt für den Umkugelradius R des Tetraeders die Ungleichheit* $R \geqq \frac{\sqrt{5}}{2}$. *Gleichheit gilt dann und nur dann, wenn die Flächen des Tetraeders kongruente Dreiecke mit den Seitenlängen* $\sqrt{3}$, $\sqrt{3}$, 2 *sind.*

BEWEIS: Wenn unter den Flächen des Tetraeders ein stumpfwinkliges oder rechtwinkliges Dreieck auftritt, so ist der Umkreisradius dieses Dreiecks mindestens $\frac{\sqrt{5}}{2}$. Wir können deshalb voraussetzen, dass sämtliche Flächen des Tetraeders spitzwinklige Dreiecke sind. Unter den durch die Ecken des Tetraeders $ABCD$ definierten sphärischen Dreiecken sei ABC das Dreieck von minimalen Flächeninhalt. Offensichtlich ist dieser Flächeninhalt höchtens πR^2. Wir bezeichnen die durch die Ecken des Dreiecks ABC hindurchgehende Kreislinie mit k und den euklidischen Radius von k mit r. Nach den Bedingungen des Hilfssatzes gilt $r \geqq \frac{3\sqrt{2}}{4}$. Wir wählen die Bezeichnung so, dass BC die längste Seite des Dreiecks ABC sei (Fig. 1.). Wir bewegen die Punkte B und C auf k gegen A, so lange bis $AB = AC = \sqrt{3}$ ist. Bei dieser Bewegung nimmt der Flächeninhalt des sphärischen Dreiecks ABC ab. Wegen $r \geqq \frac{3\sqrt{2}}{4}$ gilt für das entstandende Dreieck ABC die Relation $BC \geqq 2$. Wir können voraussetzen, dass das neue Dreieck ABC spitzwinklig ist. (S. die einleitende Bemerkung.) Wenn $BC > 2$ ist, so drehen wir den Punkt B um A auf der Kugelfläche so, dass in der Endlage der Bewegung $BC = 2$ ausfällt. Es ist leicht einzusehen, dass dabei der Flächeninhalt und der Umkreisradius des sphärischen Dreiecks ABC abnimmt und in der Endlage $r = \frac{3\sqrt{2}}{4}$ wird.

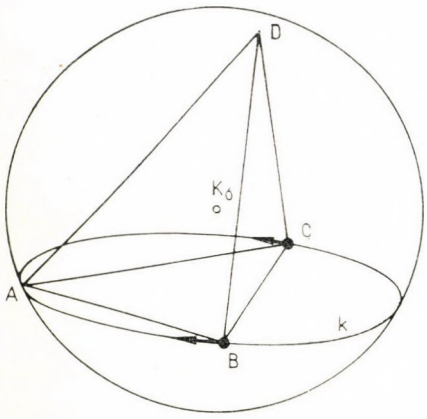


Fig. 1

Wir bezeichnen den Mittelpunkt der Umkugel des Tetraeders $ABCD$ mit K_0. Wir nehmen eine Kugel durch die Punkte ABC mit dem Radius $R^* = \frac{\sqrt{5}}{2}$ auf. Wir bezeichnen den Mittelpunkt dieser Kugel mit K^* und die um K_0 und K^* geschlagenen Einheitskugeln mit $\mathbf{E}_0$ bzw. $\mathbf{E}^*$. Wir proji-

zieren die Punkte, A, B und C auf $\mathbf{E}_0$ und $\mathbf{E}^*$ und bezeichnen die Projektionen mit A_0, B_0, C_0, bzw. mit A^*, B^*, C^* (Fig. 2.). Wegen der Wahl des ursprünglichen Dreiecks ABC ist der Flächeninhalt des sphärischen Dreiecks $A_0B_0C_0$ höchstens π. Der Flächeninhalt des sphärischen Dreiecks $A^*B^*C^*$ ist gleich π.

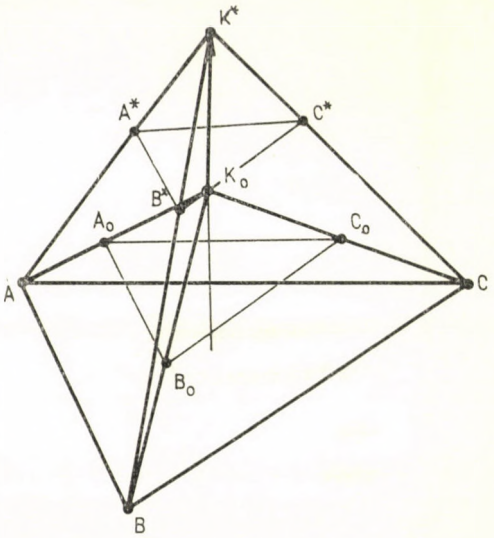


Fig. 2

Ist $R < \frac{\sqrt{5}}{2}$, so liegt der Punkt K_0 näher bei der Ebene ABC als K^*. Wir verschieben das sphärische Dreieck $A_0B_0C_0$ in der Richtung $\overrightarrow{K_0K^*}$. Nach der Verschiebung wird das sphärische Dreieck $A_0B_0C_0$ das sphärische Dreieck $A^*B^*C^*$ enthalten, was mit Rücksicht auf die Dreiecksinhalte unmöglich ist.

Es ist leicht einzusehen, dass im Falle $R = \frac{\sqrt{5}}{2}$ alle Flächen des Tetraeders miteinander kongruent und die Seitenlängen $\sqrt{3}$, $\sqrt{3}$, 2 sind.

Somit ist der Beweis des Hilfssatzes erbracht.

Es seien $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ beliebige linear unabhängige Gittervektoren. Wir bezeichnen einen der kürzesten Gittervektoren, der nicht in dem von den Vektoren $\mathbf{a}_1$, $\mathbf{a}_2$ und $\mathbf{a}_3$ aufgespannten 3-dimensionalen Unterraum Π_4^3 liegt, mit $\mathbf{a}_4$. Wir betrachten einen solchen 3-dimensionalen Unterraum Σ_4^3, der auf $\mathbf{a}_4$ orthogonal ist. Wir projizieren den Raum $\mathbf{E}_4$ sekrecht auf Σ_4^3. Diese Projektion erzeugt eine eineindeutige Abbildung zwischen den Unterräumen Π_4^3 und Σ_4^3, was wegen der linearen Unabhängigkeit von $\mathbf{a}_4$ von den Vektoren $\mathbf{a}_1$, $\mathbf{a}_2$ und $\mathbf{a}_3$ unmittelbar ersichtlich ist. Wir werden zeigen, dass die Projektionen der Kugeln Σ_4^3 nicht vollständig überdecken.

Die Projektionen der Kugelmittelpunkte auf Σ_4^3 bildet ein 3-dimensionales Punktgitter. Im diesem Gitter betrachten wir eine sogenannte Stützkugel [5], d.h. eine Kugel, die in ihrem Inneren keinen Gitterpunkt, auf ihrem Rand aber wenigsten 4 nicht koplanare Gitterpunkte C_0', C_1', C_2', C_3', enthält.

Wir bezeichnen mit $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{c}_3$ diejenigen in Π_4^3 liegenden Vektoren, die den linear unabhängigen Vektoren $\mathbf{c}_i' = \overrightarrow{C_0'C_i'}$ $(i = 1, 2, 3)$ entsprechend. Aus der Definition des Vektors $\mathbf{a}_4$ folgt, dass in denjenigen 3-dimensionalen Räumen, die entweder durch den Vektoren

$$\mathbf{c}_2 - \mathbf{c}_1, \quad \mathbf{c}_3 - \mathbf{c}_1, \quad \mathbf{a}_4,$$

oder durch den Vektoren

$$\mathbf{c}_i, \quad \mathbf{c}_k, \quad \mathbf{a}_4 \qquad (i, k = 1, 2, 3, \quad i \neq k)$$

aufgespannt werden, $\mathbf{a}_4$ ein kürzester Gittervektor ist. Deshalb ist jeder Flächen-

umkreisradius des Tetraeders $C_0'C_1'C_2'C_3'$ mindestens $\frac{3\sqrt{2}}{4}$ und jede Kante mindestens $\sqrt{3}$. (S. [7], [10].) Für $C_0'C_1'C_2'C_3'$ sind die Bedingungen des Hilfssatzes erfüllt, deshalb ist der Umkugelradius des Tetraeders $C_0'C_1'C_2'C_3' \geqq \frac{\sqrt{5}}{2}$. Wir bezeichnen den Mittelpunkt der Umkugel des Tetraeders mit C. Die um C geschlagene Kugel $\mathbf{K}_4^3\left(\frac{\sqrt{5}}{2}-1\right)$ trifft die Projektionskugeln in Σ_4^3 nicht. Deshalb trifft der Zylinder $\mathbf{Z}_3^1\left(\frac{\sqrt{5}}{2}-1\right)$ mit der um C geschlagenen Grundkugel $\mathbf{K}_4^3\left(\frac{\sqrt{5}}{2}-1\right)$, der eine zu $\mathbf{a}_4$ parallele Achsenrichtung hat, die ursprünglichen Kugeln in $\mathbf{E}_4$ nicht.

Wenn $R = \frac{\sqrt{5}}{2}$ ist, so sind die Flächen des Tetraeders $C_0'C_1'C_2'C_3'$ kongruente Dreiecke mit den Seitenlängen $\sqrt{3}$, $\sqrt{3}$, 2. Dieser Fall kommt vor, wenn wir die folgenden Gittervektoren wählen: $\mathbf{a}_1$ (2, 0, 0, 0), $\mathbf{a}_2$ (1, 1, 1, 1), $\mathbf{a}_3$ (1, −1, 1, 1), $\mathbf{a}_4$(0, 0, 0, 2). Diese Vektoren bestimmen die dichteste gitterförmige Packung des 4-dimensionalen Raumes [11].

SATZ 2. *Ist in $\mathbf{E}_n$ eine gitterförmige Packung von Einheitskugeln vorgegeben und ist $2 \leqq l \leqq n-1$, so können wir $\binom{n}{l}$ Zylinder $\mathbf{E}_l^{n-l}(r)$ mit verschiedenen Achsenstellungen angeben, die mit der Packung keinen gemeinsamen inneren Punkt haben, wobei für $l=2$ $r=\frac{3\sqrt{2}}{4}-1$ und für $3 \leqq l \leqq n-1$ $r=\frac{\sqrt{5}}{2}-1$ gibt.*

BEWEIS: Es seien $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ beliebige linear unabhängige Gittervektoren. Der Unterraum Π_n^l sei durch die Vektoren $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l$ aufgespannt. Es sei $\mathbf{b}_{l+1}$ ein kürzester nicht in Π_n^l liegender Gittervektoren. Die Gittervektoren $\mathbf{b}_{l+2}, \mathbf{b}_{l+3}, \ldots$ $\ldots, \mathbf{b}_n$ seien so gewählt, dass $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l, \mathbf{b}_{l+1}, \mathbf{b}_{l+2}, \ldots, \mathbf{b}_n$ den Raum $\mathbf{E}_n$ aufspannen. Wir bezeichnen den durch die Vektoren $\mathbf{b}_{l+1}, \mathbf{b}_{l+2}, \ldots, \mathbf{b}_n$ aufgespannten Unterraum mit Π_n^{n-l}, weiterhin den auf Π_n^{n-l} total orthogonalen l-dimensionalen Unterraum mit Σ_n^l. Da die auf Π_n^{n-l} total orthogonal liegenden l-dimensionalen Unterräume zueinander parallel sind, können wir voraussetzen, dass Σ_n^l einen Gitterpunkt O enthält. Wir projizieren $\mathbf{E}_n$ senkrecht auf Σ_n^l. Diese Projektion erzeugt zwischen den Unterräumen Π_n^l und Σ_n^l eine eineindeutige Abbildung. Wir werden zeigen, dass die Projektionen der Kugeln Σ_n^l nicht vollständig überdecken.

Die Projektionen der Kugelmittelpunkte auf Σ_n^l bildet ein l-dimensionales Punktgitter. In diesem Gitter betrachten wir eine Stützkugel [5], d.h. eine Kugel, die in ihrem Inneren keinen Gitterpunkt, auf ihrem Rand aber wenigsten $l+1$ nicht in einem $(l-1)$-dimensionalen Unterraum liegende Gitterpunkte $O, C_1', C_2', \ldots, C_l'$ enthält.

Wir bezeichnen mit $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_l$ diejenigen in Π_n^l liegenden Vektoren, die den linear unabhängigen Vektoren $\mathbf{c}_i' = \overrightarrow{OC_i'}$ $(i=1, 2, \ldots, l)$ entsprechen. Aus der Definition von $\mathbf{b}_{l+1}$ folgt, dass im Falle $l=2$ die Vektoren $\mathbf{c}_1, \mathbf{c}_2, \mathbf{b}_3$ linear unabhängig sind und in durch sie definierten 3-dimensionalen Unterraum $\mathbf{b}_3$ ein kürzester Git-

tervektor ist. Deshalb ist der Umkreisradius des Dreiecks $OC_1'C_2'$ mindestens $\frac{3\sqrt{2}}{4}$ [10]. Wenn $l \geqq 3$, so ist $\mathbf{b}_{l+1}$ ein kürzester Gittervektor in dem durch die Vektoren $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{b}_{l+1}$ aufgespannten 4-dimensionalen Unterraum. Deshalb folgt aus Satz 1., dass der Umkugelradius des Tetraeders $OC_1'C_2'C_3'$ mindestens $\frac{\sqrt{5}}{2}$ ist. Folglich ist der Umkugelradius des l-dimensionalen Simplexes $OC_1'C_2' \ldots C_l'$ mindestens $\frac{\sqrt{5}}{2}$.

Wir bezeichnen den Mittelpunkt der Umkugel dieses Simplexes mit C. Die um C geschlagene Kugel $\mathbf{K}_n^l(r)$ trifft die Projektionskugeln in Σ_n^l nicht, wobei für $l=2$ $r = \frac{3\sqrt{2}}{4} - 1$ und für $l \geqq 3$ $r = \frac{\sqrt{5}}{4} - 1$ gilt. Deshalb trifft der Zylinder $\mathbf{Z}_l^{n-l}(r)$ mit der Grundkugel $\mathbf{K}_n^l(r)$ um C, der eine zu Π_n^{n-l} parallele Achsenstellung $\overline{\Pi}_n^{n-l}$ hat, die ursprünglichen Kugeln nicht.

Damit ist der Beweis des Satzes 2. beendet.

Es sei bemerkt, dass sich in Satz 2. die Konstante $\frac{3\sqrt{2}}{4} - 1$ für keinen Wert von $n > 3$ vergrössern lässt. Wählen wir nämlich im obigen Beweis die Vektoren $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ so, dass sie ein regelmässiges Tetraeder mit der Kantenlänge 2 bestimmen. Projizieren wir $\mathbf{E}_n$ auf Σ_n^2, wo Σ_n^2 eine auf die Vektoren $\mathbf{a}_3, \mathbf{a}_4, \ldots, \mathbf{a}_n$ orthogonale Ebene ist, so bekommen wir ein Dreieck mit den Seitenlängen $\sqrt{3}, \sqrt{3}, 2$. So werden die Kugeln der Packung durch einen Zylinder $\mathbf{Z}_2^{n-2}\left(\frac{3\sqrt{2}}{4} - 1\right)$ berührt.

Ähnlich können wir zeigen, dass sich die Konstante $\frac{\sqrt{5}}{2} - 1$ im Falle $l = 3$ für keinem Wert von $n > 4$ vergrössern lässt.

Um auch für $l=4$ eine genaue Konstante zu erhalten, wäre die Lösung des folgenden Problems nötig. Gesucht wird das Minimum des Umkugelradius eines 4-dimensionalen Simplexes mit der Eigenschaft, das die Umkugelradien seiner 1-, 2-, 3-dimensionalen Zellen mindestens $\frac{\sqrt{3}}{2}, \frac{3\sqrt{2}}{4}$ bzw. $\frac{\sqrt{5}}{2}$ sind.

LITERATUR

[1] BÖRÖCZKY, K.: Über Dunkelwolke, *Proc. Coll. Convexity,* Coppenhagen, 1965. 13—17.
[2] COXETER, H. S. M.: *Regular polytopes,* London 1950.
[3] DANZER, L.: Drei Beispiele zu Lagerungsproblemen, *Arch. Math.* **11** (1960), 1959—65.
[4] FEJES TÓTH, L.: Verdeckung einer Kugel durch Kugeln, *Publ. Math.* **6** (1959), 307—13.
[5] FEJES TÓTH, L.: *Regulären Figuren,* Akad. Kiadó, Budapest, 1965.
[6] HAJÓS, G.: Über Kreiswolken, *Ann. Univ. Sci. Budapest,* **7** (1965), 55—57.
[7] HEPPES, A.: Ein Satz über gitterförmige Kugelpackung, *Ann. Univ. Sci. Budapest,* **3—4** (1960—61), 89—90.
[8] HEPPES, A.: Über Kreis- und Kugelwolken, *Achta Math. Acad. Sci. Hungar.,* **12** (1961), 209—14.
[9] HEPPES, A.: On the number of spheres which can hide a given sphere, *Can. J. Math.* **19** (1967), 413—18.
[10] HORTOBÁGYI, I.: Durchleutung gitterförmiger Kugelpackung mit Lichtbündeln, *Studia Sci. Math. Hung. (Im Druck).*
[11] KORKINE, A. and ZOLOTAREFF, G.: Sur les formes quadratique positive quaternaires, *Math. Ann.* **5** (1872), 581—3.
[12] Розенфельд, Б. А.: *Многомерные пространства,* Издательство «Наука» Москва, 1966.

Eötvös L. Universität, Budapest

(Eingegangen: 18. Juli, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 427—432.

ОБ ОЦЕНКЕ ПОГРЕШНОСТИ МЕТОДА РУНГЕ-КУТТА

О. КИШ

Решая задачу Коши

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

методом Рунге—Кутта, приближенно вычисляют решение задачи в некоторых точках, которые мы обозначим через $x_2, x_4, x_6, \ldots, x_n$. Приближенное значение y_2 величины $y(x_2)$ вычисляется с помощью формул

$$h = \frac{1}{2}(x_2 - x_0),$$

$$k_1 = hf(x_0, y_0), \tag{2}$$

$$k_2 = hf(x_0 + h, y_0 + k_1), \tag{3}$$

$$k_3 = hf(x_0 + h, y_0 + k_2), \tag{4}$$

$$k_4 = hf(x_2, y_0 + 2k_3), \tag{5}$$

$$y_2 = y_0 + \frac{1}{3}(k_1 + 2k_2 + 2k_3 + k_4). \tag{6}$$

Аналогичным образом вычисляются приближенные значения $y_4, y_6, \ldots, y_n$ величин $y(x_4), y(x_6), \ldots, y(x_n)$.

Обычно точки $x_2, x_4, x_6, \ldots, x_n$ выбирают так, чтобы выполнялись условия

$$x_{4i} - x_{4i-2} = x_{4i-2} - x_{4i-4} \qquad \left(i = 1, 2, 3, \ldots, \frac{n}{4}\right)$$

и приближенно оценивают погрешность вычисления величин $y_4, y_8, y_{12}, \ldots, y_n$. Например погрешность $y(x_4) - y_4$ оценивается следующим образом. Вычислив y_2 и y_4, вычисляют новое приближенное значение Y_4 величины $y(x_4)$, повторяя вычисления с удвоенным шагом:

$$K_1 = 2hf(x_0, y_0),$$

$$K_2 = 2hf(x_2, y_0 + K_1),$$

$$K_3 = 2hf(x_2, y_0 + K_2),$$

$$K_4 = 2hf(x_4, y_0 + 2K_3),$$

$$Y_4 = y_0 + \frac{1}{3}(K_1 + 2K_2 + 2K_3 + K_4).$$

Величину $\frac{1}{15}(y_4 - Y_4)$ можно считать приближенным значением погрешности $y(x_4) - y_4$, ибо, как известно,

$$y(x_4) - y_4 = O(h^5) = \frac{1}{15}(y_4 - Y_4) + O(h^6), \tag{7}$$

если функция $f(x, y)$ достаточно гладка в некоторой области. Эта оценка погрешности требует вычисления трех значений функции $f(x, y)$, которые фигурируют в K_2, K_3, K_4 (K_1 вычисляется по формуле $K_1 = 2k_1$).

Покажем, как можно оценить величину погрешности $y(x_4) - y_4$, вычисляя лишь два значения функции $f(x, y)$ и попутно получая приближенное значение y_1 и y_3 решения в точках

$$x_1 = x_0 + h, x_3 = x_2 + h.$$

Аналогичным образом можно оценивать погрешность вычисления $y_8, y_{12}, \ldots, y_n$ и приближенно вычислять решение в точках

$$x_i = \frac{1}{2}(x_{i+1} - x_{i-1}) \qquad (i = 5, 7, 9, \ldots, n-1).$$

Величину y_4 мы вычисляем с помощью формул

$$k_5 = hf(x_2, y_2), \tag{8}$$

$$k_6 = hf(x_3, y_2 + k_5), \tag{9}$$

$$h_7 = hf(x_3, y_2 + k_6), \tag{10}$$

$$k_8 = hf(x_4, y_2 + 2k_7), \tag{11}$$

$$y_4 = y_2 + \frac{1}{3}(k_5 + 2k_6 + 2k_7 + k_8).$$

Пусть

$$y_1 = y_0 + \frac{1}{48}(18k_1 + 19k_2 + 19k_3 + 10k_4 - 20k_5 + k_6 + k_7), \tag{12}$$

$$y_3 = y_0 + \frac{1}{16}(6k_1 + 9k_2 + 9k_3 + 6k_4 + 12k_5 + 3k_6 + 3k_7). \tag{13}$$

Ниже мы покажем, что

$$y(x_1) - y_1 = O(h^5), \tag{14}$$

$$y(x_3) - y_3 = O(h^5); \tag{15}$$

поэтому *числа* y_1 *и* y_3 *можно считать приближенными значениями величин* $y(x_1)$ *и* $y(x_3)$.

Вычисляя y_2, y_4 и y_6, мы вычисляем и $f(x_0, y_0)$, $f(x_2, y_2)$, $f(x_4, y_4)$. Вычислим еще и $f(x_1, y_1)$, $f(x_3, y_3)$. Пусть

(16)
$$\eta_4 = y_0 + \frac{2h}{45}\left(7f(x_0, y_0) + 32f(x_1, y_1) + 12f(x_2, y_2) + 32f(x_3, y_3) + 7f(x_4, y_4)\right).$$

Ниже мы покажем, что

(17)
$$y(x_4) - y_4 = \eta_4 - y_4 + O(h^6).$$

Поэтому *разность* $\eta_4 - y_4$ *можно считать приближенным значением погрешности* $y(x_4) - y_4$.

Доказательство равенства (14). Пусть имеет место (2) и

(18)
$$k_i = hf\left(x_0 + a_i h, y_0 + \sum_{j=1}^{i-1} b_{ij} k_j\right) \qquad (i = 2, 3, \ldots, r),$$

(19)
$$y_1 = y_0 + \sum_{i=1}^{r} c_i k_i.$$

Известно, что равенство (14) выполняется, если выполняются следующие условия:

(20)
$$\sum_{i=1}^{r} c_i = 1,$$

(21)
$$\sum_{i=2}^{r} c_i a_i = \frac{1}{2},$$

(22)
$$\sum_{i=2}^{r} c_i a_i^2 = \frac{1}{3},$$

(23)
$$\sum_{i=3}^{r} c_i \sum_{j=2}^{i-1} b_{ij} a_j = \frac{1}{6},$$

(24)
$$\sum_{i=2}^{r} c_i a_i^3 = \frac{1}{4},$$

(25)
$$\sum_{i=3}^{r} c_i a_i \sum_{j=2}^{i-1} b_{ij} a_j = \frac{1}{8},$$

(26)
$$\sum_{i=3}^{r} c_i \sum_{j=2}^{i-1} b_{ij} a_j^2 = \frac{1}{12},$$

(27)
$$\sum_{i=4}^{r} c_i \sum_{j=3}^{i-1} b_{ij} \sum_{l=2}^{j-1} b_{jl} a_l = \frac{1}{24},$$

(28)
$$\sum_{j=1}^{i-1} b_{ij} = a_i \qquad (i = 2, 3, \ldots, r).$$

В формулах (2)—(5) и (8)—(11), ввиду (6)

$$a_2 = b_{21} = a_3 = b_{32} = b_{65} = b_{76} = 1, \quad b_{31} = b_{41} = b_{42} = b_{75} = b_{85} = b_{86} = 0,$$

$$a_4 = b_{43} = a_5 = b_{87} = 2, \quad b_{51} = b_{54} = b_{61} = b_{64} = b_{71} = b_{74} = b_{81} = b_{84} = \frac{1}{3},$$

$$b_{52} = b_{53} = b_{62} = b_{63} = b_{72} = b_{73} = b_{82} = b_{83} = \frac{2}{3}, \quad a_6 = a_7 = 3, \quad a_8 = 4.$$

Эти значения удовлетворяют условиям (28). Подставляя эти значения в уравнения (20)—(27), переписываем их в виде

$$c_1 + c_2 + c_3 + c_4 + c_5 + c_6 + c_7 + c_8 = 1, \tag{29}$$

$$c_2 + c_3 + 2c_4 + 2c_5 + 3c_6 + 3c_7 + 4c_8 = \frac{1}{2}, \tag{30}$$

$$c_2 + c_3 + 4c_4 + 4c_5 + 9c_6 + 9c_7 + 16c_8 = \frac{1}{3}, \tag{31}$$

$$c_3 + 2c_4 + 2c_5 + 4c_6 + 5c_7 + 8c_8 = \frac{1}{6}, \tag{32}$$

$$c_2 + c_3 + 8c_4 + 8c_5 + 27c_6 + 27c_7 + 64c_8 = \frac{1}{4}, \tag{33}$$

$$c_3 + 4c_4 + 4c_5 + 12c_6 + 15c_7 + 32c_8 = \frac{1}{8}, \tag{34}$$

$$c_3 + 2c_4 + \frac{8}{3}c_5 + \frac{20}{3}c_6 + \frac{35}{3}c_7 + \frac{62}{3}c_8 = \frac{1}{12}, \tag{35}$$

$$2c_4 + \frac{4}{3}c_5 + \frac{10}{3}c_6 + \frac{16}{3}c_7 + \frac{34}{3}c_8 = \frac{1}{24}. \tag{36}$$

Решением этой системы уравнений служат числа

$$c_1 = \frac{3}{8} + c, \quad c_2 = c_3 = \frac{19}{48} - 2c, \quad c_4 = \frac{5}{24} - 3c, \quad c_5 = -\frac{5}{12} + 9c,$$

$$c_6 = c_7 = \frac{1}{48} - 2c, \quad c_8 = c,$$

где c любое вещественное число. Полагая $c = 0$, получаем:

$$c_1 = \frac{3}{8}, \quad c_2 = c_3 = \frac{19}{48}, \quad c_4 = \frac{5}{24}, \quad c_5 = -\frac{5}{12}, \quad c_5 = c_7 = \frac{1}{48}, \quad c_8 = 0.$$

Подставляя эти значения в (19), получаем (12). Поэтому определенное формулой (12) число y_1 действительно удовлетворяет условию (14).

Доказательство равенства (15). Введя обозначения

$$H = 3h,$$

$$K_i = 3k_i \qquad (i = 1, 2, \ldots, r),$$

$$A_i = \frac{a_i}{3} \qquad (i = 2, 3, \ldots, r), \tag{37}$$

$$B_{ij} = \frac{b_{ij}}{3} \qquad (i = 2, 3, \ldots, r; \quad j = 1, 2, \ldots, i-1), \tag{38}$$

$$C_i = \frac{c_i}{3} \qquad (i = 1, 2, \ldots, r), \tag{39}$$

перепишем формулы (2) и (18) в виде

$$K_1 = Hf(x_0, y_0),$$

$$K_i = Hf\left(x_0 + A_i H, y_0 + \sum_{j=1}^{i-1} B_{ij} K_j\right).$$

Пусть

$$y_3 = y_0 + \sum_{i=1}^{r} c_i k_i = y_0 + \sum_{i=1}^{r} C_i K_i.$$

Так как

$$x_3 = x_0 + H,$$

то условие (15) выполняется, если выполняются условия (20)—(28) с A_i, B_{ij} и C_i вместо a_i, b_{ij} и c_i. Ввиду (37)—(39) это равносильно выполнению условий (28), а также условий (20)—(27), правые части которых умножаются соответственно на числа 3, 9, 27, 27, 81, 81, 81, 81.

Мы уже убедились в том, что фигурирующие в (2)—(5) и (8)—(11) числа a_i, b_{ij} и c_i удовлетворяют условиям (28). Не трудно убедиться в том, что решением видоизмененных уравнений (20)—(27) служат числа

$$c_1 = \frac{3}{8} + c, \quad c_2 = c_3 = \frac{9}{16} - 2c, \quad c_4 = \frac{3}{8} - 3c, \quad c_5 = \frac{3}{4} + 9c, \quad c_6 = c_7 = \frac{3}{16} - 2c, \quad c_8 = c,$$

где c любое вещественное число; полагая здесь $c = 0$, получаем:

$$c_1 = \frac{3}{8}, \quad c_2 = c_3 = \frac{9}{16}, \quad c_4 = \frac{3}{4}, \quad c_5 = \frac{3}{4}, \quad c_6 = c_7 = \frac{3}{16}, \quad c_8 = 0.$$

Эти значения фигурируют в формуле (13), поэтому определенное этой формулой число y_3 действительно удовлетворяет условию (15).

Доказательство равенства (17). Ввиду (14), известного равенства

$$y(x_2) - y_2 = O(h^5),$$

(15) и (7) равенство (16) можно переписать в виде

$$\eta_4 = y_0 + \frac{2h}{45}\left(7f(x_0, y_0) + 32f(x_1, y(x_1)) + 12f(x_2, y(x_2)) + 32f(x_3, y(x_3)) + \right.$$
$$\left. + 7f(x_4, y(x_4))\right) + O(h^6).$$

Отсюда и из (1) следует:

$$\eta_4 = y_0 + \frac{2h}{45}\left(7y'(x_0) + 32y'(x_1) + 12y'(x_2) + 32y'(x_3) + 7y'(x_4)\right) + O(h^6).$$

Известно, что стоящая справа сумма Котэса может быть записана в виде

$$\int_{x_0}^{x_4} y'(x)\,dx + O(h^7) = y(x_4) - y(x_0) + O(h^7).$$

Поэтому

$$\eta_4 = y(x_4) + O(h^6),$$

что и требовалось доказать.

Пример. В случае задачи Коши

$$y' = y,\ y(0) = 1$$

пусть $h = 0{,}1$. Тогда

$y_1 = 1{,}105\,169\,36$; $y(0{,}1) = 1{,}105\,170\,92$; $y(0{,}1) - y_1 = 0{,}000\,001\,56$;

$y_2 = 1{,}221\,400\,00$; $y(0{,}2) = 1{,}221\,402\,76$; $y(0{,}2) - y_2 = 0{,}000\,002\,76$;

$y_3 = 1{,}349\,854\,26$; $y(0{,}3) = 1{,}349\,858\,81$; $y(0{,}3) - y_3 = 0{,}000\,004\,55$;

$y_4 = 1{,}491\,817\,96$; $y(0{,}4) = 1{,}491\,824\,70$; $y(0{,}4) - y_4 = 0{,}000\,006\,74$;

$Y_4 = 1{,}491\,733\,33$; $\frac{1}{15}(y_4 - Y_4) = 0{,}000\,005\,64$;

$\eta_4 = 1{,}491\,823\,47$; $\eta_4 - y_4 = 0{,}000\,005\,51$.

В этом примере погрешности чисел y_1, y_2, y_3, y_4 растут почти линейно и приближенные значения $\frac{1}{15}(y_4 - Y_4)$ и $\eta_4 - y_4$ погрешности $y(x_4) - y_4$ почти совпадают.

Политехнический Институт, Будапешт

(Поступило 12. 9. 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 433—435.

О МЕТОДЕ РУНГЕ-КУТТА

О. КИШ

Для приближенного решения задачи Коши

$$y' = f(x, y), \qquad y(x_0) = y_0$$

Р. Х. Мерсон (см. [1], стр. 76) предложил формулы

$$k_1 = hf(x_n, y_n),$$

$$k_2 = hf\left(x_n + \frac{h}{3}, y_n + \frac{k_1}{3}\right),$$

$$k_3 = hf\left(x_n + \frac{h}{3}, y_n + \frac{k_1}{6} + \frac{k_2}{6}\right),$$

$$k_4 = hf\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{8} + \frac{3k_3}{8}\right),$$

$$k_5 = hf\left(x_n + h, y_n + \frac{k_1}{2} - \frac{3k_3}{2} + 2k_4\right),$$

$$y_{n+1} = y_n + \frac{1}{10}(k_1 + 3k_3 + 4k_4 + 2k_5),$$

$$Y_{n+1} = y_n + \frac{1}{6}(k_1 + 4k_4 + k_5),$$

где $h = x_{n+1} - x_n$. Обозначим через $Y(x)$ решение задачи Коши

$$Y' = f(x, Y), \quad Y(x_n) = y_n.$$

Известно (см. [2]), что

(1) $$Y(x_{n+1}) - Y_{n+1} = O(h^5),$$

(2) $$Y(x_{n+1}) - y_{n+1} = O(h^4)$$

(если функция $f(x, y)$ достаточно гладкая). В связи с этим возникает следующий вопрос: существуют ли формулы вида

$$k_1 = hf(x_n, y_n),$$
$$k_1 = hf(x_n + a_2 h, y_n + b_{21} k_1),$$
$$k_3 = hf(x_n + a_3 h, y_n + b_{31} k_1 + b_{32} k_2),$$
$$k_4 = hf(x_n + a_4 h, y_n + b_{41} k_1 + b_{42} k_2 + b_{43} k_3),$$
$$y_{n+1} = y_n + c_1 k_1 + c_2 k_2 + c_3 k_3 + c_4 k_4,$$
$$Y_{n+1} = y_n + d_1 k_1 + d_2 k_2 + d_3 k_3 + d_4 k_4,$$

для которых выполняются условия (1)—(2) и хотя бы одно d_i отличается от соответствующего c_i? Докажем, что такие формулы не существуют: если выполняются условия (1)—(2), то

$$c_1 = d_1, c_2 = d_2, c_3 = d_3, c_4 = d_4.$$

Если выполняется условие (2), то, как известно

$$c_1 + c_2 + c_3 + c_4 = 1 \tag{3}$$

$$c_2 a_2 + c_3 a_3 + c_4 a_4 = \frac{1}{2}, \tag{4}$$

$$c_2 a_2^2 + c_3 a_3^2 + c_4 a_4^2 = \frac{1}{3}, \tag{5}$$

$$c_3 b_{32} a_2 + c_4 (b_{42} a_2 + b_{43} a_3) = \frac{1}{6}. \tag{6}$$

Если числа a_i, b_{ij}, d_i таковы, что выполняется условие (1), то (см. [3], стр. 309) $a_4 = 1$. Поэтому в (4) и (5) можно опустить a_4. Умножая (4) на a_2 и вычитая результат из (5), получаем:

$$c_3 a_3 (a_3 - a_2) + c_4 (1 - a_2) = \frac{1}{3} - \frac{1}{2} a_2. \tag{7}$$

Из (1) следует также (см. [3], стр. 309), что

$$2 b_{32} a_2 (2a_2 - 1) = a_3 (a_2 - a_3). \tag{8}$$

Умножая (6) на $2(2a_2 - 1)$ и принимая во внимание (8), имеем:

$$c_3 a_3 (a_2 - a_3) + 2c_4 (b_{42} a_2 + b_{43} a_3)(2a_2 - 1) = \frac{1}{3} (2a_2 - 1). \tag{9}$$

Складывая (7) и (9), получаем:

$$c_4 [1 - a_2 + 2(b_{42} a_2 + b_{43} a_3)(2a_2 - 1)] = \frac{1}{6} a_2. \tag{10}$$

Очевидно числа d_i также удовлетворяют условиям (3)—(7) и (9)—(10). Кроме того (см. [3], стр. 306) $a_2 \neq 0$. Поэтому из (10) получаем: $c_4 = d_4$. Если $a_3 \neq 0$ и $a_2 \neq a_3$, то из (9) следует: $c_3 = d_3$; тогда из (4) и (3) получаем: $c_2 = d_2$, $c_1 = d_1$, т. е. все c_i и d_i действительно совпадают. Если $a_3 = 0$ или $a_2 = a_3$, то $b_{32} \neq 0$ (см. [3], стр. 306) и $a_2 = \frac{1}{2}$ (см. [3], стр. 309) и поэтому из (7), (6), (4) и (3) снова получаем:

$$c_4 = d_4, c_3 = d_3, c_2 = d_2, c_1 = d_1 .$$

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] Дж. Н. Ланс: *Численные методы для быстродействующих вычислительных машин.* Москва, издательство иностранной литературы, 1962.

[2] Békéssy, A., Kis, O., Tarnay, Gy.: Taulmány R. H. Merson módszeréről, MTA. Automat. Kutató Int. Közl. **4** (1969). 3—31.

[3] Березин, И. С., Жидков, Н. П.: *Методы вычислений,* том 2. Москва, Физматгиз, 1962.

Политехнический Институт, Будапешт

(Поступило 6. 8. 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 437—439.

ON RATIONAL APPROXIMATION OF DIFFERENTIABLE FUNCTIONS

by

G. FREUD

In this paper we extend the results of our paper [1] to differentiable functions. Let us denote by AC_k the class of functions $f(x)$ defined for $x \in [0, 1]$, which are $k-1$ times continuously differentiable and for which $f^{(k-1)}(x)$ is absolutely continuous. Further let us denote by V_k the class of the functions $f(x) \in AC_k$ for which $f^{(k)}(x)$ has bounded variation $V(f^{(k)})$ in $[0, 1]$. $f \in AC_k$ implies that $f^{(k)}(x) \in L$ exists almost everywhere and

$$\Delta(f^{(k)}; h) \stackrel{\text{def}}{=} \int_0^{1-h} |f^{(k)}(x+h) - f^{(k)}(x)|\, dx \to 0 \tag{1}$$

for $h \to +0$. We introduce the usual notation

$$\omega_k(f; \delta) \stackrel{\text{def}}{=} \max_{\substack{x \in [0,\, 1-kh] \\ |h| \leqq \delta}} \left| \sum_{\nu=0}^{k} (-1)^{\nu} \binom{k}{\nu} f(x+\nu h) \right| \tag{2}$$

for the modulus of continuity of order k. We will make essential use of the following

LEMMA 1. *For every $f \in AC_l$ there exists a sequence of functions $\{\varphi_\nu(x)\}$, so that $\varphi_\nu \in V_l$*

$$|f(x) - \varphi_\nu(x)| \leqq a_l \omega_{l+1}(f; \nu^{-1}) \tag{3}$$

and

$$V(\varphi_\nu^{(l)}) \leqq b_l [\nu \Delta(f^{(l)}; \nu^{-1}) + \nu^l \omega_{l+1}(f; \nu^{-1})]. \tag{4}$$

Here a_l, b_l depend on l only.

The proof of this Lemma can be deduced from formulas (6) and (8) of the paper G. FREUD—V. POPOV [3] by setting $\varphi_n(x) = f_{l+1, n-1}(x)$, $h = \nu^{-1}$.

The optimal rational approximation of a function $g \in C[0, 1]$ we denote by

$$R_n(g) = \min_{a_i, b_i} \max_{x \in [0, 1]} \left| \frac{a_0 + a_1 x + \ldots + a_n x^n}{b_0 + b_1 x + \ldots + b_n x^n} - g(x) \right|. \tag{9}$$

From the investigations of P. SZÜSZ and P. TURÁN [4] one can conclude the existence of sequences $\{\lambda_n^{(k)}\}$ $(k = 1, 2, \ldots)$ for which

$$R_n(f) \leqq \lambda_n^{(k)} V(f^{(k)}) \qquad (f \in V_k) \tag{10}$$

holds. In [2] the author proved the upper estimate

$$\lambda_n^{(k)} \leqq c_k \frac{\log^2 n}{n^{k+1}} \tag{11}$$

and in [1] we proved the validity of the lower estimate

$$\lambda_n^{(k)} \geqq c_k n^{-k-1} \tag{12}$$

for $k=1$. The principal aim of the present paper is to extend (12) for integers $k \geqq 2$.

LEMMA 2. *For every $f \in AC_l$ and every positive integer n we have*

$$R_n(f) \leqq a_l \omega_{l+1}(f; v^{-1}) + b_l [v\Delta(f^{(l)} v^{-1}) + v^l \omega_{l+1}(f; v^{-1})] \lambda_n^{(l)}. \tag{13}$$

PROOF. Clearly

$$R_n(f) \leqq \|f - \varphi_v\| + R_n(\varphi_v)$$

and we now apply to this (3), (4) and (10).

THEOREM. *We have for $k=1, 2, \ldots$*

$$\lambda_n^{(k)} \geqq A_k n^{-k-1} \qquad (n = 1, 2, \ldots) \tag{14}$$

with absolute constants $A_1, A_2, \ldots$.

PROOF (Generalization of the construction in our paper [1], where we dealt the case $k=1$.) Let

$$f_k(x) = \sum_{v=1}^{\infty} \frac{T_{9^{2v}}(x - 1/2)}{9^{(2k+1)v}} \tag{15}$$

where $T_n(x) = \cos[n \arccos x]$ is the Chebyshev's polynomial of degree n. The partial sums of degree n approximate $f_k(x)$ with an error $O(n^{-k-1/2})$ at most, so that by Bernstein's theorem $f_k(x)$ is k-times continuously differentiable and $f_k^{(k)} \in \text{Lip } 1/2$. We conclude (see (2) and (1))

$$\omega_{k+1}(f_k; \delta) \leqq \alpha_k \delta^{k+1/2} \tag{16}$$

and

$$\Delta(f^{(k)}; h) \leqq \beta_k h^{1/2} \tag{17}$$

with positive constants α_k, β_k.

Now let N be any positive integer. At the $3 \cdot 9^{2N+1} + 1$ points

$$x_r = \frac{1}{2} + \cos \frac{r\pi}{9^{2(N+1)}} \qquad (3 \cdot 9^{2N+1} \leqq r \leqq 6 \cdot 9^{2N+1})$$

we have

$$f_k(x_r) - \sum_{v=1}^{N} \frac{T_{9^{2v}}(x - 1/2)}{9^{(2k+1)v}} = (-1)^r \sum_{v=N+1}^{\infty} \frac{1}{9^{(2k+1)v}}.$$

By Chebyshev's theorem we conclude

$$R_{9^{2N}}(f_k) = \sum_{\nu=N+1}^{\infty} \frac{1}{9^{(2k+1)\nu}} > \frac{1}{9^{(2k+1)(N+1)}}. \tag{18}$$

From Lemma 2, (16), (17) and (18) we get inserting $n=9^{2N}$, $l=k$

$$\frac{1}{9^{(2k+1)(N+1)}} \leqq a_k \alpha_k \nu^{-k-\frac{1}{2}} + b_k(\alpha_k+\beta_k)\nu^{1/2}\lambda_{9^{2N}}^{(k)}. \tag{19}$$

We take now $\nu=\varrho_k 9^{2N}$, where ϱ_k is the smallest integer satisfying $\varrho_k^{k+1/2} > 2a_k\alpha_k$ and obtain from (19)

$$\lambda_{9^{2N}}^{(k+1)} > [2\cdot 9^{2k+1} + b_k(\alpha_k+\beta_k)\varrho_k^{1/2}]^{-1}(9^{-2N})^{k+1}. \tag{20}$$

This is (14) for $n=9^{2N}$; for the other values of n we obtain (14) by a monotony argument. Q.e.d.

LITERATURE

[1] FREUD, G.: On rational approximation of absolutely continuous functions, *Studia Sci. Math. Hung.* **3** (1968), 383—386.

[2] FREUD, G.: Über die Approximation reeller Funktionen durch rationale gebrochene Funktionen. *Acta. Math. Acad. Sci. Hung.* **17** (1966), 313—324.

[3] Фройд, Г.—Попов, В.: Аппроксимация сплайн-функциями. *Studia Sci. Math. Hung.* **5** (1970).

[4] SZÜSZ P.—TURÁN, P.: On the constructiv theory of functions, II. *Studia Sci. Math. Hung.* **1** (1966).

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received September 6, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 1(970) 441—444.

ONCE MORE ON THE POISSON PROCESS

by

D. O. H. SZÁSZ

In memoriam Professor A. Rényi

Let us consider a point process on the real line and denote by $v(I)$ the number of points, which fall into the interval I. The (stationary) Poisson process can be characterized by two properties:

$\mathfrak{P}$. For any interval I

$$\mathsf{P}(v(I) = k) = \frac{(\lambda|I|)^k}{k!} e^{-\lambda|I|} \qquad (k = 0, 1, 2, \ldots)$$

where $|I|$ denotes the length of the interval I.

$\mathfrak{I}$. For disjoint intervals $I_1, \ldots, I_n$ the random variables $v(I_1), \ldots, v(I_n)$ are independent $(n = 1, 2, \ldots)$.

In 1965 I asked, whether it is true that the only condition $\mathfrak{P}$ also characterizes the Poisson process and the condition $\mathfrak{I}$ is a consequence of $\mathfrak{P}$. In 1967—68 indepently of each other SHEPP (see GOLDMAN [1]), MORAN [2] and LEE [3] proved that this is not true, they gave examples for point processes, for which $\mathfrak{P}$ is satisfied but $\mathfrak{I}$ is not. A. RÉNYI posed the question in another way, he substituted the condition $\mathfrak{P}$ with a stronger one:

$\mathfrak{P}^*$. For any set E, which is the union of a finite number of intervals,

$$\mathsf{P}(v(E) = k) = \frac{(\lambda|E|)^k}{k!} e^{-\lambda|E|} \qquad (k = 0, 1, 2, \ldots)$$

where $|E|$ denotes the Lebesgue measure of the set E. ($v(E)$ denotes the number of points, which fall into the set E.)

He proved that $\mathfrak{P}^*$ is already sufficient for the characterization of the Poisson process, i.e. $\mathfrak{I}$ is a consequence of $\mathfrak{P}^*$ [4].

Now the question arises, whether it is possible to require more than condition $\mathfrak{P}$ in another manner, namely to assume $\mathfrak{P}$ and something else, which requires less than $\mathfrak{I}$, in order to ensure the poissonity. A natural weakening of $\mathfrak{I}$ is the condition

$\mathfrak{I}_N$. For disjoint intervals $I_1, \ldots, I_n$ the random variables $v(I_1), \ldots, v(I_n)$ are independent $(n = 1, 2, \ldots, N)$.

Using SHEPP's idea we construct a point process, that is not a Poisson one, and for that the conditions $\mathfrak{P}$ and $\mathfrak{I}_N$ are satisfied.

THEOREM: *There exists a point process, for which $\mathfrak{P}$ and $\mathfrak{I}_N$ are valid, but it is not a Poisson process.*

PROOF: We construct the process only for the interval [0, 1]. If in the remaining part of the real line we take a Poisson process, independently of the already constructed process, then by the union of the two processes we get a process on the real line, for which the assertion of the theorem is true.

In the interval [0, 1] we start from a Poisson process of parameter λ. We denote the measure of this Poisson process by $\tilde{\mathsf{P}}$, and that of the process to be constructed by P. Let

$$\mathsf{P}(\nu([0,1]) = k) = \tilde{\mathsf{P}}(\nu([0,1]) = k) \left[= \frac{\lambda^k}{k!} e^{-\lambda} \right] \qquad (k = 0, 1, 2, \ldots) \tag{1}$$

and

$$\mathsf{P}(\cdot \mid \nu([0,1]) = k) = \tilde{\mathsf{P}}(\cdot \mid \nu([0,1]) = k) \tag{2}$$

only if $k \neq M = 2N+1$. It is known that if under the condition $\nu([0,1]) = k$ we take a random permutation of the k points of a Poisson process, which fall into the interval [0, 1], then the distribution of the so obtained k-dimensional vector is the same as the distribution of a vector, the components of which are independent and uniformly distributed in [0, 1] i.e. its conditional distribution function under the condition $\nu([0,1]) = k$ has the form

$$\tilde{F}_k(x_1, \ldots, x_k) = x_1 \ldots x_k$$

only if $0 \leqq x_1, \ldots, x_k \leqq 1$. From (2) it follows that for $k \neq M$

$$F_k = \tilde{F}_k. \tag{3}$$

For $k = M$ we define

$$\begin{aligned} F_M(x_1, \ldots, x_M) &= x_1 \ldots x_M + \varepsilon \cdot x_1 \ldots x_M (1-x_1) \ldots (1-x_M) \cdot \\ &\cdot \prod_{1 \leqq i < j \leqq M} (x_j - x_i) = \tilde{F}_M(x_1, \ldots, x_M) + H(x_1, \ldots, x_M) \end{aligned} \tag{4}$$

only if $0 \leqq x_1, \ldots, x_M \leqq 1$. It is easy to see that for $\varepsilon > 0$ small enough $\frac{\partial^M}{\partial x_1 \ldots \partial x_M} F_M(x_1, \ldots, x_M)$ is a probability density and so F_M is a distribution function. It is obvious that our process P is determined by (1), (2), (3) and (4), and because of (4) it is not a Poisson process.

We prove that for P the conditions $\mathfrak{P}$ and $\mathfrak{I}_N$ are satisfied. Obviously they are satisfied for $\tilde{\mathsf{P}}$, so it is sufficient to prove that for disjoint intervals $I_1, \ldots, I_N$

$$\mathsf{P}(\nu(I_1) = k_1, \ldots, \nu(I_N) = k_N) = \tilde{\mathsf{P}}(\nu(I_1) = k_1, \ldots, \nu(I_N) = k_N)$$

From the definition of P it follows that we have to prove only that

$$\begin{aligned} &\mathsf{P}(\nu(I_1) = k_1, \ldots, \nu(I_N) = k_N \mid \nu([0,1]) = M) = \\ &= \tilde{\mathsf{P}}(\nu(I_1) = k_1, \ldots, \nu(I_N) = k_N \mid \nu([0,1]) = M) \end{aligned} \tag{5}$$

The probability standing on the left hand side can be expressed by a finite sum of form

$$\sum (\pm F_M(\alpha_1, \ldots, \alpha_M))$$

where the possible values of the α's are 0, 1 and the endpoints of the intervals $I_1, \ldots, I_N$. So the difference of the two sides of (5) is equal to

$$\sum (\pm H(\alpha_1, \ldots, \alpha_M)).$$

But every term of this sum vanishes. Really, if 0 or 1 occurs among the α's this is obvious, if not, then at least two of the α's are equal, so H vanishes again.

The proof is finished.

REMARK: The theorem remains true also in the r-dimensional case. If a realisation of the above constructed process in [0, 1] is $\{\xi_1^{(1)}, \ldots, \xi_k^{(1)}\}$ and $\xi_1^{(i)}, \xi_2^{(i)}, \ldots$ $(i = 2, 3, \ldots, r)$ are independent random variables, uniformly distributed in [0, 1], and they are independent of $(\xi_1^{(1)}, \ldots, \xi_k^{(1)})$, then the points

$$(\xi_1^{(1)}, \ldots, \xi_1^{(r)})$$

$$\ldots\ldots\ldots\ldots$$

$$\ldots\ldots\ldots\ldots$$

$$(\xi_k^{(1)}, \ldots, \xi_k^{(r)})$$

form a realization of an r-dimensional point process in the unit cube, and this process is not a Poisson one. We denote by J r-dimensional intervals, by $\nu^{(r)}(J)$ the number of points, which fall into the interval J, by $\tilde{\mathsf{R}}$ the Poisson process in the r-dimensional unit cube with parameter λ and by R the r-dimensional point process defined above. We have to prove only that

$$\begin{aligned} &\mathsf{R}\left(\nu^{(r)}(J_1) = k_1, \ldots, \nu^{(r)}(J_N) = k_N \,|\, \nu([0, 1]) = M\right) = \\ &= \tilde{\mathsf{R}}\left(\nu^{(r)}(J_1) = k_1, \ldots, \nu^{(r)}(J_N) = k_N \,|\, \nu([0, 1]) = M\right) \end{aligned} \tag{6}$$

The probability on the left hand side can be expressed as a sum of terms of form

$$\mathsf{P}\left(\nu(I_1) = l_1, \ldots, \nu(I_s) = l_s, B \,|\, \nu([0, 1]) = M\right) \tag{7}$$

where the endpoints of the I's can be only the endpoints of the first components of the J's, B is an event depending only on the random variables

$$\xi_1^{(i)}, \xi_2^{(i)}, \ldots, \xi_M^{(i)} \qquad (i = 2, 3, \ldots)$$

Thus (7) equals to

$$\mathsf{P}\left(\nu(I_1) = l_1, \ldots, \nu(I_s) = l_s \,|\, \nu([0, 1]) = M\right) \mathsf{P}\left(B \,|\, \nu([0, 1]) = M\right)$$

Non we can see again that the first factor is equal to

$$\tilde{\mathsf{P}}\left(\nu(I_1) = l_1, \ldots, \nu(I_s) = l_s \,|\, \nu([0, 1]) = M\right)$$

and the validity of (6) already follows.

LITERATURE

[1] GOLDMAN, J. R.: Stochastic point processes: limit theorems, (Appendix), *Annals of Mathematical Statistics* **38** (1967), 771—779.
[2] MORAN, P. A. P.: A non-Markovian quasi-Poisson process. *Studia Sci. Math. Hung.* **2** (1967), 425—429.
[3] LEE, P. M.: Some examples of infinitely divisible point processes. *Studia Sci. Math. Hung.* **3** (1968), 219—224.
[4] RÉNYI, A.: Remarks on the Poisson process. *Studia Sci. Math. Hung.* **2** (1967), 119—224.

Eötvös L. University, Budapest

(Received September 15, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 445—456.

ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ОЦЕНКИ ПАРАМЕТРА ЗАТУХАНИЯ СТАЦИОНАРНОГО ГАУССОВСКОГО—МАРКОВСКОГО ПРОЦЕССА

М. АРАТО и А. БЕНЦУР

1. Рассматривается стационарный гауссовский марковский процесс, удовлетворяющий стохастическому дифференциальному уравнению

$$d\xi(t) = -\lambda\xi(t)dt + dw(t) \tag{1}$$

где $w(t)$ стандартный винеровский процесс, $\mathsf{M}dw = 0$, $\mathsf{M}(dw)^2 = \sigma_w^2\, dt$, $\mathsf{M}\xi(t) = 0$, и корреляционная функция $\mathsf{M}\xi(s)\xi(s+t) = \sigma_\xi^2 e^{-\lambda|t|}$, где λ называется параметром затухания и $\sigma_w^2 = 2\lambda\sigma_\xi^2$. Так как параметр σ_w^2 оценивается с вероятностью 1 по единственной реализации $\xi(t)$, $0 \leqq t \leqq T$, (см. Дуб [1]) единственным неизвестным параметром является λ. В настоящей заметке мы будем рассматривать задачу определения функции распределения оценки наибольшего правдоподобия параметра затухания. Оценками этого параметра занимаются статьи Стрибел [1] и Арато [1], [2]. В статье Арато [1] и ставилась задача определения функции распределения оценок.

Напомним, что подобная задача решалась в статье Арато [3] для двухмерного случая.

Преобразование

$$t = Tt', \quad \xi(t) = \xi'(t)\sqrt{T\sigma_w^2} \tag{2}$$

приводит общую задачу к случаю

$$T = 1, \quad \sigma_w^2 = 1$$

при этом $\lambda' = \lambda T = \varkappa$, т. е. с точностью до выбора масштабов распределение интересующих нас реализаций процесса характеризуется (при известном $\mathsf{M}\xi(t) = 0$) единственным параметром $\varkappa$. В дальнейшем мы предположим, что $\sigma_w^2 = 1$.

2. Пространство реализаций $\xi(t)$, $0 \leqq t \leqq T$, можно рассматривать как произведение числовой прямой $\xi(0)$ на пространство реализаций процесса

$$\eta(t) = \xi(t) - \xi(0).$$

Если L — обычная лебеговская мера на прямой, а W — известная условная мера Винера, тогда введём в пространстве реализаций стандартную меру $V = L \times W$. Мера P_λ процесса $\xi(t)$ в том же пространстве R_ξ абсолютно непрерывно относительно V и производное Радона—Никодима даётся плотностью

$$\frac{dP_\lambda}{dV} = \sqrt{\frac{\lambda}{\pi}} \exp\left\{-\lambda\left[s_1^2 - \frac{1}{2}T + \frac{1}{2}\lambda T s_2^2\right]\right\}, \tag{3}$$

где

$$s_1^2 = \frac{1}{2}[\xi^2(0)+\xi^2(T)], \quad s_2^2 = \frac{1}{T}\int_0^T \xi^2(t)\,dt$$

(см. Стрибел [1], Арато [4]). Формула (3) показывает, что статистики s_1^2, s_2^2 образуют достаточный набор статистик. Так как

$$\log \frac{dP_\lambda}{dV} = \log \sqrt{\frac{\lambda}{\pi}} + \frac{1}{2}\log\lambda - \lambda\left[s_1^2 - \frac{1}{2}T + \frac{1}{2}\lambda T s_2^2\right],$$

то уравнение наибольшего правдоподобия имеет вид

$$\frac{1}{2\lambda} - \left[s_1^2 - \frac{1}{2}T\right] - \lambda T s_2^2 = 0$$

т. е.

$$\lambda^2 T s_2^2 + \lambda\left[s_1^2 - \frac{1}{2}T\right] - \frac{1}{2} = 0. \tag{4}$$

Уравнение имеет единственное положительное решение

$$\hat{\lambda} = \frac{-\left[s_1^2 - \frac{1}{2}T\right] + \sqrt{\left[s_1^2 - \frac{1}{2}\right]^2 + 2Ts_2^2}}{2Ts_2^2}$$

и

$$P_\lambda\{\hat{\lambda} > z\} = P_\lambda\left\{z^2 T s_2^2 + z s_1^2 - \frac{1}{2}Tz - \frac{1}{2} < 0\right\}.$$

С помощью обозначений $z = \lambda x$, $\eta_\lambda = \lambda^2 x^2 T s_2^2 + \lambda x s_1^2$ получим

$$P_\lambda\{\hat{\lambda} > \lambda x\} = P_\lambda\left\{\eta_\lambda < \frac{1}{2} + \frac{1}{2}T\lambda x\right\} \tag{5}$$

Теорема 1. *Характеристическая функция статистик* s_1^2, Ts_2^2 *имеет вид*

$$\varphi(\alpha_1, \alpha_2) = \mathsf{M}\exp\{i\alpha_1 s_1^2 + i\alpha_2 T s_2^2\} = \tag{6}$$

$$= \frac{2\sqrt{\varkappa}\, e^{\varkappa/2}(\varkappa^2 - 2T^2 i\alpha_2)^{1/4}}{\left[(\varkappa - Ti\alpha_1 + \sqrt{\varkappa^2 - 2T^2 i\alpha_2})^2 e^{\sqrt{\varkappa^2 - 2T^2 i\alpha_2}} - (\varkappa - Ti\alpha_1 - \sqrt{\varkappa^2 - 2T^2 i\alpha_2})^2 e^{-\sqrt{\varkappa^2 - 2T^2 i\alpha_2}}\right]^{1/2}}$$

Доказательство. Рассмотрим условную характеристическую функцию

$$u(T, x) = \mathsf{M}\{\exp[i\alpha_1 s_1^2 + iT\alpha_2 s_2^2] \mid \xi(0) = x\}.$$

Функция $u(T, x)$ дифференцируемая по x бесконечно много раз (см. например Гихман—Скороход [1]) и

$$u(T+\Delta T) =$$

$$= \int_{-\infty}^{\infty} \mathsf{M}\{\mathsf{M}\{e^{i(\alpha_1 s_1^2+\alpha_2(T+\Delta T)s_2^2)}|\xi(0)=x\}|\xi(\Delta T)=x_1\}\, p\big(\xi(\Delta T)=x_1|\xi(0)=x\big)\,dx_1 =$$

$$= \int_{-\infty}^{\infty} \mathsf{M}\{\mathsf{M}\{e^{i\alpha_1\frac{\xi^2(0)-\xi^2(\Delta T)}{2}}\cdot e^{i\alpha_2\int_0^{\Delta T}\xi^2(t)\,dt}|\xi(0)=x,\,\xi(\Delta T)=x_1\}\}\,u(T,x_1)\,dx_1 =$$

$$= \frac{1}{\sqrt{2\pi\Delta T}}\int_{-\infty}^{\infty} e^{\frac{-(x_1-x+\lambda x\Delta T)^2}{2\Delta T}}\cdot$$

$$\cdot\left[u(T,x)+\frac{\partial u}{\partial x_1}\bigg|_{x_1=x}(x_1-x)+\frac{\partial^2 u}{\partial x_1^2}\bigg|_{x_1=x}\frac{(x_1-x)^2}{2}+\ldots\right]\cdot$$

$$\cdot\left\{1-\frac{i\alpha_1}{2}\left((x_1-x)^2+2x(x_1-x)\right)-\frac{\alpha_1^2}{8}\left(4x^2(x_1-x)^2+\ldots\right)+\ldots\right\}(1+i\alpha_2 x^2\Delta T)\,dx_1.$$

Отсюда при предельном переходе $\Delta T\to 0$, и используя соотношения

$$\frac{1}{\sqrt{2\pi\Delta T}}\int_{-\infty}^{\infty} e^{-\frac{(x_1-x+\lambda x\Delta T)^2}{2\Delta T}}(x_1-x)\,dx_1 = -\lambda x\Delta T$$

$$\frac{1}{\sqrt{2\pi\Delta T}}\int_{-\infty}^{\infty} e^{-\frac{(x_1-x+\lambda x\Delta T)^2}{2\Delta T}}(x_1-x)^2\,dx_1 = \Delta T+\lambda^2 x^2(\Delta T)^2$$

и старшие моменты имеют порядок $o(\Delta T)$, приходим к уравнению

$$\frac{\partial u}{\partial T} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2}+\frac{\partial u}{\partial x}[-x(\lambda+i\alpha_1)]+u\left[x^2\left(\lambda i\alpha_1+i\alpha_2-\frac{\alpha_1^2}{2}\right)-\frac{i\alpha_1}{2}\right]$$

с начальным условием $u(0, x) = e^{-i\alpha_1 x^2}$.

Функция $u_1(T, x)$, где

$$u_1(T,x) = u(T,x)\,e^{-i\alpha_1\frac{x^2}{2}}$$

удовлетворяет уравнению

$$\frac{\partial u_1}{\partial T} = \frac{1}{2}\frac{\partial^2 u_1}{\partial x^2}-\lambda x\frac{\partial u_1}{\partial x}+x^2 i\alpha_2 u_1, \tag{7}$$

$u_1(0, x) = e^{i\alpha_1 \frac{x_2}{2}}$. Умножив (7) на $e^{-\lambda x^2}$ и интегрируя его от $-\infty$ до $+\infty$ мы приходим к следующему соотношению

$$\frac{\partial v}{\partial T} = -\frac{\partial v}{\partial \gamma}[2\gamma^2 - 2\lambda\gamma + i\alpha_2] + v(\lambda - \gamma) \tag{8}$$

$$v(0, \gamma) = \left(\gamma - \frac{i\alpha_1}{2}\right)^{-1},$$

где $v(T, \gamma) = \int_{-\infty}^{\infty} e^{-\gamma x^2} u_1(T, x)\, dx$. Заметим, что характеристическая функция статистик $s_1^2, \overline{Ts_2^2}$ равна

$$\varphi(\alpha_1, \alpha_2) = v\left(T, \lambda - \frac{i\alpha_1}{2}\right) \tag{9}$$

Пусть

$$z(s, \gamma) = \int_0^{\infty} e^{-sT}(T, \gamma)\, dT,$$

тогда (8) перепишется в следующий вид

$$sz(s, \gamma) - \frac{1}{\gamma - \frac{i\alpha_1}{2}} = -\frac{\partial z(s, \gamma)}{\partial \gamma}[2\gamma^2 - 2\lambda\gamma + i\alpha_2] + z(s, \gamma)(\lambda - \gamma),$$

т. е.

$$\frac{\partial z}{\partial \gamma} = z\frac{\lambda - \gamma - s}{2\gamma^2 - 2\lambda\gamma + i\alpha_2} + \frac{1}{\left(\gamma - \frac{i\alpha_1}{2}\right)(2\gamma^2 - 2\lambda\gamma + i\alpha_2)}, \tag{10}$$

Решение уравнения (10)

$$z(s, \gamma) = \exp\left\{\int_0^{\gamma} \frac{\lambda - \gamma - s}{2\gamma^2 - 2\lambda\gamma + i\alpha_2}\right\} \cdot \left[c - \int_0^{\gamma} \frac{1}{\gamma - \frac{i\alpha_1}{2}} \exp\left\{\int_0^{\gamma} \frac{\lambda - \gamma - s}{2\gamma^2 - 2\lambda\gamma + i\alpha_2}\, d\gamma\right\} d\gamma\right].$$

Пусть $\gamma_{1,2}$ являются корнями уравнения $2\gamma^2 - 2\lambda\gamma + i\alpha_2 = 0$, т. е.

$$\gamma_{1,2} = \frac{1}{2} \pm [\lambda \sqrt{\lambda^2 - 2i\alpha_2}],$$

тогда

$$z(s, \gamma) = \exp\{a(s, \gamma_1, \gamma_2) \ln(\gamma - \gamma_1) + b(s, \gamma_1, \gamma_2) \ln(\gamma - \gamma_2)\} \times \tag{11}$$

$$\times \left[c - \int \frac{1}{\gamma - \frac{i\alpha_1}{2}} \exp\{a \ln(\gamma - \gamma_1) - b \ln(\gamma - \gamma_2)\}\, d\gamma\right],$$

где

$$a(s, \gamma_1, \gamma_2) = -\frac{1}{2}\left(1 + \frac{\lambda + \gamma_2 - s}{\gamma_1 + \gamma_2}\right), \quad b(s, \gamma_1, \gamma_2) = -\frac{1}{2}\frac{\lambda + \gamma_2 - s}{\gamma_1 + \gamma_2}.$$

Из (11) для $v(T, \gamma)$ мы получаем

$$v(T,\gamma)=\frac{(\gamma_1-\gamma_2)^{1/2}\,e^{1/2\,[\lambda T-T(\gamma_1-\gamma_2)]}}{\left\{(\gamma-\gamma_2)\left(\gamma_1-\frac{i\alpha_1}{2}\right)+(\gamma-\gamma_1)\left(\frac{i\alpha_1}{2}-\gamma_2\right)e^{-2T(\gamma_1-\gamma_2)}\right\}^{1/2}}.$$

Отсюда и из соотношения (9) получается доказательство теоремы.

Следствие. Случайная величина $\eta_\lambda(x)$ имеет характеристическую функцию — при $T=1$ —

$$\varphi_{\eta_\lambda(x)}(\alpha)=$$

$$(12)\qquad =\frac{2e^{\lambda/2}\,[1-2i\alpha x^2]^{1/4}}{\left[\left(1-i\alpha x+\sqrt{1-2i\alpha x^2}\right)^2 e^{\lambda\sqrt{1-2i\alpha x^2}}-\left(1-i\alpha x-\sqrt{1-2i\alpha x^2}\right)^2 e^{-\lambda\sqrt{1-2i\alpha x^2}}\right]^{\frac{1}{2}}}$$

что следует сразу из (6).

3. Преобразование Лапласа функции распределения случайной величины $\eta_\lambda(x)$ получим из (12) с заменой $p=-i\alpha$ и умноженем на $\frac{1}{p}$

$$F_\eta^*(p)=$$

$$(13)\qquad =\frac{2e^{\lambda/2}(1+2px^2)^{1/4}}{p\left[\left(1+px+\sqrt{1+2px^2}\right)^2 e^{\lambda\sqrt{1+2px^2}}-\left(1+px-\sqrt{1+2px^2}\right)^2 e^{-\lambda\sqrt{1+2px^2}}\right]^{1/2}}.$$

Функция распределения

$$(14)\qquad F_\eta(z)=\frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty}e^{pz}F^*(p)\,dp$$

Вычисления значений функции $F_\eta(z)$ производились непосредственно с помощью формулы обращения преобразования Лапласа. При $p=\sigma+is$ (14) имеет вид

$$F_\eta(z)=$$

$$=\frac{e^{z\sigma}}{\pi}\int_{-\infty}^{\infty}\frac{e^{isz}\,(1+2x^2\sigma+i2x^2s)^{\frac{1}{4}}e^{\lambda/2\,(1-\sqrt{1+2x^2\sigma+i2x^2s})}\,ds}{(\sigma+is)\left[\left(1+\sigma x+ixs+\sqrt{1+2x^2\sigma+i2x^2s}\right)^2-\left(1+\sigma x+ixs-\sqrt{1+2x^2\sigma+i2x^2s}\right)^2 e^{-2\lambda\sqrt{1+2x^2\sigma+i2x^2s}}\right]^{\frac{1}{2}}}$$

Так как нас интересует действительная часть этого интеграла с обозначениями

$$r=[(1+2x^2\sigma)^2+(2x^2s)^2]^{1/4},\quad \varphi=\frac{1}{2}\operatorname{arc\,tg}\frac{2x^2s}{1+2x^2\sigma}$$

мы приходим к

$$F_\eta(z)=$$

$$=\frac{e^{\sigma z}}{\pi}\int_{-\infty}^{\infty}\frac{e^{i\left(sz+\frac{\varphi}{2}-\frac{\lambda}{2}r\sin\varphi\right)}\sqrt{r}\;e^{\frac{\lambda}{2}(1-r\cos\varphi)}}{(\sigma+is)\left[(A_1+iA_2)^2-(B_1+iB_2)^2(\cos 2\lambda r\sin\varphi-i\sin 2\lambda\sin\varphi)\,e^{-2\lambda r\cos\varphi}\right]^{\frac{1}{2}}}$$

где $A_1 = A_1(s) = 1 + \sigma x + r\cos\varphi \qquad A_2 = A_2(s) = xs + r\sin\varphi$

$B_1 = B_1(s) = 1 + \sigma x - r\cos\varphi \qquad B_2 = B_2(s) = xs - r\sin\varphi$

Введём обозначения

$$\alpha_1 = A_1^2 - A_2^2 - \{(B_1^2 - B_2^2)\cos 2\lambda r\sin\varphi + 2B_1 B_2 \sin 2\lambda r\sin\varphi\} e^{-2\lambda r\cos\varphi}$$

$$\beta_1 = 2A_1 A_2 + \{(B_1^2 - B_2^2)\sin 2\lambda r\sin\varphi - 2B_1 B_2 \cos 2\lambda r\sin\varphi\} e^{-2\lambda r\cos\varphi}$$

$$r_1 = \sqrt{\alpha_1^2 + \beta_1^2} \quad \varphi_1 = \frac{1}{2}\operatorname{arc\,tg}\frac{\beta_1}{\alpha_1} \quad r_2 = s^2 + \sigma^2 \quad \varphi_2 = \operatorname{arc\,tg}\frac{s}{\sigma}$$

Действительная часть интеграла имеет вид

$$F(z) = \frac{e^{\sigma z}}{\pi}\int_{-\infty}^{\infty}\cos\left(sz + \frac{\varphi}{2} - \frac{\lambda}{2}r\sin\varphi - \varphi_1 - \varphi_2\right)\sqrt{\frac{r}{r_1 r_2}}\, e^{\frac{\lambda}{2}(1-r\cos\varphi)}\, ds.$$

Из вычислительных соображений мы взяли $\sigma = \frac{1}{z}$. Подинтегральная функция является четным, и верхнюю границу A — интеграла быбрали таким образом чтобы на интервале $(A, 2A)$ значение интеграла была меньше чем 10^{-4}. В большинстве случаев граница $A = \frac{100}{\lambda x}$ достаточна, но в некоторых случаях пришлось четыре раза увеличить эту границу. На машине УРАЛ-2 вычисление одного интеграла, с необходимой точностью, требовало 1,5—25 минут. В таблице даны значения x (и λx), являющиеся решением уравнения

$$F\left(\frac{\lambda x}{2} + \frac{1}{2}\right) = p \tag{15}$$

при данных λ и p. Значения x определились с помощью итерации.

Не вычислены значения x при $p \geqq 0{,}9$, $\lambda < 1$, так как в этом случае вычисление интегралов очень длительное и, с другой стороны, x с хорошим приближением является линейной функцией λ ($0 \leqq \lambda \leqq 1$ и $p \geqq 0{,}9$).

Теорема 2. *Случайная величина* $\frac{2\eta_\lambda}{x}$ *при* $\lambda \to 0$ *является* χ^2 *распределённой величиной.*

Доказательство. Как легко проверить при $\lambda \to 0$

$$\varphi_\eta(\alpha) \to \frac{1}{\sqrt{1 - i\alpha x}}$$

т. е. $\frac{2\eta_\lambda}{x}$ является асимптотически χ^2 распределенной величиной с одной степенью свободы

$$\lim_{\lambda\to 0} \mathrm{P}\left\{\frac{2\eta_\lambda}{x} < z^2\right\} = \sqrt{\frac{2}{\pi}}\int_0^z e^{-\frac{u^2}{2}}\, du.$$

Значения $x(0, p)$ определились по таблице χ^2 распределения и $x(0, p) = \frac{1}{z_p}$, где z_p p-квантиль χ^2 распределения.

Теорема 3. *При $\lambda \to \infty$ случайная величина η_λ будет нормально распределенной*:

$$\lim_{\lambda\to\infty} \mathsf{P}\{\hat{\lambda} < \lambda x\} = \lim_{\lambda\to\infty} P\{\hat{\lambda} < \lambda + z\sqrt{\lambda}\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}}\, du,$$

т. е. оценка $\hat{\lambda}$ асимптотически нормальна с $\mathsf{M}\hat{\lambda} = \lambda$, $\mathsf{D}^2\hat{\lambda} = \lambda$.

Доказательство сразу следует из вида характеристической функции (12).

К вычислениям приближений для маленьких и больших значений λ мы вернёмся на другом месте. Заметим только, что нормальное приближение с хорошей точностью имеет место только при $\lambda > 10\,000$. χ^2 приближение ведет к хорошим результатам при $\lambda < 0{,}1$.

Нами установлено, что функция распределения $F_\lambda(z)$ при фиксированном z представляет собой монотонно возрастающую непрерывную функцию от λ $(0 \leqq \lambda < \infty)$, принимающую значения от 0 до 1. Таким образом для неизвестного параметра λ можно построить доверительные границы. В таблице даны z_p квантили, решение уравнения

$$F_\lambda(z) = P_\lambda\{\hat{\lambda} > z\} = p$$

и установлено, что квантиль z_p оценки $\hat{\lambda}$ монотонно возрастающая функция от λ. Решение уравнения

$$F_\lambda(\hat{\lambda}) = p \qquad (0 < p < 1)$$

(относительно λ) $\lambda = \lambda_p(\hat{\lambda})$ представляет собой нижний доверительный предел, причём коэффициент доверия равен $1 - p$. $\lambda_p(\lambda)$ является обратной функции $z_p(\lambda)$. Для отыскания доверительного предела $\lambda_p(\hat{\lambda})$ следует решить уравнение $z_p(\lambda) = \hat{\lambda}$ с помощью обратной интерполяции этой таблицы.

4. В заключение заметим, что наша задача тесно связана с задачей оценки параметров гауссовского марковского процесса с дискретным временем. Рассматривая процесс $\xi(t)$ с непрерывным временем только в дискретные моменты мы приходим к реализации $\xi(\Delta T)$, $\xi(2\Delta T)$, ..., $\xi(N\Delta T)$ $\left(\Delta T = \frac{T}{N}\right)$ дискретного стационарного гауссовского процесса, где процесс $\xi(k\Delta T)$ удовлетворяет уравнению

$$\xi(k\Delta T) = \varrho\xi\big((k-1)\Delta T\big) + \varepsilon(k\Delta T),$$

с параметрами $\varrho = e^{-\lambda\Delta T}$, $\sigma_\varepsilon^2 = (1-\varrho^2)\sigma_\xi^2 = \frac{(1-e^{-2\lambda\Delta T})}{2\lambda}\sigma_w^2$. При известном σ_ε^2 оценка $\hat{\varrho}$ наибольшего правдоподобия параметра ϱ обладает со следующими свойствами.

Теорема 4. *При фиксированном T и $N \to \infty$ (т. е. $\varrho \to 1$) оценка наибольшего правдоподобия $\hat{\varrho}$ имеет предельное распределение, определенное с соотношением*

$$\lim_{N\to\infty} \mathsf{P}\{\hat{\varrho} < x\} = \mathsf{P}\{e^{-\hat{\lambda}T} < x\} = \mathsf{P}\{\hat{\lambda}T > -\log x\},$$

где $\hat{\varkappa} = \hat{\lambda}T$ является оценкой наибольшего правдоподобия параметра затухания процесса с непрерывным временем.

Доказательство сразу следует из «принципа инвариантности» для марковских процессов (см. например Гихман—Скороход [2]), другое доказательство можно найти в статье Арато [3].

Теорема 5. *Оценка наибольшего правдоподобия $\hat{\varrho}$ параметра ϱ имеет асимптотически нормальное распределение с математическим ожиданием ϱ и с дисперсией $\dfrac{1-\varrho^2}{1-N}$, при $N(1-\varrho^2) \to \infty$.*

Доказательство следует из теорем 3. и 4., так как в этом случае $\varkappa = -N \log \varrho \to \infty$.

Теоремы 4. и 5. дают частичный ответ на вопрос А. Н. Колмогорова относительно оценок параметров стационарного гауссовского марковского процесса, поставленный ещё в 1948 г. Полный ответ, когда все три параметра процесса являются неизвестными, будет дан на другом месте.

Другое доказательство теоремы 5. можно найти в статьях Линник [1], Лувсанцерен [1], T. W. Anderson [1], Арато [3]. В последней статье показаны пределы равномерной асимптотической нормальности.

В дальнейшем рассмотрим некоторые следствия теоремы 4. Многие авторы предлагали (см. например White [1], [2], или Orcutt и Winokur [1], где дана подробная литература) асимптотику для оценки наибольшего правдоподобия $\hat{\varrho}$ параметра ϱ. По предложению White доверительные границы параметра ϱ даются с помощью t-распределения:

$$\hat{\varrho} - t_p \sqrt{\frac{1-\hat{\varrho}^2}{N+1}} < \varrho < \hat{\varrho} + t_p \sqrt{\frac{1-\hat{\varrho}^2}{N+1}}$$

где t_p p-квантиль t-распределения с $N+1$ степенями свободы.

В таблице 1. даются доверительные границы (при $p = 0{,}05$) для параметра ϱ при $\hat{\varrho} = 0{,}5$, 0,9, 0,99, и $N = 10$, 20, 40 по формуле White, по нормальному приближению (на основе теоремы 5.) и по приближению с непрерывным временем ($\hat{\varkappa} = -N \log \hat{\varrho}$).

По данным таблицы 1. видно, что ни t-распределение ни нормальное распределение не является хорошим приближением, даже при $N = 40$, особенно для верхней доверительной границы. Заранее известно, что $\varrho < 1$.

В статье Orcutt и Winokur даются таблицы распределения оценки $\hat{\varrho}$ на основе 10 тысячных экспериментов на Э. В. М., при разных N. Для сравнения в таблице 2. мы здесь приводим некоторые результаты этих экспериментов и данные, вычисленные по таблице 3. В таблице даны верхние и нижние квантили оценки $\hat{\varrho}$, т. е. при $p = 0{,}9$ и $p = 0{,}1$. Границы по методу Монте-Карло

вычисляются из эмпирических среднего и дисперсии с нормальным приближением.

Из данных таблицы 2. видно, что доверительные интервалы, полученные по приближению с непрерывным временем будут более узким, чем настоящие. Совпадение результатов Монте-Карло с результатами теоретических данных можно считать удовлетворительным.

В таблице 3. даны значения z, в скобках $x = \frac{z}{\lambda}$, для которых

$$P_\lambda\{\hat{\lambda} > z\} = p$$

Таблица 1

Доверительные границы, полученные по разным приближениям

ϱ	N	Приближение t-распределение			Нормальное приближение			Приближение с непр. временем		
		10	20	40	10	20	40	10	20	40
0,5	Н.	0,03	0,17	0,27	0,02	0,17	0,29	0,301	0,340	0,375
	В.	0,97	0,83	0,73	0,98	0,83	0,71	0,914	0,790	0,684
0,9	Н.	0,66	0,73	0,79	0,66	0,73	0,78	0,778	0,795	0,819
	В.	1,14	1,07	1,01	1,14	1,07	1,02	0,999	0,998	0,995
0,99	Н.	0,91	0,94	0,95	0,91	0,94	0,95	0,963	0,968	0,972
	В.	1,07	1,04	1,03	1,07	1,04	1,03	0,99999	0,99999	0,99998

Таблица 2

Квантили оценки наибольшего правдоподобия

ϱ \ N	Метод Монте-Карло			Приближение с непр. временем		
	10	20	40	10	20	40
0,5 э. ср.	0,232	0,347	0,441			
дисп.	0,121	0,046	0,021			
Н. кв.	—0,19	0,12	0,28	0,23	0,30	0,37
В. кв.	0,65	0,64	0,60	0,66	0,66	0,61
0,9 э. ср.	0,482	0,680	0,796			
дисп.	0,109	0,039	0,013			
Н. кв.	0,07	0,42	0,77	0,597	0,694	0,767
В. кв.	0,91	0,94	0,93	0,955	0,951	0,944

Таблица 3

Квантили случайной величины $\hat{\lambda}(T=1)$

λ \ p	0,001	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,999
0	0 (637 000)	0 (6370)	0 (1020)	0 (255,0)	0 (63,60)	0 (0,369)	0 (0,260)	0 (0,199)	0 (0,151)	0 (0,092)
0,01	10,60 (1060)	4,232 (423,2)	2,274 (227,4)	1,170 (117,0)	0,4734 (47,34)					
0,05	11,195 (263,9)	6,330 (126,6)	4,0065 (80,13)	2,5130 (50,26)	1,3375 (26,75)					
0,1	14,38 (143,8)	7,344 (73,44)	4,879 (48,79)	3,268 (32,68)	1,908 (19,08)					
0,2	15,664 (78,32)	8,468 (72,34)	5,902 (29,51)	4,154 (20,77)	2,624 (13,12)					
0,3	16,488 (54,96)	9,207 (30,69)	6,561 (21,87)	4,746 (15,82)	3,120 (10,40)					
0,4	17,080 (72,70)	9,756 (24,39)	7,080 (17,70)	5,208 (13,02)	3,517 (8,793)					
0,5	17,670 (35,34)	10,230 (20,46)	7,515 (15,03)	5,605 (11,21)	3,8610 (7,722)	0,2085 (0,417)	0,1510 (0,302)			
0,6	18,108 (30,18)	10,638 (17,73)	7,896 (13,16)	5,9532 (9,922)	4,1676 (6,946)					
0,7	18,522 (26,46)	11,011 (15,73)	8,239 (11,77)	6,2713 (8,959)	4,4471 (6,353)					
0,8	18,896 (23,62)	11,360 (14,20)	8,560 (10,70)	6,5648 (8,206)	4,7103 (5,887)					
0,9	19,260 (21,40)	11,682 (12,98)	8,8587 (9,843)	6,8409 (7,601)	4,9446 (5,494)					
1	19,60 (19,60)	11,98 (11,98)	9,140 (9,140)	7,103 (7,103)	5,188 (5,188)	0,445 (0,445)	0,332 (0,332)	0,269 (0,269)	0,205 (0,205)	0,130 (0,130)
1,5	21,060 (14,04)	13,3080 (8,872)	10,3845 (6,923)	8,2590 (5,506)	6,2325 (4,155)	0,7005 (0,467)	0,5325 (0,355)	0,4275 (0,285)	0,3360 (0,224)	0,2145 (0,143)
2	22,32 (11,16)	14,462 (7,231)	11,426 (5,713)	9,272 (4,636)	7,156 (3,278)	0,972 (0,486)	0,750 (0,375)	0,606 (0,303)	0,480 (0,240)	0,310 (0,155)
2,5	23,4750 (9,390)	15,515 (6,206)	12,4575 (4,983)	10,2050 (4,082)	8,0100 (3,204)	1,2575 (0,503)	0,9850 (0,394)	0,8000 (0,320)	0,6500 (0,256)	0,4125 (0,165)
3	24,342 (8,114)	16,506 (5,502)	13,389 (4,463)	11,082 (3,694)	8,811 (2,937)	1,557 (0,519)	1,233 (0,411)	1,111 (0,337)	0,807 (0,269)	0,525 (0,175)
3,5	25,5850 (7,310)	17,4440 (4,984)	14,2800 (4,080)	11,9105 (3,403)	9,5970 (2,742)	1,8655 (0,533)	1,4910 (0,426)	1,2355 (0,353)	0,9975 (0,285)	0,6405 (0,183)
4	26,576 (6,644)	18,352 (4,588)	15,136 (3,784)	12,732 (3,183)	10,352 (2,588)	2,180 (0,545)	1,760 (0,440)	1,468 (0,367)	1,192 (0,298)	0,776 (0,194)
4,5	27,5355 (6,119)	19,2285 (4,273)	15,9705 (3,549)	13,5225 (3,005)	11,0835 (2,463)	2,5020 (0,556)	2,0430 (0,454)	1,7100 (0,380)	1,3905 (0,309)	0,9135 (0,203)
5	28,470 (5,694)	20,090 (4,018)	16,795 (3,359)	14,395 (2,879)	11,755 (2,351)	2,835 (0,567)	2,325 (0,0465)	1,965 (0,393)	1,615 (0,323)	1,070 (0,214)

Таблица 3

λ \ p	0,001	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,999
5,5	29,3865 (5,343)	20,9275 (3,805)	17,5835 (3,197)	15,0535 (2,737)	12,5125 (2,275)	3,1680 (0,576)	2,6235 (0,477)	2,4640 (0,448)	1,8315 (0,333)	1,2265 (0,223)
6	30,318 (5,053)	21,750 (3,625)	18,366 (3,061)	15,792 (2,632)	13,282 (2,212)	3,510 (0,585)	2,922 (0,487)	2,490 (0,415)	2,061 (0,347)	1,392 (0,232)
6,5	31,1610 (4,794)	22,5615 (3,471)	19,1295 (2,943)	16,5295 (2,543)	13,8905 (2,137)	3,8545 (0,593)	3,2305 (0,497)	2,7690 (0,426)	2,3140 (0,356)	1,5730 (0,242)
7	32,025 (4,575)	23,359 (3,337)	19,894 (2,842)	17,248 (2,464)	14,574 (2,082)	4,207 (0,601)	3,542 (0,506)	3,052 (0,436)	2,555 (0,365)	1,764 (0,252)
7,5	32,8882 (4,385)	24,1500 (3,220)	20,6475 (2,753)	17,9550 (2,394)	15,2475 (2,033)	4,5600 (0,608)	3,8550 (0,514)	3,3375 (0,445)	2,8200 (0,376)	1,9575 (0,261)
8	33,728 (4,216)	24,920 (3,115)	21,384 (2,673)	18,672 (2,334)	15,912 (1,989)	4,920 (0,615)	4,176 (0,522)	3,632 (0,454)	3,088 (0,386)	2,168 (0,271)
8,5	34,5610 (4,066)	25,6870 (3,022)	22,1170 (2,602)	19,3715 (2,279)	16,5750 (2,950)	5,2700 (0,620)	4,5050 (0,530)	3,9270 (0,462)	3,3490 (0,394)	2,3630 (0,278)
9	35,388 (3,932)	26,451 (2,939)	22,689 (2,521)	20,070 (2,230)	17,226 (1,914)	5,643 (0,627)	4,842 (0,538)	4,230 (0,470)	3,618 (0,402)	2,592 (0,288
9,5	36,1855 (3,809)	27,1700 (2,860)	23,5790 (2,482)	20,7480 (2,184)	17,8790 (1,882)	6,0135 (0,633)	5,1870 (0,546)	4,5315 (6,477)	3,8950 (0,410)	2,812 (0,296)
10	37,04 (3,704)	27,55 (2,755)	24,28 (2,428)	21,47 (2,147)	18,53 (1,853)	6,38 (0,638)	5,50 (0.550)	4,84 (0,484)	4,20 (0,420)	3,04 (0,304)
20	52,200 (2,610)	42,040 (2,102)	37,800 (1,890)	34,4360 (1,7218)	30,8960 (1,5448)	14,1780 (0,7089)	12,7140 (0,6357)	11,580 (0,579)	10,380 (0,519)	8,320 (0,416)
30	66,270 (2,209)	55,230 (1,841)	50,520 (1,684)	46,7370 (1,5579)	42,7080 (1,4236)	22,4310 (0,7477)	20,4960 (0,6832)	18,960 (0,632)	17,340 (0,578)	14,400 (0,480)
40	79,800 (1,995)	67,920 (1,698)	62,800 (1,570)	58,6800 (1,4670)	54,2320 (1,3558)	30,9400 (0,7735	28,5920 (0,7148)	26,720 (0,668)	24,680 (0,617)	21,000 (0,525)
50	92,900 (1,858)	80,3200 (1,6064)	74,8400 (1,4968)	70,3950 (1,4079)	65,5800 (1,3116)	39,6150 (0,7923)	36,9000 (0,7380)	34,7100 (0,6972)	32,350 (0,647)	27,950 (0,559)
60	105,780 (1,763)	92,520 (1,542)	86,6820 (1,4447)	81,9540 (1,3659)	76,7940 1,2799)	48,4140 (0,8069)	45,3660 (0,7561)	42,8880 (0,7148)	40,200 (0,670)	30,300 (0,585)
70	118,370 (1,691)	104,510 (1,493)	98,3770 (1,4054)	93,3800 (1,3340)	87,9130 (1,2559)	57,3020 (0,8186)	53,9560 (0,7708)	51,2120 (0,7316)	48,2300 (0,689)	42,490 (0,607)
80	130,800 (1,635)	116,40 (1,455)	109,960 (1,3745)	104,7120 (1,3089)	98,9600 (1,2370)	66,2722 (0,8284)	62,6240 (0,7828)	59,3320 (0,7454)	56,400 (0,0705)	50,080 (0,326)
90	143,100 (1,590)	128,160 (1,424)	121,4550 (1,3495)	115,9650 (1,2885)	109,9350 (1,2215)	75,2940 (0,8366)	71,3880 (0,7932)	68,1570 (0,7573)	64,620 (0,718)	57,870 (0,643)
100	155,32 (1,6332)	139,79 (1,3979)	132,86 (1,3286)	127,15 (1,2715)	120,86 (1,2086)	84,37 (0,8737)	80,21 (0,8021)	76,8 (0,768)	73,0 (0,730)	65,7 (0,657)
500	609,000 (1,218)	580,000 (1,160)	567,0500 (1,1341)	555,800 (1,1116)	543,15 (1,0833)	462,05 (0,9241)	451,4500 (0,9029)	442,600 (0,8852)	432,600 (0,8652)	412,00 (0,824)
1000	1149 (1,149)	1110,6 (1,1106)	1092,6 (1,0926)	1077,3 (1,0773)	1060,00 (1,0600)	945,3 (0,9453)	929,9 (0,9299)	917,00 (0,9170)	902,3 (0,9023)	872,00 (0,872)
10 000	10477 (1,0477)	10336 (1,03360)	10282,1 (1,02821)	10236,3 (1,02333)	10183,9 (1,01839)	9821,4 (0,98214)	9771,1 (0,97711)	9727,6 (0,97276)	9377,5 (0,93775)	9274 (0,9574)

ЛИТЕРАТУРА

ANDERSON, T. W.: On asymptotic distributions of estimates of parameters of stochastic difference equations. *Ann. M. Stat.*, (1959), 676—87.

Арато М.: [1] Оценка параметров стационарного гауссовского марковского процесса *Д. А. Н.* **145,** Но. 1, (1959) 13—16.

[2] Folytonos állapotú Markov folyamatok statisztikai vizsgálatáról. II. *MTA III. Oszt. Közl.* **14** (1964), 137—159.

[3] Folytonos állapotú Markov folyamatok statisztikai vizsgálatáról. III. *MTA III. Oszt. Közl.* **14** (1964), 317—330.

[4] Вычисление доверительных границ для параметра «затухания» комплексного стационарного гауссовского марковского процесса, *Теория вероятностей и её; прим.* **13** (1968) 326—333.

[5] Точные формулы для плотностей мер элементарных гауссовских процессов, *Studia Sci. Math.* **5** (1970).

DOOB, J. L.: *Вероятностные процессы,* (1953).

Гихман И. И., Скороход, А. В.: [1] *Стохастические дифференциальные уравнения,* Киев, (1968).

[2] *Введение в теорию случайных процессов,* Москва, (1965).

Линник, Ю. В. Об одном вопросе статистики зависимых наблюдений, *Известия А.Н. СССР* **14** (1950) 501—522.

Лувсанцерен, Ш.: Оценки наибольшего правдоподобия и доверительные множества для неизвестных параметров стационарного процесса марковского типа, *Д. А. Н.* (1954), 723—726.

ORCUTT, G. H.—WINOKUR, H. S.: First order autoregression; inference estimation and prediction, *Econometrica* (1969), 1—14.

STRIEBEL CH.: Densities for stochastic processes. *Ann. Math. Stat.* **30** (1959), 559—567.

WHITE, J. S.: [1] The limiting distribution of the serial correlation coefficient, *Ann. Math. Stat.* **29** (1958), 1188—1197.

[2] A *t*-test for the serial correlation coefficient, *Ann. Math. Stat.* **28** (1957), 1046—48.

[3] Approximate moments for the serial correlation coefficients, *Ann. Math. Stat.* **27** (1956), 798.

Вычислительный Центр А. Н. Венгерской
Народной Республики Будапешт

(Поступила 3. X. 1969)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 457—460.

SHORT NOTE ON THE MOST INFORMATIVE DECISION

by

T. O. H. NEMETZ

Introduction

The aim of this short note is to prove exactly that the most informative decision, obtained in a former paper [2] is always a non-randomized one, except for pathological situations. In this investigation the BAYES' point of view is accepted and LINDLEY's concept [1] of the information provided by an experiment (with respect to a parameter) is used.

We shall deal with alternative hypotheses. Let $\{\Omega, \mathscr{M}, \mathsf{P}\}$ be a probability space, and let ϑ be a random variable on it, with values 0 and 1, and let us denote by w_i the probability $\mathsf{P}(\vartheta = i)$, $i = 0, 1$. We can observe a random variable ξ which takes its values in an (abstract) measurable sample space $\{X, \mathscr{B}\}$; for every $B \in \mathscr{B}$ we know the probabilities $\mathsf{P}_i(B) = \mathsf{P}\{\omega : \xi(\omega) \in B \mid \vartheta = i\}$, $i = 0, 1$.

We can also observe a random variable η with values from a measurable space $\{Y, \mathscr{Y}\}$, where η is independent of the pair (ϑ, ξ), and, in addition, for every $0 \leqq \gamma \leqq 1$ there exists a set $A_\gamma \in \mathscr{Y}$ such that

$$\gamma = \mathsf{P}\{\eta \in A_\gamma\}$$

holds.

Let $d = d(x, y)$ be a 0—1 valued measurable function on $\{X * Y, \mathscr{B} * \mathscr{Y}\}$. Such a function is called a decision. (For the sake of simplicity we assume that $\xi(\Omega) = X$.)

In paper [2] the following information-theoretical variant of the Neyman—Pearson lemma was proved:

THEOREM 1. *There exist constants $c > 0$ and $0 \leqq \gamma \leqq 1$ such that the decision $\delta(\xi, \eta)$ defined by*

$$\delta(\xi, \eta) = \begin{cases} 0 & \text{if} \quad \dfrac{d\mathsf{P}_0}{dQ}(\xi) > c \cdot \dfrac{d\mathsf{P}_1}{dQ}(\xi). \\ 0 & \text{if} \quad \dfrac{d\mathsf{P}_0}{dQ}(\xi) = c \cdot \dfrac{d\mathsf{P}_1}{dQ}(\xi), \quad \text{and} \quad \eta \in A_\gamma \\ 1 & \text{if} \quad \dfrac{d\mathsf{P}_0}{dQ}(\xi) = c \cdot \dfrac{d\mathsf{P}_0}{dQ}(\xi) \quad \text{and} \quad \eta \notin A_\gamma \\ 1 & \text{if} \quad \dfrac{d\mathsf{P}_0}{dQ}(\xi) < c \cdot \dfrac{d\mathsf{P}_1}{dQ}(\xi) \end{cases}$$

contains maximal information concerning ϑ, where Q is an arbitrary measure on $\mathscr{B}$,

with respect to that the measures P_0 *and* P_1 *are both absolutely continuous, and* $\frac{d\mathsf{P}_k}{dQ}$ *denotes the Radon—Nikodym derivative.*

It is well-known that the information contained in the pair (ξ, η) with respect to ϑ is the same as that in the random variable ξ alone. It is intuitively evident that a similar statement is valid for the decision but one must prove this. This is that we shall do now.

Existence of a non-randomized most informative decision

THEOREM 2. *There exists a most informative decision, which is not randomized i.e. the constant* γ *in theorem* 1 *is* 0 *or* 1.

PROOF: Let $\mathsf{I}(\vartheta, d)$ denote the amount of information contained in the decision d concerning the parameter ϑ. Following Lindley, we can write

$$\mathsf{I}(\vartheta, d) = \mathsf{H}(\vartheta) - \mathsf{E}\{\mathsf{H}(\vartheta|d)\}$$

where $\mathsf{H}(\vartheta)$ is the Shannon-entropy of the parameter ϑ, and $\mathsf{H}(\vartheta|d)$ denotes the conditional entropy of ϑ given the value of d. ($\mathsf{E}\{\cdot\}$ denotes the expectation.)

It is easy to obtain the following identity [3]:

$$\mathsf{I}(\vartheta, d) = h(w_0 y_0 + w_1 y_1) - w_0 h(y_0) - w_1 h(y_1)$$

where

$$y_i = \mathsf{P}\{d=0|\vartheta=i\}, \qquad i=0, 1$$

and

$$h(p) = \begin{cases} -p\log p - (1-p)\log(1-p) & \text{if} \quad 0 < p < 1 \\ 0 & \text{if} \quad p(1-p) = 0 \end{cases}$$

In the case of the most informative decision given in theorem 1, there exist non-negative numbers p_i and r_i, $p_i + r_i \leqq 1$, $i=0, 1$ such that

$$y_i = p_i + \gamma r_i,$$

where

$$p_i = \mathsf{P}_i\left\{\omega: \frac{d\mathsf{P}_0}{dQ} > c\frac{d\mathsf{P}_1}{dQ}\right\} \text{ and } r_i = \mathsf{P}\left\{\omega: \frac{d\mathsf{P}_0}{dQ} = c\frac{d\mathsf{P}_1}{dQ}\right\}$$

depend only on c, thus by the theorem they are constant. We have to prove that the function

$$i(\gamma) = h[w_0 p_0 + w_1 p_1 + \gamma(w_0 r_0 + w_1 r_1)] - w_0 h(p_0 + \gamma r_0) - w_1 h(p_1 + \gamma r_1)$$

can take on its maximal value when γ varies from 0 to 1 only at the points $\gamma=0$ or $\gamma=1$. We shall prove the convexity of the function $i(\gamma)$ or, what is the same, $\frac{d^2 i(\gamma)}{d\gamma^2} \geqq 0$ in [0, 1], where the equality holds only in the pathological cases $y_0 = y_1$ or $w_0 w_1 = 0$ or in the case $r_0 = r_1 = 0$.

It is routine work to obtain

$$\frac{d^2 i(\gamma)}{d\gamma^2} = -(w_0 r_0 + w_1 r_1)^2 \times$$

$$\times \left\{ \frac{1}{1 - w_0 p_0 - w_1 p_1 - \gamma(w_0 r_0 + w_1 r_1)} + \frac{1}{w_0 p_0 + w_1 p_1 + \gamma(w_0 r_0 + w_1 r_1)} \right\} +$$

$$+ w_0 r_0^2 \left\{ \frac{1}{1 - p_0 - \gamma r_0} + \frac{1}{p_0 + \gamma r_0} \right\} + w_1 r_1^2 \left\{ \frac{1}{1 - p_1 - \gamma r_1} + \frac{1}{p_1 + \gamma r_1} \right\}$$

It suffices to prove the inequalities

(1) $$\frac{w_0 r_0^2}{p_0 + \gamma r_0} + \frac{w_1 r_1^2}{p_1 + \lambda r_1} \geqq \frac{(w_0 r_0 + w_1 r_1)^2}{w_0(p_0 + \gamma r_0) + w_1(p_1 + \gamma r_1)}$$

and

(2) $$\frac{w_0 r_0^2}{1 - p_0 - \gamma r_0} + \frac{w_1 r_1^2}{1 - p_1 - \gamma r_1} \geqq \frac{(w_0 r_0 + w_1 r_1)^2}{w_0(1 - p_0 - \gamma r_0) + w_1(1 - p_1 - \gamma r_1)};$$

these inequalities follow from the convexity of the function $f(x, y) = \frac{x^2}{y}$, but we prefer to give a direct proof now.

It is easy to see that (for any value of the constant c in the theorem 1) neither y_0 nor y_1 equals 0 in $0 < \gamma < 1$. We can obviously assume that one of r_0 and r_1, say r_1, is not 0, and $w_0 \cdot w_1 \neq 0$.

Introducing the new variables $y = \frac{y_0}{y_1}$, $r = \frac{r_0}{r_1}$ and $w = \frac{w_0}{w_1}$ we arrive at the equivalent form of (1):

(3) $$\frac{w \cdot r^2}{y} + 1 \geqq \frac{(wr + 1)^2}{wy + 1};$$

this is equivalent to

$$w(y - r)^2 \geqq 0.$$

It is easy to see, that here the equality holds only, if $p_0 = p_1 = 0$.

(2) can be reduced to (3) in a perfectly similar way, with $y = \frac{1 - y_0}{1 - y_1}$, and the equality holds now only if $1 - p_0 = 1 - p_1 = 0$. Thus we have proved our assertion.

We remark that a similar statement holds for the case when ϑ has any finite number of different possible values.

REFERENCES

[1] LINDLEY, D. V.: On the Measure of Information Provided by an Experiment, *Annals of Math. Stat.* **27** (1956), 986—1005.
[2] NEMETZ, T.: Maximális információt tartalmazó döntésfüggvények, *MTA III. Oszt. Közl.* XVII. (1967), 454—465.
[3] NEMETZ, T.: Information Theory and Testing Hypotheses, *Proc. of Coll. on Inf. Theory,* Debrecen, 1967. 283—294.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received October 27, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 461—470.

ON THE MAXIMAL NUMBER OF EDGES OF CRITICAL k-CHROMATIC GRAPHS*

by

B. TOFT

Abstract. It is proved for all integers k, $k \geqq 4$, that there exists a positive constant c_k such that $c_k n^2 < f_k(n)$, where $f_k(n)$ denotes the largest integer for which there exists a critical k-chromatic graph with n vertices and $f_k(n)$ edges. This answers a question of P. ERDŐS.

0. The graphs considered in this paper are finite, undirected, without loops and multiple edges. If Γ is a graph then $V(\Gamma)$ denotes the set of vertices of Γ and $E(\Gamma)$ denotes the set of edges of Γ. The valency of a vertex x in Γ is the number of vertices of Γ joined to x by edges in Γ. If $e \in E(\Gamma)$ then $\Gamma - e$ denotes the graph obtained from Γ by deleting the edge e, but no vertex from Γ. An edge e joining the two vertices x and y is denoted (x, y). If $e = (x, y)$ and $x \in M_1 \subseteq V(\Gamma)$ and $y \in M_2 \subseteq V(\Gamma)$, then e is a $(M_1) \times (M_2)$-edge. If $M \subseteq V(\Gamma)$, then $\Gamma(M)$ denotes the subgraph of Γ spanned by M, i.e.:

$$V(\Gamma(M)) = M$$

$$E(\Gamma(M)) = \{e \in E(\Gamma) | e \quad \text{is a} \quad (M) \times (M)\text{-edge}\}.$$

If $E(\Gamma(M)) = \emptyset$ then M is said to be an independent set of vertices in Γ. Two graphs Γ_1 and Γ_2 are said to be disjoint if $V(\Gamma_1) \cap V(\Gamma_2) = \emptyset$. If S is a set then $|S|$ denotes the number of elements in S.

1. Let k be an integer $\geqq 1$. A graph Γ is said to be k-colourable if $V(\Gamma)$ can be divided into k mutually disjoint (colour) classes, such that no two vertices in the same class are joined by an edge. The smallest integer k for which a graph Γ is k-colourable is called the chromatic number of Γ. A connected k-chromatic graph Γ is called critical k-chromatic if for every edge $e \in E(\Gamma)$ $\Gamma - e$ is $(k-1)$-colourable.

There exist no critical k-chromatic graphs with $k+1$ vertices ([2] Theorem 6). A complete k-graph (i.e. a graph with k vertices and $\frac{1}{2} \cdot k \cdot (k-1)$ edges) is a critical k-chromatic graph, and for $k = 1$ and $k = 2$ it is the only one. A graph is critical 3-chromatic if and only if it is an odd circuit, hence any critical 3-chromatic graph has the same number of vertices and edges.

Let us now suppose that $k \geqq 4$ and let n be an integer such that $n \geqq k$ and $n \neq k+1$. G. A. DIRAC ([3] Theorem 2, see also [7] Theorem 11. 7. 5) proved that there exists a critical k-chromatic graph with n vertices. Let $f_k(n)$ denote the largest integer for which there exists a critical k-chromatic graph with n vertices and $f_k(n)$ edges.

* This work was carried out during the authors visit at the Hungarian Academy of Sciences, Budapest, autumn 1969.

P. TURÁN ([8], [9]) proved that a graph with n vertices and more than

$$\frac{1}{2}\frac{k-2}{k-1}(n^2-r^2)+\binom{r}{2}$$

edges contains a complete k-graph, where $n \equiv r \pmod{k-1}$ and $1 \leqq r \leqq k-1$. He also proved that a graph with n vertices and precisely the above mentioned number of edges either contains a complete k-graph or is $(k-1)$-chromatic. Hence:

(I) *If* $n \geqq k+2 \geqq 6$, *then* $f_k(n) < \frac{1}{2}\frac{k-2}{k-1}n^2$.

Using other forbidden subgraphs than the complete graphs and using other extremal-results than the theorem of TURÁN, M. SIMONOVITS obtained very recently an improvement of the constant $\frac{1}{2}\frac{k-2}{k-1}$ in this upper bound for $f_k(n)$ (oral communication).

P. ERDŐS asked this question: Does there for a given k, $k \geqq 4$, exist a positive constant c_k such that

$$c_k n^2 < f_k(n)$$

for infinitely many values of n? G. A. DIRAC [1] proved the existence of the constant c_k for $k \geqq 6$. The proof was based on the following construction. Let C_1 and C_2 denote two odd disjoint circuits of equal length. The graph Γ obtained from C_1 and C_2 by joining all vertices of C_1 to all vertices of C_2 by edges is critical 6-chromatic. If n denotes the number of vertices of Γ then the number of edges is $\frac{1}{4}n^2+n$.

For the cases $k=4$ and $k=5$ the question of P. ERDŐS remained unsolved for a long time [5]. The purpose of the present paper is to prove by explicit constructions that also in these cases there exists a positive constant c_k such that $c_k n^2 < f_k(n)$ for infinitely many values of n. This result shows that there exist critical 4- and 5-chromatic graphs of arbitrary high genus. We shall prove:

THEOREM 1. *Let* $k \geqq 4$. *Then there exists a positive constant* c_k *such that for all* n, $n \geqq k$ *and* $n \neq k+1$,

$$c_k n^2 < f_k(n).$$

THEOREM 2. *Let* $k \geqq 4$, $k = 3q+r$, *where* $r=0$, 1 *or* 2. *Then for infinitely many values of* n *the following inequality holds:*

a) $$\frac{q-1}{2q}n^2 < f_k(n) \qquad \text{if} \quad r=0$$

b) $$\frac{7q-6}{14q+2}n^2 < f_k(n) \qquad \text{if} \quad r=1$$

c) $$\frac{23q-15}{46q+16}n^2 < f_k(n) \qquad \text{if} \quad r=2.$$

Theorem 2a) is due to G. A. DIRAC and P. ERDŐS [5]. Theorem 2b) and c) disprove a conjecture of P. ERDŐS [5]. In the proof of Theorem 2 we shall for each

of the three inequalities explicitely give an infinite set of n-values for which the inequality holds. In the cases $k=4$ and $k=5$ Theorems 2b) and c) state

$$\frac{1}{16}n^2 < f_4(n) \quad \text{and} \quad \frac{4}{31}n^2 < f_5(n).$$

We shall prove in Propositions 1 and 2 that these inequalities hold not only for infinitely many values of n, but for all values of n. I do not know whether the constants $c_4 = \frac{1}{16}$ and $c_5 = \frac{4}{31}$ are best possible.

In § 2 we shall prove two lemmas. In § 3 a general construction of critical k-chromatic graphs is described. Special cases with $k=4$ and $k=5$ are considered in § 4 and § 5 respectively. Theorem 1 is proved in § 4. In § 6 we shall prove Theorem 2, and in § 7 we conclude the paper by stating some unsolved problems.

I wish to thank G. A. DIRAC and P. ERDŐS for having drawn my attention to this problem and I. T. JAKOBSEN for fruitful discussions. I also express my thanks to T. GALLAI, to whom Lemma 1 is due. This lemma helped me to generalize my original constructions.

2. LEMMA 1. *Let Γ be a critical k-chromatic graph ($k \geqq 2$) with vertices $a_1, \ldots, a_n$. For each i, $1 \leqq i \leqq n$, let a_i be joined to a vertex b_i not in Γ by an edge. The vertices $b_1, \ldots, b_n$ are not necessarily all distinct, but their number is $\geqq 2$. They are each given one of the two colours 1 and 2, such that they do not all have the same colour. Then Γ is k-colourable with the colours 1, 2, ..., k in such a way that a_i gets a colour different from the colour of b_i for all i.*

PROOF of Lemma 1. Let for $j=1$ and $j=2$ M_j denote the set of vertices a_i for which b_i has the colour j. $M_1 \cap M_2 = \emptyset$, $M_1 \cup M_2 = V(\Gamma)$ and $M_1 \neq \emptyset \neq M_2$. Since Γ is connected it contains a $(M_1)\times(M_2)$-edge $e=(x_1, x_2)$, where $x_1 \in M_1$ and $x_2 \in M_2$. Γ is critical k-chromatic, hence $\Gamma - e$ is $(k-1)$-colourable. Let K denote a $(k-1)$-colouring of $\Gamma - e$ with the colours 2, ..., k, such that x_1 and x_2 both have colour 2. Change in the $(k-1)$-colouring K of the graph $\Gamma - e$ the colour of all vertices of M_2 of colour 2, and give them the new colour 1. The result is a k-colouring of Γ, where no vertex of M_j has the colour j, $j=1, 2$. This proves Lemma 1.

Lemma 1 and its proof are due to T. GALLAI. T. GALLAI asked in this connection: Is the lemma still true if we allow the vertices $b_1, \ldots, b_n$ to have $\geqq 2$ different colours? The question originates from a theorem of J. B. KELLY and L. M. KELLY ([6] Lemma 4. 1), who proved that the answer is yes if $k=3$. G. A. DIRAC ([4], Lemma 7) proved this when Γ is a complete k-graph.

LEMMA 2. *Let $k \geqq 4$ and suppose for all n, $n \geqq k$ and $n \neq k+1$, that $c_k n^2 < f_k(n)$. Then for all n, $n \geqq k+1$ and $n \neq k+2$,*

$$c_k n^2 < f_{k+1}(n).$$

PROOF of Lemma 2. Let $n \geqq k$ and $n \neq k+1$. Let Γ' denote a critical k-chromatic graph with n vertices and $f_k(n)$ edges. Let Γ denote a graph obtained from Γ' by

joining a new vertex to all vertices of Γ' by edges. Γ is critical $(k+1)$-chromatic and has $n+1$ vertices and $f_k(n)+n$ edges, hence

$$f_{k+1}(n+1) \geqq f_k(n)+n > c_k n^2 + n.$$

By (I) $c_k < \frac{1}{2}\frac{k-2}{k-1}$. A simple computation shows that

$$c_k n^2 + n > c_k (n+1)^2,$$

hence

$$f_{k+1}(n+1) > c_k (n+1)^2$$

and Lemma 2 follows.

3. Let k and m be positive integers, such that $k \geqq 2m$ and $k \geqq 3$. We shall define a class of graphs denoted $H_{k,m}$. A graph Γ belongs to $H_{k,m}$ if and only if it has the structure described in a), ..., g):

a) $V(\Gamma) = A_1 \cup \ldots \cup A_m \cup A_{m+1} \cup \ldots \cup A_{2m} \cup A_{2m+1}$, where $A_i \cap A_j = \emptyset$ for $i \neq j$.

b) For all i, $1 \leqq i \leqq m$, $\Gamma(A_i)$ is critical $(k-1)$-chromatic.

c) For all i, $1 \leqq i \leqq m$, A_{m+i} is an independent set of vertices in Γ, and $2 \leqq |A_{m+i}| \leqq |A_i|$.

d) $\Gamma(A_{2m+1})$ is critical $(k-2m)$-chromatic (if $k = 2m$ then $A_{2m+1} = \emptyset$).

e) For all i, $1 \leqq i \leqq m$, each vertex of A_i is joined to precisely one vertex of $V(\Gamma) - A_i$, and this vertex is contained in A_{m+i}.

f) For all i, $1 \leqq i \leqq m$, each vertex of A_{m+i} is joined to $\geqq 1$ vertices of A_i and to all vertices of A_{2m+1}.

g) For all pairs i, j satisfying $m+1 \leqq i < j \leqq 2m$ all $(A_i) \times (A_j)$-edges are in Γ.

A diagram of a graph in $H_{k,m}$ in the case $m = 2$ is shown in Figure 1.

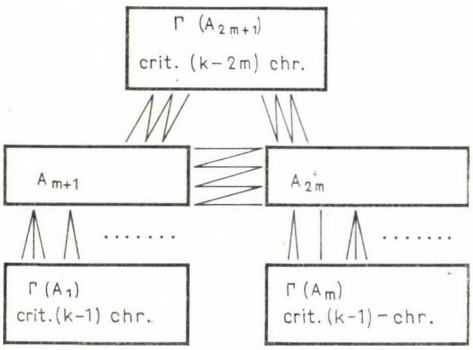


Fig. 1

(II) *If $\Gamma \in H_{k,m}$ then Γ is critical k-chromatic.*

PROOF of (II). Let $\Gamma \in H_{k,m}$. Suppose that Γ is $(k-1)$-colourable. For each i, $1 \leqq i \leqq m$, all $k-1$ colours of the $(k-1)$-colouring are used in $\Gamma(A_i)$, hence by e) the vertices of A_{m+i} have $\geqq 2$ different colours. Then because of g) the vertices of $A_{m+1} \cup \cdots \cup A_{2m}$ have $\geqq 2m$ different colours. It follows that $k-1 \geqq 2m$. $\Gamma(A_{2m+1})$ is $(k-2m)$-chromatic and all $(A_{m+1} \cup \cdots \cup A_{2m}) \times (A_{2m+1})$-edges are in Γ. Hence the $(k-1)$-colouring has $\geqq (k-2m)+2m$ different colours. This is a contradiction.

It follows that Γ has chromatic number $\geqq k$. The chromatic number of a graph decreases by at most one when an edge is deleted from the graph. In order to prove that Γ is critical k-chromatic it is therefore sufficient to prove for all edges e of Γ that $\Gamma - e$ is $(k-1)$-colourable. The symmetry of Γ permits us to consider only the following five cases:

(1) $$e \in E(\Gamma(A_{2m+1})).$$

$E(\Gamma(A_{2m+1}))\neq\emptyset$ implies $k\geqq 2m+2$. Colour for each i, $1\leqq i\leqq m$, the vertices of A_{m+i} with the colours i and $m+i$ so that not all vertices of A_{m+i} get the same colour. Colour $\Gamma(A_{2m+1})-e$ with the $k-2m-1$ colours $2m+1, \ldots, k-1$. This $(k-1)$-colouring of $\Gamma(A_{m+1}\cup\cdots\cup A_{2m+1})-e$ can by Lemma 1 be extended to a $(k-1)$-colouring of $\Gamma-e$.

(2) $$e=(x_{2m+1}, x_{2m}),$$

where $x_{2m+1}\in A_{2m+1}$ and $x_{2m}\in A_{2m}$. $A_{2m+1}\neq\emptyset$ implies $k\geqq 2m+1$. Colour for each i, $1\leqq i\leqq m$, the vertices of A_{m+i} as in 1), so that x_{2m} as the only vertex of A_{2m} gets the colour $2m$. Colour $\Gamma(A_{2m+1})$ with the $k-2m$ colours $2m, \ldots, k-1$ so that x_{2m+1} as the only vertex of A_{2m+1} gets the colour $2m$. This $(k-1)$-colouring of $\Gamma(A_{m+1}\cup\cdots\cup A_{2m+1})-e$ can by Lemma 1 be extended to a $(k-1)$-colouring of $\Gamma-e$.

(3) $$e=(x_{2m-1}, x_{2m}),$$

where $m\geqq 2$ and $x_{2m-1}\in A_{2m-1}$ and $x_{2m}\in A_{2m}$. Colour for each i, $1\leqq i\leqq m-1$, the vertices of A_{m+i} as in 1), so that x_{2m-1} as the only vertex of A_{2m-1} gets the colour $m-1$. Colour the vertices of A_{2m} with the colours $m-1$ and m, so that x_{2m} as the only vertex of A_{2m} gets the colour $m-1$. If $k\geqq 2m+1$ then colour $\Gamma(A_{2m+1})$ with the $k-2m$ colours $2m, \ldots, k-1$. This $(k-1)$-colouring of $\Gamma(A_{m+1}\cup\cdots\cup A_{2m+1})-e$ can by Lemma 1 be extended to a $(k-1)$-colouring of $\Gamma-e$.

(4) $$e=(x_m, x_{2m}),$$

where $x_m\in A_m$ and $x_{2m}\in A_{2m}$. If $m\geqq 2$ then colour for each i, $1\leqq i\leqq m-1$, the vertices of A_{m+i} as in 1). Colour all vertices of A_{2m} with the colour m. If $k\geqq 2m+1$ then colour $\Gamma(A_{2m+1})$ as in 3). Colour $\Gamma(A_m)$ with the colours $1, \ldots, k-1$, so that x_m as the only vertex of A_m gets the colour m. If $m=1$ then the result is a $(k-1)$-colouring of $\Gamma-e$. If $m\geqq 2$ then by Lemma 1 the $(k-1)$-colouring of $\Gamma(A_m\cup\cdots\cup A_{2m+1})-e$ can be extended to a $(k-1)$-colouring of $\Gamma-e$.

(5) $$e\in E(\Gamma(A_m)).$$

Colour $\Gamma(A_{m+1}\cup\cdots\cup A_{2m+1})$ as in 4). Colour $\Gamma(A_m)-e$ with the colours $1, \ldots, k-1$ so that no vertex of A_m gets the colour m. If $m=1$ then the result is a $(k-1)$-colouring of $\Gamma-e$. If $m\geqq 2$ then by Lemma 1 the $(k-1)$-colouring of $\Gamma(A_m\cup\cdots\cup A_{2m+1})-e$ can be extended to a $(k-1)$-colouring of $\Gamma-e$. (II) is then proved.

4. Let m_1, m_2, m_3 and m_4 be positive integers, so that m_1 and m_2 are odd and $\geqq 3$, $2\leqq m_3\leqq m_1$ and $2\leqq m_4\leqq m_2$. The set of graphs in $H_{4,2}$ with $|A_i|=m_i$ for $i=1, 2, 3$ and 4 shall be denoted $H_{4,2}(m_1, m_2, m_3, m_4)$. From this definition and (II) we get:

(III) *Let $\Gamma\in H_{4,2}(m_1, m_2, m_3, m_4)$. Then Γ is critical 4-chromatic and*

$$|V(\Gamma)| = m_1+m_2+m_3+m_4$$
$$|E(\Gamma)| = 2(m_1+m_2)+m_3\cdot m_4.$$

An example of a graph contained in $H_{4,2}(3, 7, 3, 4)$ is shown in Figure 2.

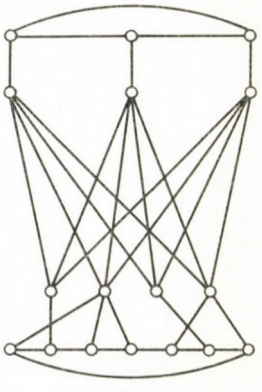


Fig. 2

PROPOSITION 1. *Let* $n \geqq 4$ *and* $n \neq 5$. *Then*

$$\frac{1}{16} n^2 < f_4(n).$$

PROOF of Proposition 1. Consider the following two cases:

(1) $$n \geqq 24.$$

Let $n = 8q+r$ where $q \geqq 3$ and $0 \leqq r \leqq 7$. Let $\Gamma \in H_{4,2}(m_1, m_2, m_3, m_4)$, where $m_1 = 2q+3$, $m_2 = 2q+1$, $m_3 = 2q+(r-4)$ and $m_4 = 2q$. From (III) we get:

$$f_4(n) \geqq |E(\Gamma)| = 2(4q+4)+2q(2q+r-4).$$

A simple computation with $q = \frac{1}{8}(n-r)$ shows that

$$2(4q+4)+2q(2q+r-4) > \frac{1}{16} n^2.$$

(2) $$n < 24.$$

All vertices of a critical 4-chromatic graph have valency $\geqq 3$, hence

$$f_4(n) \geqq \frac{3}{2} n > \frac{1}{16} n^2.$$

Hence Proposition 1 has been proved. The constant $\frac{1}{16}$ is the best possible constant c_4 that can be obtained from examples of type $H_{k,m}$.

Proposition 1 proves Theorem 1 in the case $k = 4$. The cases $k \geqq 5$ of Theorem 1 follow by induction over k using Lemma 2.

5. Let m_1, m_2 and m_3 be positive integers, so that $m_1 \geqq 4$ and $m_1 \neq 5$, $2 \leqq m_2 \leqq \leqq m_1$, and m_3 odd and $\geqq 3$. The set of graphs in $H_{5,1}$ with $|A_i| = m_i$ for $i = 1, 2$ and 3 shall be denoted $H_{5,1}(m_1, m_2, m_3)$. From this definition and (II) we get (IV):

(IV) *Let* $\Gamma \in H_{5,1}(m_1, m_2, m_3)$. *Then* Γ *is critical 5-chromatic and*

$$|V(\Gamma)| = m_1+m_2+m_3$$

$$|E(\Gamma)| = |E(\Gamma(A_1))|+m_1+m_3+m_2 \cdot m_3.$$

PROPOSITION 2. *Let* $n \geqq 5$ *and* $n \neq 6$. *Then*

$$\frac{4}{31} n^2 < f_5(n).$$

PROOF of Proposition 2. Consider the following three cases:

(1) $$n \geqq 24.$$

Let $8n = 31q+r$, where $q \geqq 6$ and $0 \leqq r \leqq 30$. If n is odd then let $\Gamma \in H_{5,1}(q, q, n-2q)$.

If n is even then let $\Gamma \in H_{5,1}(q, q-1, n-2q+1)$. In both cases we may assume by Proposition 1 that $|E(\Gamma(A_1))| > \frac{1}{16} q^2$. From (IV) we get

$$f_5(n) \geqq |E(\Gamma)| > \frac{1}{16} q^2 + q + (n-2q) + (q-1)(n-2q).$$

A simple computation with $q = \frac{1}{31}(8n-r)$ shows that

$$\frac{1}{16} q^2 + q(n-2q+1) > \frac{4}{31} n^2.$$

$$10 \leqq n \leqq 23. \tag{2}$$

If n is odd then let $\Gamma \in H_{5,1}(4, 4, n-8)$.
If n is even then let $\Gamma \in H_{5,1}(4, 3, n-7)$.

By (IV) and a simple computation we get

$$f_5(n) \geqq |E(\Gamma)| \geqq 6+4+(n-8)+3(n-8) > \frac{4}{31} n^2$$

$$5 \leqq n \leqq 9 \quad \text{and} \quad n \neq 6. \tag{3}$$

All vertices of a critical 5-chromatic graph have valency $\geqq 4$. Hence

$$f_5(n) \geqq 2n > \frac{4}{31} n^2$$

Hence $f_5(n) > \frac{4}{31} n^2$ for all n and Proposition 2 has been proved. The constant $c_5 = \frac{4}{31}$ is the best possible constant that can be obtained using the graphs of types $H_{5,1}$ and $H_{4,2}$. This can be seen by finding the maximal value of the function $h(m_2, m_3) = m_2 \cdot m_3 + \frac{1}{16}(n-m_2-m_3)^2$ under the conditions $m_i \geqq 0$, $i = 2$ and 3, and $m_3 + 2m_2 \leqq n$ (and this was the way in which the constant $\frac{4}{31}$ was first determined).

6. PROOF of Theorem 2. If $k = 4$ or $k = 5$ then the theorem follows from Proposition 1 and 2. We may therefore suppose that $q \geqq 2$.

a) Let $k = 3q$ and let $n = m \cdot q$, where m is odd and $\geqq 3$. We shall prove that

$$\frac{q-1}{2q} n^2 < f_k(n).$$

Let $\Gamma_1, \ldots, \Gamma_q$ denote q pairwise disjoint odd circuits each of length m. Let Γ denote the graph obtained from $\Gamma_1, \ldots, \Gamma_q$ by joining all vertices of Γ_i to all vertices of Γ_j for all pairs i, j with $i \neq j$. Γ is critical k-chromatic and has n vertices. Hence:

$$f_k(n) \geqq |E(\Gamma)| > \binom{q}{2} m^2 = \frac{q-1}{2q} n^2.$$

b) Let $k = 3q+1$ and let $n = (7q+1)\cdot m$, where m is odd and $\geqq 1$. We shall prove that

$$\frac{7q-6}{14q+2}n^2 < f_k(n).$$

Let Γ_1 denote a critical $3(q-1)$-chromatic graph with $7(q-1)\cdot m$ vertices and $> \frac{q-2}{2(q-1)}(7(q-1)\cdot m)^2$ edges. If $q=2$ then Γ_1 is an odd circuit of length $7m$. If $q \geqq 3$ then the existence of a graph Γ_1 with the required properties follows from Theorem 2a, since $7(q-1)\cdot m$ is one of the infinitely many n-values for which Theorem 2a holds.

Let Γ_2 denote a critical 4-chromatic graph with $8m$ vertices and $> \frac{1}{16}(8m)^2$ edges. The existence of such a Γ_2 follows from Proposition 1.

We may suppose that Γ_1 and Γ_2 are disjoint. Let Γ denote a graph obtained from Γ_1 and Γ_2 by joining all vertices of Γ_1 to all vertices of Γ_2 by edges. Γ is critical k-chromatic. It has n vertices and

$$|E(\Gamma)| > \frac{1}{2}49m^2(q-1)(q-2)+4m^2+56(q-1)m^2.$$

A simple computation with $m = \frac{n}{7q+1}$ shows that the right-hand side of the above inequality is equal to

$$\frac{7q-6}{2(7q+1)}n^2.$$

Since $f_k(n) \geqq |E(\Gamma)|$ Theorem 2b) follows.

c) Let $k = 3q+2$ and let $n = (23q+8)\cdot m$ where m is odd and $\geqq 1$. We shall prove that

$$\frac{23q-15}{46q+16}n^2 < f_k(n).$$

Let Γ_1 denote a critical $3(q-1)$-chromatic graph with $23(q-1)\cdot m$ vertices and $> \frac{q-2}{2(q-1)}(23(q-1)m)^2$ edges. If $q=2$ then Γ_1 is an odd circuit with $23\cdot m$ vertices. If $q \geqq 3$ then the existence of a graph Γ_1 with the required properties follows from Theorem 2a), since $23(q-1)\cdot m$ is one of the infinitely many n-values for which Theorem 2a) holds.

Let Γ_2 denote a critical 5-chromatic graph with $31\cdot m$ vertices and $> \frac{4}{31}(31m)^2$ edges. The existence of such a Γ_2 follows from Proposition 2.

We may suppose that Γ_1 and Γ_2 are disjoint. Let Γ denote a graph obtained from Γ_1 and Γ_2 by joining all vertices of Γ_1 to all vertices of Γ_2 by edges. Γ is critical k-chromatic and has n vertices and

$$|E(\Gamma)| > \frac{1}{2}23^2(q-1)(q-2)m^2+4\cdot 31\cdot m^2+23\cdot 31\cdot(q-1)m^2.$$

A simple computation with $m = \dfrac{n}{23q+8}$ shows that the right-hand side of the above inequality is equal to

$$\frac{23q-15}{2(23q+8)}n^2.$$

Since $f_k(n) \geqq |E(\Gamma)|$ Theorem 2c) follows.

Theorem 2 is then proved. The constants c_k obtained in Theorem 2 are the best possible constants that can be obtained using the constructions of this paper. I do not know whether these constructions are best possible.

If we instead of critical k-chromatic graphs consider the larger class of all vertex-critical k-chromatic graphs then the constants of Theorem 2 are not best possible in the cases where $r \neq 0$. In order to see this it is sufficient to consider the case $k=4$. Let $\Gamma \in H_{4,2}(h, h, h, h)$, where h is odd and $\geqq 3$. Identify the two odd circuits of length h spanned by A_1 and A_2 respectively (this idea was proposed by L. LOVÁSZ). The obtained graph Γ' is vertex-critical 4-chromatic and $|E(\Gamma')| > \dfrac{1}{9}(|V(\Gamma')|)^2$.

7. There are several unsolved problems connected with the function $f_k(n)$. By (I) and Theorem 1 the order of magnitude of $f_k(n)$ is known, but almost no information on exact values of $f_k(n)$ exists. Only in the very few cases where all critical k-chromatic graphs with n vertices have been determined $f_k(n)$ is known. Let us conclude by asking the following questions, some of which are due to P. ERDŐS [5]:

1) Are the constants $c_4 = \dfrac{1}{16}$ and $c_5 = \dfrac{4}{31}$ best possible?

2) $\dfrac{1}{4}n^2 < f_6(n)$ for infinitely many values of n. Is $\dfrac{1}{4}$ a best possible constant? Is $\dfrac{1}{4}n^2 < f_6(n)$ for all values of n, $n \geqq 6$ and $n \neq 7$?

3) If $n \equiv 2 \pmod 4$ and $n \geqq 6$ then $f_6(n) \geqq \dfrac{1}{4}n^2 + n$. Is $f_6(n) = \dfrac{1}{4}n^2 + n$?

4) Does $\lim\limits_{n\to\infty} \dfrac{f_k(n)}{n^2}$ exist?

REFERENCES

[1] DIRAC, G. A.: A property of 4-chromatic graphs and some remarks on critical graphs, *J. London Math. Soc.* **27** (1952), 85—92.
[2] DIRAC, G. A.: Some theorems on abstract graphs, *Proc. London Math. Soc.* **2** (1952), 69—81.
[3] DIRAC, G. A.: Circuits in critical graphs, *Monatsh. Math.* **59** (1955), 178—187.
[4] DIRAC, G. A.: A theorem of R. L. Brooks and a conjecture of H. Hadwiger, *Proc. London Math. Soc.* **7** (1957), 161—195.
[5] ERDŐS, P.: *Problems and results in chromatic graph theory.* Proof techniques in graph theory (Ed. Frank Harary). Academic Press (1969), p. 27—35.
[6] KELLY, J. B. and KELLY, L. M.: Paths and circuits in critical graphs, *Amer. J. Math.* **76** (1954), 786—792.
[7] ORE, O.: *The four-color problem.* Academic Press (1967).
[8] TURÁN, P.: Egy gráfelméleti szélsőértékfeladatról, *Mat. Fiz. Lapok* **48** (1941), 436—452.
[9] TURÁN, P.: On the theory of graphs, *Coll. Math.* **3** (1954), 19—30.

Aarhus Universitet, Matematisk Institut, Ny Munkegade 8000, *Aarhus* C, *Denmark*

(Received November 24, 1969.)

16

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 471—473.

ON A PROBLEM OF FEJES TÓTH

by

G. C. SHEPHARD

A finite set of convex bodies in Euclidean space is said to form a *packing* if no two have interior points in common. A packing is called *relatively stable* if it is impossible to displace any one of the bodies to a position exterior to the convex hull of the others without, at some intermediate position during this displacement, two of the bodies having interior points in common. Thus, in physical terms, if we regard the bodies as solid, it is impossible to remove any one from a relatively stable packing without disturbing at least one of the others.

In an interesting paper [2], L. FEJES TÓTH and A. HEPPES proved that there were no relatively stable packings in E^2, but a relatively stable packing of twelve tetrahedra in E^3 exists. At the Colloquim on Convexity held in Copenhagen during August 1965, Professor FEJES TÓTH asked the following question.

Does there exist a relatively stable packing of centrally symmetric convex bodies in E^3?

The purpose of this note is to answer this question in the affirmative by constructing such a packing of twelve convex bodies. We remark that by DE BRUIJN's Theorem [1], such a packing cannot be *absolutely stable*, that is to say, it is possible to "take apart" the packing by simultaneous displacement of all the bodies.

The construction is as follows. Let C be a cube of edge-length c and let $A_1, \ldots, A_{12}$ be the mid-points of its edges. These points may be joined by twelve line segments $[A_i A_j]$, each of length $\frac{1}{2}\sqrt{6}c$ in such a way that we obtain a configuration of four interlinked equilateral triangles (see figure 1). The symmetry group G^* of this configuration has order 24 being the group of rotational symmetries of the cube. Let O be the centre of C. Then each line OA_i meets one of the segments $[A_j A_k]$ in its mid-point, which will be denoted by B_i, and B_i is the mid-point of $[OA_i]$. Let C_i be the mid-point of $[OB_i]$. Let the points A_i be numbered in such a way that B_1 is the mid-point of $[A_2 A_3]$ and B_5 is the mid-point of $[A_1 A_4]$. Let π be the plane through A_1, C_5 and parallel to the direction of the line $A_2 A_3$. Applying the operations of the group G^* to π we obtain a set Π of 24 planes, two through each of the points A_i, and two through each of the points C_i.

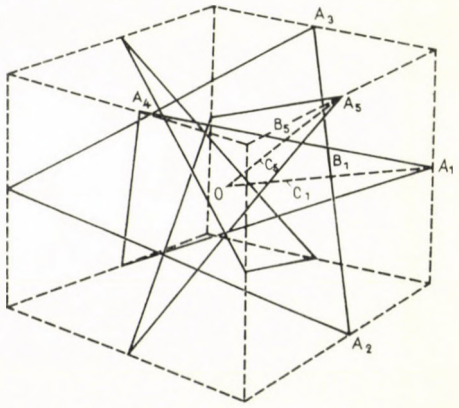


Fig. 1

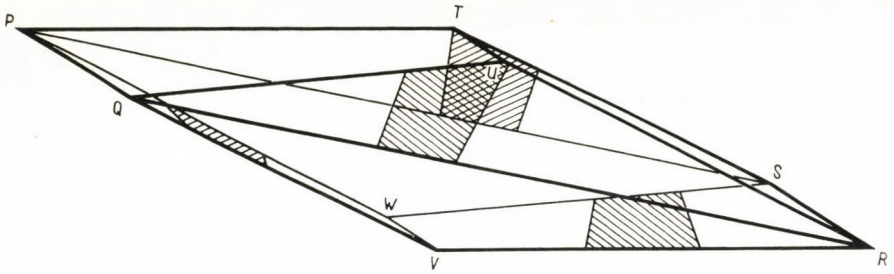


Fig. 2

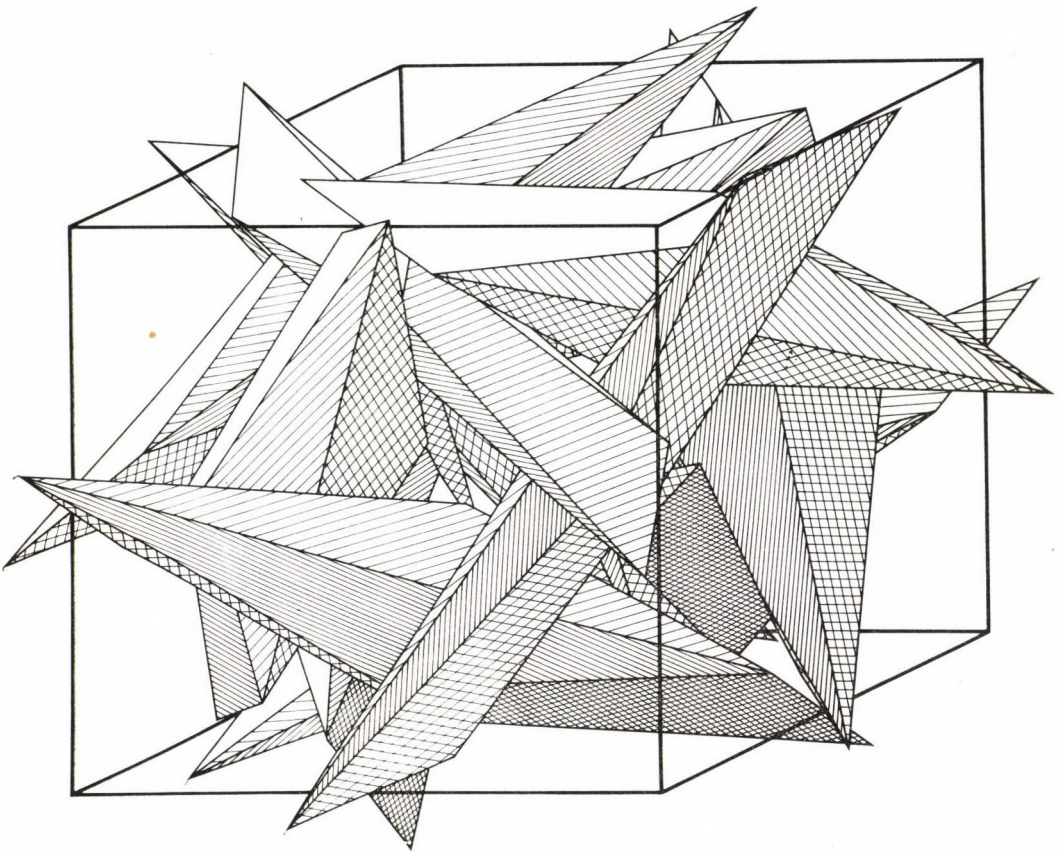


Fig. 3

Now let T_1 be the tetrahedron with B_1 as interior point and bounded by the two planes of Π that pass through A_1 and the two planes that pass through C_1. Let T_1^* be the tetrahedron produced by reflecting T_1 in the point B_1, and let $K_1 = T_1 \cap T_1^*$. Then K_1 is centrally symmetric and has OA_1 as an axis of 2-fold rotational symmetry (see figure 2). The set of twelve convex bodies obtained from K_1 by applying the operations of the group G^* form a packing, which, as we shall now show, is relatively stable. In figure 3 an isometric projection of the packing is shown. It is apparent from this diagram, and is easily verified analytically, that four faces of K_1 are in contact with faces of neighbouring K_i, the regions of contact being indicated in figure 2. Since these regions are parts of the faces of the tetrahedron T_1, it is evident that K_1 cannot be translated in any direction without disturbing at least one of its four neighbours. It is slightly more difficult, but perfectly straightforward, to verify that infinitesimal rotations of K_1 also violate the constraints imposed by the neighbouring bodies. Hence the packing has the required properties.

NOTE: The following numerical information will be found helpful in building a model of the packing. If the original cube has edge-length 112, then (see figure 2),

PQ, RS, TU and WV are parallel and $|PQ| = |RS| = 2|TU| = 2|VW| = 29.39$.

PS and QR are parallel and $|PS| = |QR| = 137.2$.

$|PT| = |UR| = |PW| = |VR| = 66.07$, $\quad |QU| = |TS| = |QV| = |WS| = 77.03$.

The coordinates of the 96 vertices of the twelve bodies, relative to axes through the centre of the cube and parallel to its edges may be found by applying the group of rotational symmetries of the cube to

(66, 60, 10) (52, 46, 10) (47, 37, 2) and (19, 9, 2).

REFERENCES

[1] DE BRUIJN, N. G.: Aufgaben 17 und 18, *Nieuw Archief voor Wiskunde* **2** (1954), 67; *Wiskundige Opgaven met de oplossingen* **20** (1955), 19—20.

[2] FEJES TÓTH, L. and HEPPES, A.: Über stabile Körpersysteme, *Compositio Math.* **15** (1962), 119—126.

Michigan State University East Lansing, Michigan.

(Received July 1, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 475—478.

ÜBER DIE NEWTONSCHE ZAHL EINER SCHEIBE KONSTANTER BREITE

von

J. SCHOPP

Das Problem der dreizehn Kugeln stellt die Frage, wieviele Einheitskugeln sich auf eine Einheitskugel auflegen lassen. Wir wollen uns hier mit einer der verschiedenen Varianten dieses berühmten Problems beschäftigen.

Unter *Scheiben* verstehen wir abgeschlossene ebene Bereiche, die keine gemeinsame innere Punkte haben. Wir sagen, dass zwei Scheiben einander *berühren*, wenn sie mindestens einen gemeinsamen Randpunkt haben. Es sei S eine Scheibe und $N(S)$ ihre Newtonzahl [1], d.h. die Maximalzahl der zu S kongruenten Scheiben, die mit S in Berührung gebracht werden können. Ist z. B. S ein Kreis, ein Quadrat, oder ein gleichseitiges Dreieck, so ist $N(S)$ gleich 6, 8 bzw. 12.

L. FEJES TÓTH stellte die Frage [2], was das Maximum N von $N(S)$ ist, wenn S alle Scheiben konstanter Breite durchläuft. Er zeigte, dass $7 \leqq N \leqq 8$, und sprach die Vermutung aus, dass $N = 7$ ist. Unser Hauptergebnis, das diese Vermutung bestätigt, ist enthalten im folgenden

SATZ 1. *Eine Scheibe konstanter Breite lässt sich mit höchstens 7 kongruenten Scheiben in Berührung bringen.*

Es sei S eine Scheibe mit der konstanten Breite 1. Bekanntlich hat S einen einzigen Inkreis, der mit dem Umkreis konzentrisch ist. Den gemeinsamen Mittelpunkt des In-, und Umkreises nennen wir *Mittelpunkt* von S. Ist R der Umkreisradius und r der Inkreisradius, so gelten die bekannten Ungleichungen [3]

$$1/2 \leqq R \leqq 1/\sqrt{3} \approx 0{,}57735 \tag{1a}$$

$$1/2 \geqq r \geqq (\sqrt{3}-1)/\sqrt{3} \approx 0{,}42265 \tag{1b}$$

wo die linksstehenden Schranken für den Kreis, die rechtsstehenden Schranken für das REULEAUX Dreieck erreicht werden. Wegen der ebenfalls bekannten Relation

$$R + r = 1 \tag{2}$$

sind (1a) und (1b) gleichwertig.

Ist O der Mittelpunkt, und G ein beliebiger Randpunkt von S so gilt offenbar

$$R \geqq OG \geqq r \tag{3}$$

Es seien $S_1, \ldots, S_n$ kongruente Scheiben mit den Mittelpunkten $O_1, \ldots, O_n$, die alle mit S je einen gemeinsamen Randpunkt haben. Wir betrachten den äusseren Parallelbereich S_ϱ von S vom Abstand ϱ. Aus (3) folgt, dass O_i $(i = 1, \ldots, n)$ nicht innerhalb S_r, und nicht ausserhalb S_R liegen kann. Wir zeigen, dass $O_1, \ldots, O_n$

Ecken eines konvexen n-Eckes sind. (Fig. 1) Da zwei Durchmesser von S stets einen gemeinsamen Punkt haben [4], gibt es durch O_i nur eine einzige Gerade, die S in einem Durchmesser $Q'Q$ schneidet, wo Q' die Punkte Q und O_i trennt. Die Konvexität wird bewiesen, wenn wir zeigen, dass $\alpha = \sphericalangle QO_iO_{i+1} < \pi/2$, wobei wir annehmen, dass $S_1, \ldots, S_n$ in zyklischer Reihenfolge angeordnet sind ($O_{n+1} = O_1$). Aus dieser Ungleichung folgt nähmlich, dass alle Innenwinkel des n-Eckes $O_1 \ldots O_n$ kleiner als π sind.

Es gelten offenbar die Ungleichungen

$$2(1-r) \geqq O_iO_{i+1} \geqq 2r \tag{4a}$$

$$O_iQ \geqq 1+r \tag{4b}$$

$$O_{i+1}Q \leqq 2-r \tag{4c}$$

woraus sich

$$2 \cdot O_iQ \cdot O_iO_{i+1} \cdot \cos\alpha = QO_i^2 + O_iO_{i+1}^2 - O_{i+1}Q^2 \geqq$$

$$\geqq (1+r)^2 + 4r^2 - (2-r)^2 \quad = 4r^2 + 6r - 3$$

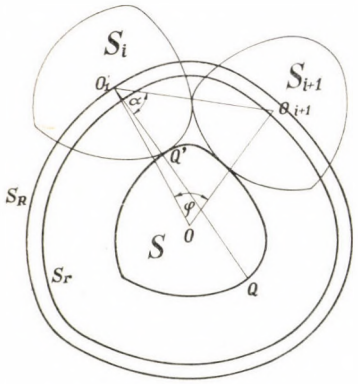


Fig. 1

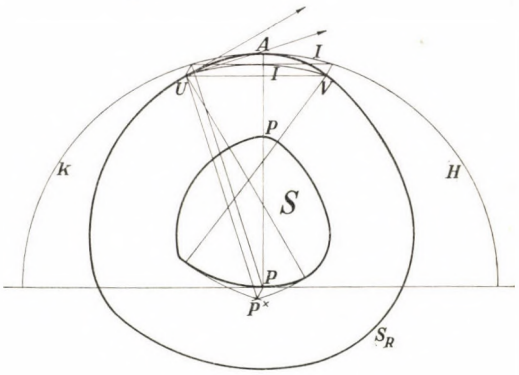


Fig. 2

ergibt. Es lässt sich aber leicht zeigen, dass die rechte Seite im Interval (1b) positiv ist, woraus tatsächlich $\alpha < \pi/2$ folgt. Der Umfang L von S_R ergibt sich aus einem bekannten Satz von BARBIER [5]:

$$L = (1+2R)\pi = (3-2r)\pi \tag{5}$$

Anderseits haben wir $U \geqq 2r \cdot n$, wo U den Umfang des n-Eckes bedeutet.

Hieraus ergibt sich wegen (1b) und (5)

$$n \leqq U/2r \leqq L/2r = (3-2r)\pi/2r \leqq \left(5+\sqrt{27}\right)\pi/4 \approx 8{,}0081,$$

also ein Wert, der noch grösser ist als 8.

Um eine bessere Schranke zu erhalten, konstruieren wir einen konvexen Bereich B, der das n-Eck $O_1 \ldots O_n$ enthält, jedoch in S_R liegt. Es sei A ein beliebiger Randpunkt von S_R (Fig. 2). Da S_R ein äusserer Parallelbereich ist, kann A kein singu-

lärer Punkt sein, und es gibt einen einzigen Durchmesser von S_R mit dem Endpunkt A. Dieser Durchmesser enthält gleichzeitig einen Durchmesser $P'P$ von S, wo P' die Punkte A und P trennt. Wir setzen voraus, dass der Vektor $\overrightarrow{PA}$ vertikal nach oben gerichtet ist. Zeichnen wir um P einen Kreis k mit dem Radius $\overrightarrow{PA} = 1+R = 2-r$. Dieser Kreis enthält offensichtlich S_R. Betrachten wir nun denjenigen S enthaltenden Halbkreisbereich H, den k und die durch P hindurchgehende, zu AP senkrechte Stützgerade von S begrenzt. Zeichnen wir in S_R eine zu AP senkrechte, innerhalb H liegende Sehne UV von der Länge $\leqq 3-2r$. Zeichnen wir jetzt die zu UV parallele, kongruente Sehne s von H, und betrachten wir diejenige Translation, die s in die Strecke UV überführt. Wir behaupten, dass diese Translation den zu s gehörigen offenen Kreisbogen der Länge I in das Innere von S_R überführt. Der Mittelpunkt des, den verschobenen Kreisbogen UV enthaltenden Kreises liegt ausserhalb H. Anderseits geht von U ein einziger Durchmesser von S_R aus, der gleichzeitig einen Durchmesser von S enthält, und den Durchmesser PP' offensichtlich schneidet. Hieraus folgt, dass die Tangente von S_R in U steiler ist, als die Kreistangente in U. Dasselbe gilt auch für V. Da dies für jede Sehne UV der Länge $\leqq 3-2r$ gilt, ist unsere Behauptung bewiesen. (Ist P ein singulärer Punkt von S, so kann vorkommen, dass die betrachtete Translation die Identität ist.) Durch Verschiebungen und Drehungen der Seiten des n-Ecks $O_1 \dots O_n$ können wir einen $O_1 \dots O_n$ enthaltenden konvexen Bereich konstruieren, der durch S_R und durch die Seiten von $O_1 \dots O_n$ kongruente Sehnen von S_R begrenzt ist. Wir verschieben diese Sehnen nach aussen, so dass ihre Länge gleich $2r$ wird und ersetzen jede verschobene Sehne durch je einen Kreisbogen vom Typ UV. Ist Λ der Umfang des so erhaltenen konvexen Bereiches B, so gilt $n \cdot I \leqq \Lambda \leqq L$, also

$$L \geqq n \cdot I. \tag{6}$$

Wegen $I \geqq (2-r)\varphi$ und $\varphi = 2 \cdot \arcsin [r/(2-r)]$ (Fig. 2.) haben wir

$$I \geqq 2 \cdot (2-r) \arcsin [r/(2-r)].$$

Aus (5) und (6) ergibt sich also

$$n \leqq \frac{L}{I} \leqq \frac{\pi}{2 \cdot \arcsin [r/(2-r)]} \cdot \frac{3-2r}{2-r} = f(r), \quad \frac{\sqrt{3}-1}{\sqrt{3}} \leqq r \leqq \frac{1}{2}, \tag{7}$$

da aber $f(r)$ eine monoton abnehmende Funktion ist, haben wir

$$n \leqq f\left(\frac{\sqrt{3}-1}{\sqrt{3}}\right) \approx 7{,}94, \quad \text{also} \quad N = 7 \quad \text{w. z. b. w.}$$

Aus der Monotonie von $f(r)$ folgt wegen $f(0{,}46) < 6{,}997$, dass für eine Scheibe von konstanter Breite 1, und mit einem Inkreisradius $r \geqq 0{,}46$ die Newtonzahl $N(S) = 6$ ist.

Wir können in einem gewissen Sinne entgegengesetzte Frage stellen bezüglich der Minimalzahl $M(S)$ derjenigen mit S kongruenten Scheiben, die S einschliessen [6, 7]. Bekanntlich gilt für jede konvexe Scheibe S, $M(S) \leqq 6$. Für einen Kreis bzw. für ein Dreieck haben wir $M(S) = 6$ bzw. $M(S) = 3$. Wir wollen hier das minimum für M von $S(M)$ für sämtliche Scheiben konstanter Breite bestimmen. Es gilt folgender

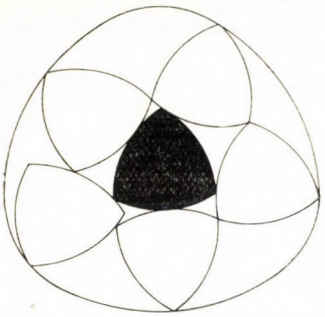


Fig. 3

SATZ 2. *Um eine Scheibe konstanter Breite mit kongruenten Scheiben umzuschliessen, sind mindestens 5 Scheiben nötig.*

Es lässt sich zeigen, dass für ein REULEAUX Dreieck $M(S)=5$ ist (Fig. 3). Folglich ist die gesuchte Minimalzahl $M=5$. Zum Beweis von Satz 2 genügt es zu zeigen, dass mit der obigen Bezeichnungen $\psi = \sphericalangle O_i O O_{i+1} < \pi/2$ gilt (s. Fig. 1). Wegen

$$(8) \quad 2 \cdot OO_i \cdot OO_{i+1} \cdot \cos\psi = OO_i^2 + OO_{i+1}^2 - O_iO_{i+1}^2$$

folgt aus (4*a*) und den trivialen Ungleichungen $OO_i \geqq 2r$ $(i=1, \ldots, n)$

$$OO_i \cdot OO_{i+1} \cdot \cos\psi = 2(r^2+2r-1).$$

Da aber die rechte Seite nach (1*b*) positiv ist, haben wir $\cos\psi > 0$, also tatsächlich $\psi < \pi/2$.

Aus (8) ergibt sich

$$\cos\psi = \frac{1}{2}\left(\frac{OO_i}{OO_{i+1}} + \frac{OO_{i+1}}{OO_i}\right) - \frac{O_iO_{i+1}^2}{2 \cdot OO_i \cdot OO_{i+1}}.$$

Hieraus folgt wegen (4a) und der obigen trivialen Ungleichungen $OO_i \geqq 2r$ $(i=1, \ldots, n)$

$$\cos\psi \geqq 1 - \frac{1}{2}\left(\frac{1-r}{2r}\right)^2 = g(r).$$

Ist $\cos\psi > \cos(2\pi/5) \approx 0{,}30902$, so ist $\psi < 2\pi/5$, und $M(S) > 5$. Aus der Monotonie von $g(r)$ folgt aber wegen $g(0,46) \approx 0{,}31096$, dass für eine Scheibe S von konstanter Breite 1, und mit einem Inkreisradius $r \geqq 0{,}46$ $M(S)=6$ gilt.

LITERATURVERZEICHNIS

[1] FEJES TÓTH, L.: Remarks on a theorem of R. M. Robinson, *Studia Sci. Math. Hung.* **4** (1969), 441—445.

[2] FEJES TÓTH, L.: On the number of equal discs that can touch another of the same kind, *Studia Sci. Math. Hung.* **2** (1967), 363—367.

[3] BONNESEN, T.—FENCHEL, W.: *Theorie der konvexen Körper,* Berlin 1934. S. 134.

[4] JAGLOM, J. M.—BOLTJANSKI, W. G.: *Konvexe Figuren* (aus der russischen Sprache übersetzt). Berlin 1956. S. 60 u. S. 207.

[5] BARBIER, E.: Note sur le probleme de l'aiguille et le jeu du joint couvert. *Journal de Mathématiques pures et appliquées.* 2^e série, **5** (1860), 273—286.

[6] FEJES TÓTH, L.—HEPPES, A.: Regions enclosed by convex domains. *Studia Sci. Math. Hung.* **1** (1966), 413—417.

[7] FEJES TÓTH, L.: Über das Didosche Problem. *Elem. Math.* **23** (1968), 97—101.

Technische Universität, Budapest

(Eingegangen: 27. Oktober, 1969.)

Studia Scientiarum Mathematicarum Hungarica 5 (1970) 479.

REMARK ON MY PAPER "DISPROOF OF SOME CONJECTURES ON DIOPHANTINE APPROXIMATIONS"

by

W. M. SCHMIDT

In a paper under the above title in this Journal, Studia Scientiarum Mathematicarum Hungarica **3** (1968), 137—144, I disproved a conjecture attributed to H. T. CROFT. It was brought to my attention that this already had been disproved in problem 4605 by NEWMAN and WEISSBLUM in Am. Math. Monthly **62** (1955), p. 738, and later by LEKKERKERKER in Indag. Math. **20** (1958), 197—205.

INDEX

Printed in Hungary

A kiadásért felel az Akadémiai Kiadó igazgatója – Műszaki szerkesztő: Várhelyi Tamás
A kézirat a nyomdába érkezett: 1970. VIII. 14. – Terjedelem: 23,5 (A/5) ív, 33 ábra

70-4282 – Szegedi Nyomda

Die *Studia Scientiarum Mathematicarum Hungarica* ist eine Halbjahrsschrift der Ungarischen Akademia der Wissenschaften. Sie veröffentlicht Originalbeiträge aus dem Bereiche der Mathematik in deutscher, englischer, französischer oder russischer Sprache. Es erscheint jährlich ein Band.

Adresse der Reaktion: Budapest V., Reáltanoda u. 13—15, Ungarn.
Technischer Redaktor: Gy. Petruska

Abonnementspreis pro Band (pro Jahr): $ 16.00. 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 téchique: Gy. Petruska

Le prix de l'abonnement: $ 16.00 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 Mathémat que (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 и немеском, английском, французском и русском языках. Отдольвные вышуски составляют ежегодно один том.

Адрес редакции: Budapest V., Reáltanoda u. 13—15, Венгрия.
Технический редактор: Gy. Petruska.

Подпусная цена на год (за один том): $ 16.00. Подписка на журнал принимаетця Внешнеторговым предприятием „Культура" Budapest 62, P. O. B. 149) или его представительствами за границей.
По поводу отношения обмена просим обрашаться к Библиотеке Института Математики (Budapest V. Reáltanoda u. 13—15).
Работы, предназначенные для опубликования в журнале следует направлять по адресу редакции в двух экземплярах.

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 Librairie
48, rue Gay Lussac
Paris 5

GERMAN DEMOCRATIC REPUBLIC

Deutscher Buchexport 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 Sanson
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 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

30. XI. 1971. **Index: 26 748**