Studia Scientiarum Mathematicarum Hungarica

AUXILIO
CONSILII INSTITUTI MATHEMATICI
ACADEMIAE SCIENTIARUM HUNGARICAE

OMUS XIII.
ASC. 1—2.
78

AKADÉMIAI KIADÓ, BUDAPEST

Studia Scientiarum Mathematicarum Hungarica

A Magyar Tudományos Akadémia matematikai folyóirata

Szerkesztőség: 1053 Budapest V., Reáltanoda u. 13—15.

Technikai szerkesztő: Merza József

Kiadja az Akadémiai Kiadó, 1054 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 (1011 Budapest I., Fő u. 32.).

Cserekapcsolatok felvétele ügyében kérjük az MTA Matematikai Kutató Intézete Könyvtárához (1053 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: 1053 Budapest V., Reáltanoda u. 13—15, Hungary.
Technical Editor: J. Merza

Orders may be placed with *Kultura* 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 (1053 Budapest V., Reáltanoda u. 13—15.)

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

Studia Scientiarum Mathematicarum Hungarica

Auxilio
Consilii Instituti Mathematici
Academiae Scientiarum Hungaricae

Tomus XIII

Akadémiai Kiadó, Budapest
1978

STUDIA SCIENTIARUM MATHEMATICARUM HUNGARICA

Tomus XIII

INDEX

80-2235 — Szegedi Nyomda — F. v.: Dobó József igazgató

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 1—17.

ON A GENERALIZATION OF DYADICITY

by

J. GERLITS

Introduction

D denotes the two-point discrete space; a continuous Hausdorff image of a product space $\mathbf{D}^\varkappa$ is said to be a *dyadic compactum*.

It is well-known that the class $\mathscr{D}\mathscr{C}$ of dyadic compacta behaves very nicely with respect to the topological cardinal functions. Roughly speaking, $\mathscr{D}\mathscr{C}$ is a one-parameter class; if $R\in\mathscr{D}\mathscr{C}$, the weight of R is τ, and φ is one of the "usual" topological cardinal functions then $\varphi(R)$ depends only on τ. Although this is only a heuristic principle (e.g., $|R|$ can be equal to 2^τ or $2^{\underset{\smile}{\tau}}$), I think it gives a fairly proper description of the situation.

In 1970 S. MRÓWKA introduced the class of λ-adic compacta [13]. For a cardinal λ $\mathbf{A}_\lambda$ denotes the one-point compactification of a discrete space of cardinality λ; a continuous Hausdorff image of a product space $\mathbf{A}_\lambda^\varkappa$ is said to be a *λ-adic compactum.* He proved that a compactum is dyadic iff it is ω-adic and raised the question: is every compactum λ-adic for a suitable cardinal λ?

The first counterexample was due to R. MARTY [12]; he proved, e.g., that a separable compactum can be λ-adic iff it is metrizable. The author proved [7] that the character and the weight of a λ-adic compactum coincide.

The aim of this paper is the continuation of these investigations; we should like to establish the relationship between the usual topological cardinal functions in the class of λ-adic spaces. As this will be more obvious in the sequel, the really interesting class is the class of compacta which are λ-adic for some cardinal λ. Such a space is said to be a polyadic compactum; $\mathscr{P}\mathscr{C}$ denotes the class of polyadic compacta.

Considering that the elements of $\mathscr{P}\mathscr{C}$ are the continuous Hausdorff images of the product spaces $\mathbf{A}_\lambda^\varkappa$, it can be suspected that $\mathscr{P}\mathscr{C}$ is a two-parameter class; there exist two topological cardinal functions φ_0, φ_1 such that if φ is one of the "usual" topological cardinal functions then the values of φ in $\mathscr{P}\mathscr{C}$ can be computed from the values of φ_0 and φ_1. This is indeed the case; the *cellularity c* and the *tightness t* can be chosen as these two functions.

The paper is divided into five sections, numbered 0 to 4. § 0 contains an account of terminology, notation, and some basic results which are for the most part known from the literature and are stated without proof.

§ 1 contains two theorems; the first gives a characterization of $\mathscr{P}\mathscr{C}$, the second states that a compact G_δ-set of a polyadic compactum is polyadic, too.

AMS (MOS) subject classifications (1970). Primary 54D30; Secondary 54A25.

Key words and phrases. Dyadic space, $\mathfrak{m}$-adic space, topological cardinal functions.

§ 2 is rather technical looking; it studies the behaviour of the tightness with respect to the products and generalized Σ-products of compacta.

The main task of § 3 is to prove the main theorems (3.9), (3.13) and (3.16).

Finally, in § 4 we show how the values of the usual topological cardinal functions can be computed in $\mathscr{P}\mathscr{C}$ from the cellularity and the tightness.

The reader familiar with the works of B. EFIMOV on dyadic compacta [4] [5], will certainly notice the similarities between his papers and our presentation. In the building up of the theory and in the proof of some theorems we relied considerably on the methods developed by B. Efimov. However, since the situation in the case of polyadic compacta is much more difficult, we were often compelled to more subtle considerations or to a completely different treatment of the subject.

Finally, I would express my sincerest gratitude to my colleagues A. HAJNAL and I. JUHÁSZ for suggestions which have improved the present work.

§ 0

$|H|$ denotes the cardinality of the set H.

We assume that each ordinal is the set of the smaller ordinals.

ξ, η, μ, ν denote ordinals;

$\varkappa, \lambda, \tau, \sigma$ denote cardinals (i.e. initial ordinals);

$\varkappa^+$ denotes the immediate successor of the cardinal $\varkappa$;

$\mathrm{cf}(\varkappa)$ denotes the cofinality of the cardinal $\varkappa$.

The closure of a set A in the topological space R is denoted by $\bar{A}$.

The *weight* of the topological space R is denoted by $w(R)=\min\{|\mathscr{B}|;\ \mathscr{B}$ is an open base of $R\}$. If x is a point in the space R, the character of x in R, resp. the character of R is defined by $\chi(x, R)=\min\{|\mathscr{U}|;\ \mathscr{U}$ is a local base of x in $X\}$ $\chi(R)=\sup\{\chi(x, R);\ x\in R\}$.

If $x\in R$, $A\subset R$, $x\in\bar{A}$, put

$$a(x, A)=\min\{|B|;\ B\subset A,\ x\in\bar{B}\}$$

$$t(x, R)=\sup\{a(x, A);\ A\subset R,\ x\in\bar{A}\}$$

$$t(R)=\sup\{t(x, R);\ x\in R\}$$

$$\hat{t}(R)=\min\{\varkappa;\ x\in\bar{A}\subset R \text{ implies } a(x, A)<\varkappa\}.$$

A system $\mathscr{U}$ of non-empty open sets in the space R is called a local π-base at the point $x\in R$ if given any neighbourhood V of x in R there exists a $U\in\mathscr{U}$ with $U\subset V$.

The π-character of x in R, resp. the π-character of R is defined by

$$\pi\chi(x, R)=\min\{|\mathscr{U}|;\ \mathscr{U} \text{ is a local } \pi\text{-base of } x \text{ in } R\},\quad \pi\chi(R)=\sup\{\pi\chi(x, R);\ x\in R\}.$$

The *cellularity* of R is defined by

$$c(R)=\sup\{|\mathscr{U}|;\ \mathscr{U} \text{ is a disjoint open system in } R\}$$

$$\hat{c}(R)=\min\{\varkappa;\ \text{if } \mathscr{U} \text{ is a disjoint open system in } R, \text{ then } |\mathscr{U}|<\varkappa\}.$$

Finally, the *density* of R is $d(R)=\min\{|S|;\ S\subset R,\ \bar{S}=R\}$.

By a mapping we mean a continuous function from a topological space to another; $f\colon X\to Y$ denotes a mapping from the space X into the space Y; $f\colon X\twoheadrightarrow Y$ indicates that the mapping is *onto*. If $A\subset X$ then $f|A$ denotes the restriction of f to the subspace A.

Let $\{X_i, i\in I\}$ be a family of topological spaces. We shall regard the points of the product space $X=\Pi\{X_i;\ i\in I\}$ as functions mapping I into $\cup\{X_i;\ i\in I\}$ with $f(i)\in X_i$ for $i\in I$. If $J\subset I, p\in X$, denote by $\pi_J(p)$ the restriction of p to the set J; it is a point of the partial product $X_J=\Pi\{X_i;\ i\in J\}$. If $J=\{i\}$ denote $\pi_{\{i\}}$ simply by π_i.

A set $U=\Pi\{U_i;\ i\in I\}$ is said to be a basic open set if the sets U_i are open in X_i and $I(U)=\{i\in I;\ X_i\neq U_i\}$ is finite; the family of basic open sets in X forms a base for the topology of X.

THEOREM A ([8]). *Assume $\{X_i;\ i\in I\}$ is a family of compacta, λ is a cardinal, $\operatorname{cf}(\lambda)>\omega$, $w(X_i)<\operatorname{cf}(\lambda)$, $f\colon X\twoheadrightarrow Y$ a mapping onto the T_2-space Y. If $w(Y)\geqq\lambda$ then Y contains a topological copy of $\mathbf{D}^\lambda$.* □

THEOREM B. (R. ENGELKING [13]; for a proof see [7].) *A polyadic compactum E is λ-adic iff $c(E)\leqq\lambda$.* □

The following, more general theorem can be proved in the same way as Engelking's theorem:

THEOREM C. *Let $\lambda>\omega$ be a regular cardinal; a polyadic compactum E is the continuous image of a product $\Pi\{\mathbf{A}_{\lambda_i};\ i\in I\}$ with $\lambda_i<\lambda$ $(i\in I)$ iff $\hat{c}(E)\leqq\lambda$.* □

COROLLARY D. *Let E be a polyadic space, $A\subset E$; then there exists a polyadic space Y, $A\subset Y\subset E$, $\hat{c}(Y)\leqq\hat{c}(A)$.* □

§ 1

The following characterization of the class $\mathscr{DC}$ is due to B. Efimov.

THEOREM. (B. Efimov [4].) *$\mathscr{DC}$ is the smallest class $\mathbb{O}$ of topological spaces such that*

a) *$\mathbb{O}$ contains each compact metrizable space;*

b) *$\mathbb{O}$ is closed for arbitrary topological products;*

c) *$\mathbb{O}$ is closed for continuous Hausdorff images (i.e., if $X\in\mathbb{O}$, Y is Hausdorff and $f\colon X\twoheadrightarrow Y$ then $Y\in\mathbb{O}$).*

THEOREM 1. *$\mathscr{PC}$ is the smallest class $\mathbb{O}$ of topological spaces s.t.*

a) *$\mathbb{O}\neq\emptyset$, $\exists X\in\mathbb{O}$, $X\neq\emptyset$;*

b) *$\mathbb{O}$ is closed for arbitrary topological products;*

c) *$\mathbb{O}$ is closed for continuous Hausdorff image;*

d) *Given any system $\{R_i;\ i\in I\}\subset\mathbb{O}$ there exists a space $R\in\mathbb{O}$ which is a compactification of the topological sum $\sum\{R_i;\ i\in I\}$.*

PROOF. First we prove that $\mathscr{PC}$ fulfils the conditions a)—d); here only d) needs a proof. We can suppose that $f_i\colon \mathbf{A}_\lambda^J\to R_i$ for $i\in I$. Denoting by R the one-point compactification of the topological sum $\sum\{R_i;\ i\in I\}$, it is immediate that R is the continuous image of the product space $\mathbf{A}_{|I|}\times\mathbf{A}_\lambda^I$. So, by b) and c), $R\in\mathscr{P}\mathscr{C}$.

On the other hand, suppose the class $\mathcal{O}$ satisfies the conditions a)—d); we prove $\mathscr{P}\mathscr{C}\subset\mathcal{O}$. By b) and c) we must prove only that $\mathbf{A}_\lambda\in\mathcal{O}$ for each cardinal λ. Using a) and c) we get $D(1)\in\mathcal{O}$ and, by d), $D(\lambda)$ has a compactification $R_\lambda\in\mathcal{O}$. Making use of the fact that $\mathbf{A}_\lambda$ is the smallest compactification of $D(\lambda)$, c) gives the assertion. □

Another interesting result in [4] is the following

THEOREM (B. Efimov [4]). *A compact G_δ-subspace of a dyadic compactum is dyadic.* □

This result applies also to polyadic spaces.

THEOREM 2. *A compact G_δ-subspace of a polyadic compactum is polyadic.*

PROOF. The inverse image of a closed G_δ-subspace by a mapping is a closed G_δ, too, so we must only prove that if $C\subset\mathbf{A}_\lambda^I$ is a compact G_δ then $C\in\mathscr{P}\mathscr{C}$; we can assume that $\lambda>\omega$.

Call a set E in $\mathbf{A}_\lambda^I$ an *elementary set* if it has the form $E=\pi_i^{-1}(H)$ where $i\in I$, $H\subset\mathbf{A}_\lambda-\{\Omega\}$, H is finite or $\mathbf{A}_\lambda-H\subset\mathbf{A}_\lambda-\{\Omega\}$, $\mathbf{A}_\lambda-H$ is finite; put $K(E)=H$ or $\mathbf{A}_\lambda-H$ accordingly; $K(E)$ is a finite set in $\mathbf{A}_\lambda-\{\Omega\}$. A *base-set* is the intersection of finitely many elementary sets; the family of base-sets form a base of $\mathbf{A}_\lambda^I$. If B is a base-set, $B=\bigcap\{E_k;\ k<n\}$, put $K(B)=\bigcup\{K(E_k);\ k<n\}$.

Choose now a finite family $\mathscr{G}_n$ of base-sets for each $n<\omega$ s.t. if $G_n=\bigcup\mathscr{G}_n$ then $C=\bigcap\{G_n;\ n<\omega\}$ and put $K=\bigcup\{K(B);\ B\in\bigcup\{\mathscr{G}_n;\ n<\omega\}\}$; now $K\subset\mathbf{A}_\lambda-\{\Omega\}$, $|K|\leqq\omega$.

CLAIM 1. *Let $p,q\in\mathbf{A}_\lambda^I$, $p\in C$. Assume that for each $i\in I$ if $p(i)\neq q(i)$ then $\{p(i),q(i)\}\cap K=\emptyset$; then $q\in C$.*

PROOF. If $q\notin C$ choose an $n<\omega$, $B\in\mathscr{G}_n$ with $p\in B$, $q\notin B$ and an $i\in I$ with $p(i)\in\pi_i(B)$, $q(i)\notin\pi_i(B)$. Now there exists an elementary set $E=\pi_i^{-1}(H)$, $K(E)\subset\subset K(B)$ s.t. $p(i)\in H$, $q(i)\in\mathbf{A}_\lambda-H$. One of the sets H, $\mathbf{A}_\lambda-H$ is contained in the set $K(E)\subset K(B)\subset K$ so $\{p(i),q(i)\}\cap K\neq\emptyset$, a contradiction. □

Let now $z_0\in\mathbf{A}_\lambda-(K\cup\{\Omega\})$ be an arbitrary point, $L=K\cup\{z_0\}\cup\{\Omega\}$, $F=C\cap L^I$. The space L^I is a dyadic compactum, F is a compact G_δ in L^I hence F is dyadic. Let $\varphi\colon F\times(\mathbf{A}_\lambda-K)^I\to\mathbf{A}_\lambda^I$ be the following function:

$$\varphi(x,y)(i)=\begin{cases} y(i) & \text{if } x(i)=z_0\\ x(i) & \text{otherwise.}\end{cases}$$

CLAIM 2. *φ is continuous.*

We must prove that the function $\varphi_i=\pi_i\circ\varphi$ is continuous, i.e., that given any isolated point x of $\mathbf{A}_\lambda$, $\varphi_i^{-1}(\{x\})$ is a clopen set in $F\times(\mathbf{A}_\lambda-K)^I$. For $i\in I$, let ${}^F\pi_i$ (resp. ${}^A\pi_i$) the projection of F (resp. $(\mathrm{A}_\lambda-K)^I$) into the i-th factor-space, i.e., ${}^F\pi_i=\pi_i|F$ and ${}^A\pi_i|(\mathbf{A}_\lambda-K)_i$.

If $x\neq z_0$, then $\varphi_i^{-1}(\{x\})=({}^F\pi_i^{-1}(\{z_0\})\cap{}^A\pi_i^{-1}(\{x\})\cup{}^F\pi_i^{-1}(\{x\})$ is a clopen set; $\varphi_i^{-1}(\{z_0\})={}^F\pi_i^{-1}(\{z_0\})\cap{}^A\pi_i^{-1}(\{z_0\})$ is also clopen. □

CLAIM 3. $\varphi(F\times(\mathbf{A}_\lambda-K)^I)=C$.

a) Let $p\in C$; denote $p'\in\mathbf{A}_\lambda^I$ the point

$$p'(i)=\begin{cases} z_0 & \text{if} \quad p(i)\notin K \\ p(i) & \text{if} \quad p(i)\in K. \end{cases}$$

Now $p'(i)\in L$ for each $i\in I$; using Claim 1 to the point $q=p'$ we get $p'\in C$ so $p'\in C\cap L^I=F$. Denote $p''\in\mathbf{A}_\lambda^I$ the point

$$p''(i)=\begin{cases} z_0 & \text{if} \quad p(i)\in K \\ p(i) & \text{if} \quad p(i)\notin K. \end{cases}$$

Evidently, for each $i\in I$ $p''(i)\notin K$ hence $(p',p'')\in F\times(\mathbf{A}_\lambda-K)^I$.

If $i\in I$ then if $p(i)\notin K$ then $p'(i)=z_0$, $p''(i)=p(i)$ and $\varphi(p',p'')(i)=p(i)$; if $p(i)\in K$ then $p'(i)=p(i), p''(i)=z_0$ and $\varphi(p',p'')(i)=p(i)$. Hence we proved that $\varphi(p',p'')=p$.

b) Let $(p,p')\in F\times(\mathbf{A}_\lambda-K)^I$. Now $p\in C$; let $q=\varphi(p,p')$. If $i\in I$ and $p(i)=p'(i)$ then $q(i)=p(i)$. If $p(i)\neq q(i)$ then $p(i)=z_0$ and $q(i)=p'(i)$. But $p'(i)\notin K$ hence $\{p(i),q(i)\}\cap K=\emptyset$; by Claim 1 $q\in C$. □

So we proved that C is the continuous image of the polyadic compactum $F\times(\mathbf{A}_\lambda-K)^I$ so C is a polyadic compactum, too.

Q.e.d

§ 2

DEFINITION 1 (A. V. ARHANGELSKIĬ [2]). Let X be a topological space, $\varkappa$ be a cardinal; the set $A\subset X$ is *$\varkappa$-closed* if $B\subset A$, $|B|<\varkappa$ implies $\bar{B}^X\subset A$.

We list without proof some easy consequences of the definition.

Any intersection of $\varkappa$-closed sets is $\varkappa$-closed; the union of finitely many $\varkappa$-closed sets is $\varkappa$-closed; a closed set is always $\varkappa$-closed.

If $B\subset A\subset X$, B is $\varkappa$-closed in A, A is $\varkappa$-closed in X then B is $\varkappa$-closed in X.

If $\varkappa$ is a regular cardinal, $A\subset X$ then the set $[A]_\varkappa=\bigcup\{\bar{H};\ H\subset A, |H|<\varkappa\}$ (the "$\varkappa$-closure of A") is $\varkappa$-closed and $x\in[A]_\varkappa$ iff $x\in\bar{A}$ and $a(x,A)<\varkappa$.

If X and Y are topological spaces, $f\colon X\twoheadrightarrow Y$ is a closed mapping and $A\subset X$ is $\varkappa$-closed then $f(A)\subset Y$ is $\varkappa$-closed; if $B\subset Y$ is $\varkappa$-closed then $f^{-1}(B)\subset X$ is $\varkappa$-closed.

The following theorem was proved by V. MALYHIN.

THEOREM (V. Malyhin [11]). *If X and Y are T_2-spaces, X is compact then $t(X\times Y)\leqq t(X)+t(Y)$.* □

We need a slightly more general result. If X is a topological space put

$$\hat{t}(X)=\min\{\varkappa;\ A\subset X, x\in\bar{A} \text{ implies } a(x,A)<\varkappa\}.$$

Evidently, always $t(X)\leqq\hat{t}(X)$, if $t(X)=\varkappa$ then $\hat{t}(X)$ is equal to $\varkappa^+$ or $\varkappa$ according to the condition that the topological cardinal function $a(x,A)$ attains or not the value $\varkappa$ in X.

It is easily seen that if $\varkappa$ is a regular cardinal then $\hat{t}(X)\leqq\varkappa$ iff each $\varkappa$-closed set in X is closed.

LEMMA 2. *If X, Y are T_2-spaces, X is compact, $\hat{t}(X)\leqq\varkappa$, $\hat{t}(Y)\leqq\varkappa$ and $\varkappa$ is regular then $\hat{t}(X\times Y)\leqq\varkappa$.*

PROOF. We must only prove that if $A\subset Z=X\times Y$ is $\varkappa$-closed in Z then A is closed. Assume $p=(a, b)\in\bar{A}$, and put $A_b=\{x\in X;\ (x, b)\in A\}\subset X$. If $a\in\bar{A}_b^X$ then, by $\hat{t}(X)\leqq\varkappa$, $a\in A_b$ hence $(a, b)\in A$. Hence we can suppose that the point a has a closed nbd U in X, $U\cap A_b=\emptyset$. Now p is in the closure (in Z) of the set $H=\pi_X^{-1}(U)\cap A$. H is $\varkappa$-closed in Z and $\pi_Y\colon Z\to Y$ is a closed mapping so b is in the closure in Y of the $\varkappa$-closed set $K=\pi_Y(H)$. Now, by $\hat{t}(Y)\leqq\varkappa$, we get $b\in K$ so $U\cap A_b\neq\emptyset$, a contradiction. □

COROLLARY 3. *Let $\{X_i;\ i\in I\}$ be a family of compacta, $X=\prod\{X_i; i\in I\}$ their topological product. If $\varkappa\geqq\omega_1$ is a regular cardinal, $\hat{t}(x_i)\leqq\varkappa$ for each $i\in I$ and $|I|<\varkappa$ then $\hat{t}(X)\leqq\varkappa$.*

PROOF. Let $A\subset X$ be a $\varkappa$-closed set and let $p\in\bar{A}^X$. If $J\subset I$, $|J|<\omega$ then $\pi_J(A)$ is $\varkappa$-closed in X_J and $\hat{t}(X_J)\leqq\varkappa$ by Lemma 2, hence $\pi_J(p)\in\pi_J(A)$. This shows that given any $J\subset I$, $|J|<\omega$, we can select a point $p_J\in A$ with $p_J|J=p|J$. Put $B=\{p_J;\ J\subset I, |J|<\omega\}\subset A$. Now $|B|\leqq|I|\cdot\omega<\varkappa$ and A is $\varkappa$-closed so $\bar{B}^X\subset A$. But $p\in\bar{B}$ and so $p\in A$. □

COROLLARY 4 (V. Malyhin [11]). *Let $\{X_i; i\in I\}$ be a family of compacta, $|X_i|\geqq 2$, $X=\prod\{X_i;\ i\in I\}$ the product; then $t(X)=\sum\{t(X_i);\ i\in I\}$.*

PROOF. Here the sign $\geqq$ is evident; and if $\lambda=\sum\{t(X_i);\ i\in I\}$, then Corollary 4 works for $\varkappa=\lambda^+$. □

LEMMA 5. *Let $\{X_i;\ i\in I\}$ be a family of compacta, $X=\prod\{X_i;\ i\in I\}$, $\varkappa\geqq\omega$ a cardinal, $A\subset X$ $\varkappa$-closed, $p\in\bar{A}$. If for each set $J\subset I$, $|J|<\varkappa$ $\hat{t}(X_J)\leqq\varkappa$ holds, then for each set $J\subset I$, $|J|<\varkappa$ there exists a $p_J\in A$ with $p_J|J=p|J$.*

PROOF. Take in X_J the point $\pi_J(p)$ and the set $\pi_J(A)$; by $\hat{t}(X_J)\leqq\varkappa$ $\pi_J(p)\in\pi_J(A)$ and this shows the validity of the assertion. □

REMARK. The condition of Lemma 5 is fulfilled if for each $i\in I$ $\hat{t}(X_i)\leqq\tau$ where τ is a regular cardinal, $\tau\leqq\varkappa$. Indeed, if $\tau=\varkappa$, this follows from Corollary 3 and if $\tau<\varkappa$ then, by Corollary 4, $t(X_J)\leqq\sum\{t(X_i);\ i\in J\}\leqq\sum\{\hat{t}(X_i);\ i\in J\}\leqq\tau\cdot|J|<\varkappa$ hence $\hat{t}(X_J)\leqq(t(X_J))^+\leqq\varkappa$. □

In the sequel $\{X_i\colon i\in I\}$ denotes a family of compacta, $X=\prod\{X_i;\ i\in I\}$ the topological product.

DEFINITION 6. If $\varkappa$ is a cardinal, $p, \vartheta\in X$, put

$$I(p, \vartheta)=\{i\in I;\ p(i)\neq\vartheta(i)\}$$

$$\sum(\varkappa, \vartheta)=\{p\in X;\ |I(p, \vartheta)|<\varkappa\}.$$

It is immediate that for $\varkappa\geqq\omega$ $\sum(\varkappa, \vartheta)$ is dense in X and if $\varkappa$ is regular then $\sum(\varkappa, \vartheta)$ is $\varkappa$-closed.

LEMMA 7. *If $\varkappa \geqq \omega_1$ is a regular cardinal and $\hat{t}(X_i) \leqq \varkappa$ for each $i \in I$ then $\hat{t}(\sum(\varkappa, \vartheta)) \leqq \varkappa$.*

PROOF. Let $A \subset \sum(\varkappa, \vartheta)$ be $\varkappa$-closed in $\sum(\varkappa, \vartheta)$; then it is also $\varkappa$-closed in X and choose a point $p \in \bar{A} \cap \sum(\varkappa, \vartheta)$.

Put $p_0 = p$; if $n < \omega$ and for each $k < n$ $p_k \in \sum(\varkappa, \vartheta)$ is defined, denote $J_n = \cup \{I(p_k, \vartheta);\ k < n\}$. Using that $p_k \in \sum(\varkappa, \vartheta)$ $(k < n)$ we get $|J_n| < \varkappa$.

Making use of Lemma 5 (and the remark following it) we can select a point $p_n \in A$ with $p_n | J_n = p | J_n$. Let now $J = \cup \{J_n;\ n < \omega\} = \cup \{I(p_n, \vartheta);\ n < \omega\}$. If $i \in J$, there exists an $n < \omega$ s.t. $i \in I(p_n, \vartheta)$ hence $p_m(i) = p(i)$ if $n < m < \omega$. If $i \in I - J$ then $p_n(i) = \vartheta(i) = p_0(i) = p(i)$ for each $n < \omega$.

Hence given any $i \in I$ we get an index $n(i) < \omega$ s.t. if $n \geqq n(i)$ then $p_n(i) = p(i)$. This shows that the sequence $\langle p_n;\ 1 \leqq n < \omega \rangle$ converges to the point p; but A is ω_1-closed so $p \in A$. □

COROLLARY 8. *If $\varkappa \geqq \omega_1$ is regular and $\hat{t}(X_i) \leqq \varkappa$ $(i \in I)$, then $\sum(\varkappa, \vartheta)$ does not contain any free sequence of cardinality $\varkappa^+$.*

PROOF. We remind the reader that $\langle p_\xi;\ \xi < \lambda \rangle$ is a *free sequence* in the topological space E if for each $\eta < \lambda$ $\overline{\{p_\xi;\ \xi < \eta\}} \cap \overline{\{p_\xi;\ \eta \leqq \xi < \lambda\}} = \emptyset$. By a theorem of W. COMFORT [3] the subspace $\sum(\varkappa, \vartheta)$ is $\varkappa$-Lindelöf and we know from Lemma 7 that $t(\sum(\varkappa, \vartheta)) \leqq \varkappa$. Finally, a theorem of A. Arhangelskiĭ [2] states that if E is $\varkappa$-Lindelöf and $t(E) \leqq \varkappa$ then E does not contain any free sequence of cardinality $\varkappa^+$. □

§ 3

In this paragraph we shall always assume that $\{X_i;\ i \in I\}$ is a family of compacta, $X = \prod \{X_i;\ i \in I\}$ is the topological product, $f\colon X \twoheadrightarrow Y$ is a mapping onto the T_2-space Y, $\tau = \sup \{t(X_i);\ i \in I\}$, $\varkappa \geqq \max(\omega, \tau)$.

DEFINITION 1. If $p \in X$, put $\operatorname{ord}(f, p) = \min \{|J|;\ J \subset I,$ if $q \in X, q | J = p | J$ then $f(p) = f(q)\}$.

DEFINITION 2. If $\vartheta \in X$, $\lambda \geqq \omega$ is a cardinal, put $H(\lambda) = \{y \in Y;$ for each $p \in f^{-1}(y)$ $\operatorname{ord}(f, p) < \lambda\}$,

$$S(\lambda, \vartheta) = f(\textstyle\sum(\lambda, \vartheta)),$$

$$S(\lambda) = \cap \{S(\lambda, \vartheta);\ \vartheta \in X\}.$$

It is easily seen that $H(\lambda) \subset S(\lambda)$.

In the proof of the following theorem we shall need a result from [8].

THEOREM ([8]). *Let $X = \prod \{X_i;\ i \in I\}$ be the topological product of the spaces $\{X_i;\ i \in I\}$, $p \in X$; $f\colon X \twoheadrightarrow Y$ be a mapping onto the T_2-space Y. If $\operatorname{ord}(f, p) \geqq \lambda > \omega$ then there exists a subspace $C \subset X$ homeomorphic to $\mathbf{D}^\lambda$ s.t. $p \in C$ and $f | C$ is an imbedding.* □

THEOREM 3. *For each point $y\in Y$ the following conditions are equivalent:*
a) $y\notin H(\varkappa)$;
b) *There exists a set* $A\subset Y$ *with* $y\in\bar{A}$, $a(y, A)\geqq\varkappa$;
c) *There exists a subspace* $C\subset Y, y\in C, C$ *is homeomorphic to* $\mathbf{D}^{\varkappa}$.

PROOF. a→c. This is an immediate consequence of the above-mentioned theorem.

c→b. Trivial.

b→a. If $B=f^{-1}(A)$, there exists a point $p\in X$ with $p\in\bar{B}^X$, $f(p)=y$ because $y\in\bar{A}^Y\subset f(\bar{B}^X)$; it is enough to show that ord $(f, p)<\varkappa$ is impossible. Choose a regular cardinal $\sigma\leqq\varkappa$ with $\tau<\sigma$, ord $(f, p)<\sigma$. If $H\subset X$ is the σ-closure of B and $J\subset I$, $|J|<\sigma$ is a set s.t. $q\in X, q|J=p|J$ implies $f(p)=f(q)$, then, by (2.5), there exists a point $q\in H$ with $q|J=p|J$, hence $y=f(p)=f(q)$. Now, for a suitable set $E\subset B, |E|<\sigma, q\in\bar{E}^X$ so $y=f(q)\in f(\bar{E}^X)=\overline{f(E)}^Y$. But $f(E)\subset A, |f(E)|\leqq|E|<$ $<\sigma\leqq\varkappa$ consequently $a(y, A)<\varkappa$, a contradiction. □

THEOREM 4. *For any infinite cardinal* λ $\{y\in Y;\ \pi\chi(y, Y)<\lambda\}\subset S(\lambda)$.

PROOF. Assume $\pi\chi(y, Y)<\lambda$; then there exists a family $\mathcal{U}$ of non-empty basic open sets in X s.t. $|\mathcal{U}|<\lambda$ and if G is a nbd of y in Y then $f(U)\subset G$ for a suitable $U\in\mathcal{U}$.

Put $J=\bigcup\{I(U);\ U\in\mathcal{U}\}$; then $J\subset I$, $|J|<\lambda$ and for any fixed $\vartheta\in X$ put $C=$ $=\bigcap\{\pi_i^{-1}(\vartheta(i));\ i\in I-J\}$. The set $C\subset X$ is compact, $C\subset\sum(\lambda, \vartheta)$ and $y\in f(C)$ because $U\cap C\neq\emptyset$ for each $U\in\mathcal{U}$ hence $f(U)$ is not contained in the open set $Y-f(C)$. This shows that $y\in f(C)\subset f\big(\sum(\lambda, \vartheta)\big)=S(\lambda, \vartheta)$. The point $\vartheta\in X$ was arbitrary hence $y\in\bigcap\{S(\lambda, \vartheta);\ \vartheta\in X\}=S(\lambda)$. □

LEMMA 5. *Let λ be an infinite cardinal, $C\subset X$ be compact, $C\cap\sum(\lambda, \vartheta)\neq\emptyset$ for each $\vartheta\in X$. Then there exists a set $J\subset I$, $|J|<\lambda$ s.t.* $\pi_{I-J}(C)=X_{I-J}$.

PROOF. Assume the Lemma is false; given any set $J\subset I$, $|J|<\lambda$, there exists a basic open set $U\neq\emptyset$ in X, $U\cap C=\emptyset, I(U)\cap J=\emptyset$.

Now, by transfinite induction, we can choose a sequence $\langle U_\xi;\ \xi<\lambda\rangle$ of non-empty basic open sets s.t. for $\xi<\eta<\lambda$ $U_\xi\cap C=\emptyset$, $I(U_\xi)\cap I(U_\eta)=\emptyset$. Select $\vartheta\in\bigcap\{U_\xi;\ \xi<\lambda\}$, $p\in C\cap\sum(\lambda, \vartheta)$. By $|I(p, \vartheta)|<\lambda$ there exists an ordinal $\xi<\lambda$ with $I(p, \vartheta)\cap I(U_\xi)=\emptyset$; but then $\vartheta\in U_\xi$ implies $p\in U_\xi$, so $p\in U_\xi\cap C=\emptyset$, a contradiction. □

COROLLARY 6. *If λ is a limit-cardinal then* $S(\lambda)=\bigcup\{S(\sigma);\ \sigma$ is a cardinal, $\sigma<\lambda\}$.

PROOF. Here the inclusion $\supset$ is evident. On the other hand, assume $y\in S(\lambda)$ and put $C=f^{-1}(y)$. Now $C\subset X$ is a compact set, $C\cap\sum(\lambda, \vartheta)\neq\emptyset$ for each $\vartheta\in X$ hence, by Lemma 5, there exists a set $J\subset I$, $|J|=\sigma<\lambda, \pi_{I-J}(C)=X_{I-J}$. Evidently, $C\cap S(\sigma^+, \vartheta)\neq\emptyset$ for each $\vartheta\in X$ so $y\in S(\sigma^+)$. The cardinal λ being a limit-cardinal, $\sigma^+<\lambda$. □

THEOREM 7. *If $\varkappa$ is regular then $S(\varkappa)$ is compact and* $\hat{t}(S(\varkappa))\leqq\varkappa$.

PROOF. If $\varkappa$ is regular, $S(\varkappa)$ is $\varkappa$-closed in Y. We must prove that if $A\subset S(\varkappa)$ is $\varkappa$-closed then A is compact; assume $y\in\bar{A}-A$. Put $B=f^{-1}(A), C=f^{-1}(y)$; B is $\varkappa$-closed in X, C is compact, $B\cap C=\emptyset$.

For each $p\in C$ the set $B\cap\sum(\varkappa, p)$ is $\varkappa$-closed and hence closed in $\sum(\varkappa, p)$ because $\hat{t}(\sum(\varkappa, p))\leqq\varkappa$ by (2.7). We can thus select a basic open set U_p with $p\in U_p$ and $\overline{U}_p\cap B\cap\sum(\varkappa, p)=\emptyset$.

The set C is compact hence $C\subset\bigcup\{U_{p_i};\ i<n\}=U$ $(n<\omega)$.

The set $K=f(X-U)$ is compact in Y, $y\notin K$ so $y\in\overline{A}$ implies the existence of a point $a\in A-K$; if $H=f^{-1}(a)$, $H\subset U$.

The set H is compact and by $a\in A\subset S(\varkappa)$, $H\cap\sum(\varkappa, \vartheta)\neq\emptyset$ for each $\vartheta\in X$. By Lemma 5 choose a set $J\subset I$ s.t. $|J|<\varkappa, \pi_{I-J}(H)=X_{I-J}$. We can also suppose that $\bigcup\{I(U_{p_i});\ i<n\}\subset J$. The set $\pi_{I-J}(\overline{U}_{p_i}\cap H)$ is compact in X_{I-J} and by $\overline{U}_{p_i}\cap H\cap\sum(\varkappa, p_i)=\emptyset$ it is nowhere dense in the space X_{I-J} $(i<n)$. But $X_{I-J}=$ $=\bigcup\{\pi_{I-J}(\overline{U}_{p_i}\cap H);\ i<n\}$ so X_{I-J} would be equal to the union of finitely many nowhere dense subsets which is a contradiction.

Q.e.d.

Corollary 8. *If $\varkappa$ is regular, the set $\{y\in Y;\ \pi\chi(y, Y)<\varkappa\}$ is closed in Y.*

Proof. Denote the above set by P; by Theorem 4 $P\subset S(\varkappa)$. If $y\in\overline{P}$ then the compactness of $S(\varkappa)$ implies $y\in S(\varkappa)$; using that $\hat{t}(S(\varkappa))\leqq\varkappa$, we can choose a set $Q\subset P$ with $|Q|<\varkappa$ and $y\in\overline{Q}$. If $\mathscr{U}_q$ is a local π-base at q for each $q\in Q$, $|\mathscr{U}_q|<\varkappa$, then $\mathscr{U}=\bigcup\{\mathscr{U}_q;\ q\in Q\}$ is a local π-base at y and $|\mathscr{U}|<\varkappa$ so $\pi\chi(y, Y)<\varkappa$, $y\in P$. □

Theorem 9. *The following conditions are equivalent:*

a) *Y contains a subspace homeomorphic to $\mathbf{D}^{\varkappa}$.*

b) *Y can be continuously mapped onto $\mathbf{D}^{\varkappa}$.*

c) *There exists a point $y\in Y$ and a set $A\subset Y$ s.t. $y\in\overline{A}, a(y, A)\geqq\varkappa$, i.e., $\hat{t}(Y)\geqq\varkappa^+$.*

d) *There exists a point $y\in Y$ with $\pi\chi(y, Y)\geqq\varkappa$.*

e) *There exists a compact set $C\subset Y$ s.t. $\pi\chi(y, C)\geqq\varkappa$ for each $y\in C$.*

f) *If $Y=\bigcup\{C_\xi;\ \xi<\varkappa\}, C_\xi$ is compact $(\xi<\varkappa)$, then $\hat{t}(C_\xi)\geqq\varkappa^+$ for a suitable $\xi<\varkappa$.*

g) *If $Y=\bigcup\{C_\xi;\ \xi<\operatorname{cf}(\varkappa)\}$, C_ξ is compact, $C_\xi\subset C_\eta$ $(\xi<\eta<\operatorname{cf}(\varkappa))$, then $t(C_\xi)\geqq\varkappa$ for a suitable $\xi<\operatorname{cf}(\varkappa)$.*

h) *$H(\varkappa)\neq Y$.*

i) *$S(\varkappa)\neq Y$.*

j) *For each $\vartheta\in X$, $S(\varkappa, \vartheta)\neq Y$.*

For the proof we need two lemmas. The first one generalizes the well-known Baire Category Theorem.

Lemma 10. *Let E be a compactum, $\varkappa\geqq\omega$ a cardinal, $E=\bigcup\{A_\xi;\ \xi<\varkappa\}$. If $\lambda<\varkappa$ is a cardinal, $\mathscr{H}(\lambda)$ denotes the family of non-empty G_λ-sets in E (i.e., $H\in\mathscr{H}(\lambda)$ iff $H\neq\emptyset$ and H is the intersection of λ many open sets). Then there exists an ordinal $\xi_0<\varkappa$ and a set $H_0\in\mathscr{H}(|\xi_0|)$ s.t. if $H\subset H_0$, $H\in\mathscr{H}(|\xi_0|)$ then $H\cap A_{\xi_0}\neq\emptyset$.*

Proof. Assume the assertion is false. We shall define by transfinite induction a sequence of sets $\langle H_\xi;\ \xi<\varkappa\rangle$. If $\mu<\varkappa$ and for $\xi<\mu$ H_ξ is defined s.t.

(1) if $\xi<\mu$ then $H_\xi\in\mathscr{H}(|\xi|)$, $H_\xi\cap A_\xi=\emptyset$;

(2) if $\xi<\eta<\mu$ then $\overline{H}_\eta\subset H_\xi$,

put $H=\cap\{H_\xi;\ \xi<\mu\}$. The compactness of E implies that $H\neq\emptyset$ and so $H\in\mathscr{H}(|\mu|)$. From the indirect hypothesis we get a set $H'\in\mathscr{H}(|\mu|)$ with $H'\subset H$, $H'\cap A_\mu=\emptyset$. Finally, by the regularity of E, we can select a set H_μ, $H_\mu\in\mathscr{H}(|\mu|)$, $\bar{H}_\mu\subset H'$. The sequence $\langle H_\xi;\ \xi<\varkappa\rangle$ fulfils conditions (1), (2) for $\mu=\varkappa$. If $x\in\cap\{H_\xi;\ \xi<\varkappa\}$ then $x\in E-\cup\{A_\xi;\ \xi<\varkappa\}=\emptyset$, a contradiction. □

Remark. If in the lemma the sets $\{A_\xi;\ \xi<\varkappa\}$ are compacta then necessarily $H_0\subset A_{\xi_0}$ for the ordinal $\xi_0<\varkappa$ since if $x\in H_0-A_{\xi_0}$ then the set $H=H_0-A_{\xi_0}$ is an element of $\mathscr{H}(|\xi_0|)$ and $H\cap A_{\xi_0}=\emptyset$. □

The following lemma can be proved by a standard argument so we omit the (easy) proof.

Lemma 11. *Let E be a compactum, λ an infinite cardinal, $F\subset E$ closed, $F\in\mathscr{H}(\lambda)$. If for each point $x\in F$ $\pi\chi(x, E)>\lambda$ then $\pi\chi(x, F)>\lambda$ for each point $x\in F$.* □

Proof of Theorem 9. We proved $a\leftrightarrow c\leftrightarrow h$ in Theorem 3. The line of our reasoning will be

$$\begin{tikzcd}
a \arrow[r] & b \arrow[r] & f \arrow[r] & g \arrow[r] & i \arrow[r] & a \arrow[r] & j \arrow[d] \\
 & & & & e \arrow[u] & & d \arrow[ll]
\end{tikzcd}$$

a→b. $\mathbf{D}^\varkappa$ evidently can be mapped onto $[0, 1]^\varkappa$. The Tietze—Uryson extension theorem implies that such a mapping can be continuously extended onto Y. □

b→f. If $Y=\cup\{C_\xi;\ \xi<\varkappa\}$, C_ξ is compact $(\xi<\varkappa)$ and $f\colon Y\twoheadrightarrow[0, 1]^\varkappa$, put $K_\xi=f(C_\xi)$. Using Lemma 10 and the remark following it, we get an ordinal $\xi<\varkappa$ and a set $H_\xi\subset K_\xi$, $H_\xi\in\mathscr{H}(|\xi|)$. It is immediate that H_ξ contains a subspace homeomorphic to $[0, 1]^\varkappa$ but then $\hat{t}(C_\xi)\geqq\hat{t}(K_\xi)\geqq\hat{t}([0, 1]^\varkappa)=\varkappa^+$. □

f→g. Trivial.

g→i. Select a set H of cardinals s.t. $|H|\leqq\mathrm{cf}\,(\varkappa)$, for $\lambda\in H$ $\tau=\sup\{t(X_i);\ i\in I\}\leqq\lambda$ and $\varkappa=\sup\{\lambda^+;\ \lambda\in H\}$. (If $\varkappa=\lambda^+$ then $H=\{\lambda\}$ works.)

By Corollary 6 $S(\varkappa)=\cup\{S(\lambda^+);\ \lambda\in H\}$; by Theorem 7 $S(\lambda^+)$ is compact and $\hat{t}\big(S(\lambda^+)\big)\leqq\lambda^+$ hence $t\big(S(\lambda^+)\big)\leqq\lambda<\varkappa$. So condition g) implies $S(\varkappa)\neq Y$. □

i→a. The implication $i\rightarrow h$ is evident $\big(H(\varkappa)\subset S(\varkappa)\big)$ and $h\rightarrow a$ is already proved. □

a→j. Assume $\vartheta\in X$, $C\subset S(\varkappa, \vartheta)$ and C is homeomorphic to $\mathbf{D}^\varkappa$. We shall distinguish two cases.

Case 1. $\varkappa=\lambda^+$. By the Hewitt—Pondiczery—Marczewski Theorem ([10] p. 48) there exists a set $A\subset C$, $|A|\leqq\lambda$, $\bar{A}=C$. Choose a set $B\subset\sum(\varkappa, \vartheta)$ with $|B|\leqq\lambda$, $f(B)=A$. The cardinal ϑ being regular, $\sum(\varkappa, \vartheta)$ is $\varkappa$-closed hence $\bar{B}^X\subset\sum(\varkappa, \vartheta)$. But then $\hat{t}(D^\varkappa)=\hat{t}(C)\leqq\hat{t}(\bar{B}^X)\leqq\hat{t}\big(\sum(\varkappa, \vartheta)\big)\leqq\varkappa$, a contradiction.

Case 2. $\varkappa$ is a limit-cardinal. If $\xi<\varkappa$ put

$$A_\xi=\begin{cases} C\cap S(\lambda, \vartheta) & \text{if } \quad \xi=\lambda^+, \\ \lambda \text{ is a regular cardinal,} & \tau<\lambda<\varkappa; \\ \emptyset & \text{otherwise.}\end{cases}$$

By Corollary 6, $C=\cup\{A_\xi;\ \xi<\varkappa\}$ hence, by Lemma 10, there exists a regular cardinal λ, $\tau<\lambda<\varkappa$ and a set $H\in\mathscr{H}(\lambda^+)$ (taken in the subspace C), s.t. if $H'\subset H$, $H'\in\mathscr{H}(\lambda^+)$ then $H'\cap S(\lambda,\vartheta)\neq\emptyset$. For simplicity, in the sequel we shall identify C with the space $\mathbf{D}^\varkappa$.

We can now assume that $H=\cap\{\pi_\xi^{-1}(p(\xi));\ \xi<\eta\}$ where $p\in\mathbf{D}^\varkappa$, $\eta<\varkappa$, $|\eta|\leqq\lambda^+<\varkappa$ and choose a set $J\subset\varkappa-\eta$, $|J|=\lambda^+$. If $\mu\in J$ put $H_\mu=\{q\in H;$ for $v\in J$ $q(v)=1$ iff $v\leqq\mu\}$.

Since $H_\mu\subset H$, $H_\mu\in\mathscr{H}(\lambda^+)$, there exists a point $p_\mu\in H_\mu\cap S(\lambda,\vartheta)$ $(\mu\in J)$. Put

$$P_\mu=\{p_v;\ v\in J,\ v<\mu\}\subset\pi_\mu^{-1}(0)$$

$$P^\mu=\{p_v;\ v\in J,\ \mu\leqq v\}\subset\pi_\mu^{-1}(1).$$

This shows that $\langle p_v;\ v\in J\rangle$ is a free sequence of cardinality λ^+ in $S(\lambda,\vartheta)$; selecting a point $q_v\in f^{-1}(p_v)\cap\sum(\lambda,\vartheta)$ for each $v\in J$ we get a free sequence of cardinality λ^+ in $\sum(\lambda,\vartheta)$ but this contradicts to (2.8). □

j→d. By Theorem 4

$$\{y\in Y;\ \pi\chi(y,Y)<\varkappa\}\subset S(\varkappa)\quad\text{hence if}\quad S(\varkappa)\neq Y,$$

there exists a point $y\in Y$ with $\pi\chi(y,Y)\geqq\varkappa$. □

d→e. Assume d) is true. If $\varkappa$ is regular, the set $Q=\{y\in Y;\ \pi\chi(y,Y)\geqq\varkappa\}$ is non-empty and by Corollary 8, it is open. If $\varkappa$ is singular, Q is the intersection of $\mathrm{cf}(\varkappa)<\varkappa$ many open sets. In both cases Q contains a compact set C which is the intersection of less then $\varkappa$ open sets. By $C\subset Q$, $\pi\chi(y,Y)\geqq\varkappa$ for each point $y\in C$; hence Lemma 11 shows that $\pi\chi(y,C)\geqq\varkappa$ for each $y\in C$. □

e→i. Assume $S(\varkappa)=Y$, $C\subset Y$ is compact and $\pi\chi(y,C)\geqq\varkappa$ for each $y\in C$. Now $Y=S(\varkappa)=\cup\{S(\lambda^+);\ \lambda$ is an infinite cardinal, $\tau\leqq\lambda<\varkappa\}$. Using Lemma 10 we get a cardinal λ and a set $H\neq\emptyset$ s.t. $\tau\leqq\lambda<\varkappa$, H is compact, it is the intersection of less than $\varkappa$ open sets and $H\subset C\cap S(\lambda^+)$. By Lemma 11 $\pi\chi(y,H)\geqq\varkappa$ for each $y\in H$ hence $\pi\chi(H)\geqq\varkappa$.

On the other hand, $t(H)\leqq t(S(\lambda^+)\leqq\lambda<\varkappa$ hence for the compactum H $t(H)<\pi\chi(H)$ holds, and this contradicts a theorem of B. SHAPIROVSKIĬ [15].

Q.e.d.

It is a natural question whether we could replace the cardinal function $\pi\chi$ with χ in condition e) of the theorem. The answer in general is negative, as the following result of V. FEDORCHUK shows.

THEOREM (V. Fedorchuk [6]). *Assume the axiom of constructibility; then there exists a compactum C s.t. $t(C)=\omega$ but $\chi(x,C)=\omega_1$ for each $x\in C$.* □

Hence we are forced to assume some extra condition on the spaces $\{X_i;\ i\in I\}$. The following definition was suggested by a paper of A. Arhangelskiĭ [1].

DEFINITION 12. The topological space E is said to be *monolitic* if $w(\overline{A})\leqq|A|$ for each subset $A\subset E$.

It is easily seen that $\mathbf{A}_\lambda$ is monolitic for each cardinal λ.

THEOREM 13. *If the spaces* $\{X_i;\ i\in I\}$ *are also monolitic then conditions* a)—j) *of Theorem 9 are equivalent to the following conditions, too:*

k) *There exists a compact subspace* $C\subset Y$ *with* $\chi(y, C)\geqq\varkappa$ *for each* $y\in C$;

l) $Y\neq\bigcup\{\overline{K(\lambda)};\ \lambda\leqq\varkappa,\ \lambda \text{ is regular}\}$ *where* $K(\lambda)=\{y\in Y;\ \chi(y, Y)<\lambda\}$.

PROOF. d→l. Let $y_0\in Y$ be a point with $\pi\chi(y_0, Y)\geqq\varkappa$. If λ is a regular cardinal, $\tau=\sup\{t(X_i);\ i\in I\}<\lambda\leqq\varkappa$ then the set $P(\lambda)=\{y\in Y;\ \pi\chi(y, Y)<\lambda\}$ is closed, $y_0\notin P(\lambda)$. Since $K(\lambda)\subset P(\lambda)$ for each cardinal λ, $y_0\notin\overline{K(\lambda)}$. □

l→k. The proof is analogous to that of $d\to e$, so it is left to the reader. □

For the proof of the implication $k\to i$ we need two lemmas.

LEMMA 14. *Let* C *be a compactum,* $t(C)\leqq\lambda$, $\chi(x, C)>\lambda$ *for each* $x\in C$; *then there exists a set* $S\subset C$ *with* $|S|\leqq\lambda$, $\chi(\bar S)>\lambda$.

PROOF. Assume the assertion is false; we shall define by transfinite induction a sequence $\langle\langle p_\xi, H_\xi\rangle;\ \xi<\lambda^+\rangle$. If $\mu<\lambda^+$ and for $\xi<\mu$ $\langle p_\xi, H_\xi\rangle$ is already defined with the properties

1) $p_\xi\in H_\xi$, H_ξ is closed, $H_\xi\in\mathscr{H}(|\xi|+\omega)$;

2) If $\xi<\eta<\mu$ then $H_\xi\supset H_\zeta$;

3) $H_\xi\cap\overline{\{p_\eta;\ \eta<\xi\}}=\emptyset$, $(\xi<\mu)$, put $S=\{p_\xi;\ \xi<\mu\}$; $S\subset C$, $|S|\leqq\lambda$. If $K=\bigcap\{H_\xi;\ \xi<\mu\}$ then K is compact, $K\in\mathscr{H}(|\mu|+\omega)$. By the indirect hypothesis $K-\bar S\neq\emptyset$; choose a compact G_δ-set Q in C with $K\cap Q\neq\emptyset$, $Q\cap\bar S=\emptyset$ and put $H_\mu=K\cap Q$, $p_\mu\in H_\mu$. It is immediate that the sequence $\langle\langle p_\xi, H_\xi\rangle;\ \xi<\mu+1\rangle$ satisfies the conditions 1), 2), 3) and so we get a sequence $\langle\langle p_\xi, H_\xi\rangle;\ \xi<\lambda^+\rangle$ with the above properties. Note that the sequence $\langle p_\xi;\ \xi<\lambda^+\rangle$ is now a free sequence of cardinality λ^+ in C but this contradicts to the theorem of A. Arhangelskiĭ mentioned in the proof of (2.8). □

LEMMA 15. *If the spaces* $\{X_i;\ i\in I\}$ *are monolitic compacta,* λ *is an infinite cardinal,* $A\subset S(\lambda, \vartheta)$, $|A|<\lambda$ *then* $w(\bar A)<\lambda$.

PROOF. The easy proof is left to the reader. □

PROOF of k→i. Assume $C\subset Y$ is compact, $\chi(y, C)\geqq\varkappa$ for each $y\in C$ but $S(\varkappa)=Y$. As we saw it many times in the proof of Theorem 9, we can now select a non-empty compact set H in C and a cardinal λ, $\tau\leqq\lambda<\varkappa$ s.t. $\chi(y, H)\geqq\varkappa$ for each $y\in H$ and $H\subset S(\lambda^+)$. Now $t(H)\leqq t(S(\lambda^+))<\varkappa$; by Lemma 14 there exists a set $S\subset H\subset S(\lambda^+)$ with $|S|\leqq\lambda$, $\chi(\bar S)\geqq\lambda^+$. Using Lemma 15 we get $w(\bar S)<\lambda^+\leqq\chi(\bar S)$, an evident contradiction. □

REMARKS. 1. If $\varkappa$ is regular but it is not strongly inaccessible (i.e., there exists a $\lambda<\varkappa$ with $2^\lambda\geqq\varkappa$) then we can add to Theorem 13 also the condition

m) There exists a set $S\subset Y$ s.t. $|S|<\varkappa$ but $w(\bar S)\geqq\varkappa$.

Indeed, if $C\subset Y$, C is homeomorphic to $D^\varkappa$ and $2^\lambda\geqq\varkappa$ then, by the Hewitt—Pondiczery—Marczewski Theorem ([10] p. 48), a set $S\subset C$ can be selected with $|S|\leqq\lambda<\varkappa$, $\bar S=C$; hence $w(\bar S)=\varkappa$. On the other hand if $S(\varkappa, \vartheta)=Y$ for a $\vartheta\in X$ and $\varkappa$ is regular then by Lemma 15 if $|S|<\varkappa$ then $w(\bar S)<\varkappa$, too. □

149. 315.930

minta: piros, f

boritő m.

Studia
Scient.
Math.
13
1978

MTA

13
1998

2. If $\varkappa$ is regular then condition l) means that the set $K(\varkappa)=\{y\in Y;\ \chi(y, Y)<\varkappa\}$ is not dense in Y.

3. If $\varkappa$ is an isolated cardinal (i.e., $\varkappa=\lambda^+$) then c) is equivalent to the condition $t(Y)\geqq\varkappa$.

Note that if λ is any cardinal then $\mathbf{A}_\lambda$ is a monolitic compactum, $t(\mathbf{A}_\lambda)=\omega$ hence conditions a)$-$l) are equivalent for any $Y\in\mathcal{PC}$ if $\varkappa\geqq\omega_1$. We would like to obtain a similar result also for the case $\varkappa=\omega$.

THEOREM 16. *Assume $\{X_i;\ i\in I\}$ is a family of monolitic compacta, $t(X_i)=\omega$ $(i\in I)$ and $f\colon X=\prod\{X_i; i\in I\}\twoheadrightarrow Y$ is a mapping onto the T_2-space Y; then the following conditions are equivalent:*

a) *Y contains a subspace homeomorphic to $\mathbf{D}^\omega$.*

b) *Y contains a compact, dense-in-itself set.*

b′) *Y contains a dense-in-itself set.*

c) *Y can be continuously mapped onto $[0, 1]^\omega$.*

c′) *Y can be continuously mapped onto $[0, 1]$.*

d) *There exists a countable set $S\subset Y$ s.t. $\bar{S}$ is not countable.*

PROOF. a$\to$b, b$\leftrightarrow$b′, c$\to$c′ are evident for any compactum Y; c′$\to$c since $[0, 1]^\omega$ is a Peano-continuum.

b$\leftrightarrow$c′ is a theorem of E. PELCZYNSKI and Z. SEMADENI for any compactum Y [13]·

c′$\to$d. Choose a countable set $S\subset Y$ s.t. the image of S is dense in $[0, 1]$.

d$\to$a. By Remark 1 to our last theorem, $w(\bar{S})=\omega$ or Y contains a subspace homeomorphic to $\mathbf{D}^{\omega_1}$ (and hence also a subspace homeomorphic to $\mathbf{D}^\omega$). In the former case $\bar{S}$ is a non-countable compact metric space so by a classical result contains a copy of $\mathbf{D}^\omega$. □

§ 4

In this paragraph we shall always assume that E is an infinite polyadic compactum.

1. *Weight.* $w(E)=\max\big(c(E), t(E)\big)$.

PROOF. Here $\geqq$ is evident; suppose $c(E)\leqq\lambda$, $t(E)\leqq\lambda$. By the theorem of R. Engelking (§ 0, Theorem B), E is the continuous image of a product space $\mathbf{A}_\lambda^I$. E does not contain a topological copy of $\mathbf{D}^{\lambda^+}$ and $w(\mathbf{A}_\lambda)<\lambda^+$; hence, by Theorem A in § 0, $w(E)<\lambda^+$. □

COROLLARY 1. *If $c(E)\leqq\lambda$, $t(E)\leqq\varkappa$ and $\lambda\leqq\varkappa$ then E is the continuous image of $\mathbf{A}_\lambda^\varkappa$.* □

PROBLEM. Is it true that $c(E)=\lambda$, $t(E)=\varkappa$ implies that E is the continuous image of $\mathbf{A}_\lambda^\varkappa$?

COROLLARY 2. *If $t(E)=\varkappa$, $A\subset E$, $c(A)\leqq\varkappa$ then $w(\bar{A})\leqq\varkappa$.*

PROOF. By Corollary D, there exists an $Y\in\mathcal{PC}$, $A\subset Y\subset E$, $c(Y)\leqq\varkappa$, $t(Y)\leqq t(E)\leqq\varkappa$ so $w(Y)\leqq\varkappa$. □

2. *Tightness*

Here we have only the sup=max problem; i.e., if $t(E)=\varkappa$, can we choose $x\in E$, $A\subset E$ with $x\in\bar{A}$, $a(x, A)=\varkappa$, i.e., is $\hat{t}(E)=\varkappa^+$ true? As the following example shows the answer is in general negative.

EXAMPLE. Let $\varkappa=\sup\{\varkappa_\xi;\ \xi<\mathrm{cf}(\varkappa)\}$, $\omega\leqq\varkappa_\xi<\varkappa$ for $\xi<\mathrm{cf}(\varkappa)$ and denote E' the one-point compactification of the discrete topological sum of the spaces $\mathbf{D}^{\varkappa_\xi}$ $(\xi<\mathrm{cf}(\varkappa))$. Now $E'\in\mathscr{PC}$, $\hat{t}(E')=t(E')=\varkappa$. If $\lambda\geqq\mathrm{cf}(\varkappa)$, E is the discrete topological sum of the spaces E' and $\mathbf{A}_\lambda$ then $E\in\mathscr{PC}$, $c(E)=\lambda$, $t(E)=\hat{t}(E)=\varkappa$. □

However, if $c(E)=\lambda$, $t(E)=\varkappa$ and $\lambda<\mathrm{cf}(\varkappa)$ then such a space does not exist.

ASSERTION. *If* $E\in\mathscr{PC}$, $t(E)=\varkappa$ *and* $\hat{c}(E)\leqq\mathrm{cf}(\varkappa)$ *then* $\hat{t}(E)=\varkappa^+$.

PROOF. By a theorem of P. ERDŐS and A. TARSKI ([10], p. 37) $\sigma=\hat{c}(E)$ is always a regular cardinal. Applying now Theorem C, E is the continuous image of a product space $\prod\{X_i;\ i\in I\}$ where the X_i's are compacta, $w(X_i)<\sigma$ $(i\in I)$. Since $w(E)\geqq t(E)=\varkappa$ and $\sigma<\mathrm{cf}(\varkappa)$, we get from Theorem A that E contains a subspace homeomorphic to $\mathbf{D}^\varkappa$. □

3. *Cellularity*

Here, as at the tightness, only the sup=max problem is interesting. We know that if $c(E)=\hat{c}(E)=\lambda$ then λ is a regular limit cardinal.

PROPOSITION. *If* $c(E)=\hat{c}(E)=\lambda$ *then* $\hat{t}(E)\geqq\lambda^+$.

PROOF. Since $\hat{c}(E)=\lambda$, E is the continuous image of a product space $\prod\{X_i;\ i\in I\}$ where the X_i's are compacta, $w(X_i)<\lambda$ for $i\in I$. Since $w(E)\geqq c(E)=\lambda$, $\lambda=\mathrm{cf}(\lambda)$, E contains a subset homeomorphic to $\mathbf{D}^\lambda$ by Theorem A. □

4. *Discrete subspaces*

LEMMA. *Let* $t(E)=\varkappa>\omega$, *then there exist cardinals* $\{\varkappa_\xi;\ \xi<\tau\}$, $\varkappa=\sup\{\varkappa_\xi;\ \xi<\tau\}$ *and a subspace of E homeomorphic to the discrete topological sum of the spaces* $\{\mathbf{D}^{\varkappa_\xi};\ \xi<\tau\}$.

PROOF. This is certainly true if E contains a subspace homeomorphic to $\mathbf{D}^\varkappa$ so let us assume such a subspace does not exist; now $\varkappa$ is a limit cardinal.

Let $\varkappa=\sup\{\varkappa_\xi;\ \xi<\tau=\mathrm{cf}(\varkappa)\}$, $\varkappa_\xi$ regular, $\omega_1\leqq\varkappa_\xi<\varkappa_\eta<\varkappa$ if $\xi<\eta<\tau$. Put $C_\xi=S(\varkappa_\xi)$; by (3.7) C_ξ is compact, $t(C_\xi)\leqq\varkappa_\xi<\varkappa$ and if $\xi<\eta<\tau$ then $C_\xi\subset C_\eta$. Using (3.9) we get that $E=\bigcup\{C_\xi;\ \xi<\tau\}$. Denote $G_\xi=\mathrm{Int}\,C_\xi(\xi<\tau)$; if $\xi<\eta<\tau$ then $G_\xi\subset G_\eta$.

We shall distinguish two cases:

Case a). There are τ many different sets among the G_ξ's. We can evidently assume then that $G_\xi\neq G_{\xi+1}$ $(\xi<\tau)$. Put $U_\xi=G_{\xi+1}-C_\xi$; $\{U_\xi;\ \xi<\tau\}$ is a family of disjoint non-empty open sets in E. If $x\in U_\xi$ then $x\notin H(\varkappa_\xi)$ so by (3.3) E contains a subspace K_ξ homeomorphic to $\mathbf{D}^{\varkappa_\xi}$, $x\in K_\xi$. Note that a non-empty open subset of $\mathbf{D}^{\varkappa_\xi}$ contains a subspace homeomorphic to $\mathbf{D}^{\varkappa_\xi}$, so we can assume that $K_\xi\subset U_\xi$ $(\xi<\tau)$; but then the subspace $\bigcup\{K_\xi;\ \xi<\tau\}$ is a suitable subspace of E.

Case b). There exists an ordinal $\xi_0<\tau$ s.t. $G_\xi=G_{\xi_0}$ if $\xi_0\leqq\xi<\tau$. This means that in the open subspace $U=E-C_{\xi_0}$ $U-S(\sigma)$ is dense for each cardinal $\sigma<\varkappa$. If now U would contain a disjoint open system of cardinality τ then we should obtain, exactly as in the proof of Case a), a suitable subspace in U and hence also in E.

Consequently we can suppose that $\hat{c}(U)\leqq\tau$; using Corollary D we get a subspace Y with $U\subset Y\subset E$, $Y\in\mathscr{P}\mathscr{C}$, $\hat{c}(Y)\leqq\tau=\mathrm{cf}\,(\varkappa)$. Since $t(E)=\varkappa$, $E=C_{\xi_0}\cup Y$ and $t(C_{\xi_0})<\varkappa$, necessarily $t(Y)=\varkappa$. Now, the assertion in 2 shows that Y contains a copy of $\mathbf{D}^\varkappa$, a contradiction. □

COROLLARY 1. *If $E\in\mathscr{P}\mathscr{C}$ then E contains a discrete subspace H, $|H|=w(E)$.*

PROOF. Since $w(E)=\max\big(c(E),t(E)\big)$ and by the lemma there is a discrete subspace of cardinality $t(E)$, we must only to prove that there is one of cardinality $c(E)$, too. This is evidently true if $\hat{c}(E)>c(E)$; and if $\hat{c}(E)=c(E)$ then, by 3, $t(E)\geqq C(E)$. □

COROLLARY 2. *For a topological space R $o(R)$ denotes the number of open sets in R* [10]. *If $E\in\mathscr{P}\mathscr{C}$ then $o(E)=2^{W(E)}$.*

PROOF. Evidently $o(R)\leqq 2^{W(R)}$ for any R; on the other hand, if $H\subset E$ is discrete, $|H|=w(E)$, for each $x\in H$ choose an open set $G_x\subset E$, $G_x\cap H=\{x\}$. For any $T\subset H$ put $G_T=\bigcup\{G_x;\ x\in T\}$; the family $\mathscr{G}=\{G_T;\ T\subset H\}$ consists of open sets in E and $|\mathscr{G}|=2^{|H|}=2^{W(E)}$. □

5. *Density*

If λ is an infinite cardinal denote $\log\lambda=\min\{\sigma;\ 2^\sigma\geqq\lambda\}$.

PROPOSITION. $d(E)=\max\big(c(E),\log t(E)\big)$.

PROOF. a) $d(E)\geqq c(E)$; since $t(R)\leqq w(R)\leqq 2^{d(R)}$ is true for any regular space R, $d(E)\geqq\log t(E)$.

b) Let $\sigma=\max\big(c(E),\log t(E)\big)$; then $c(E)\leqq\sigma$, $t(E)\leqq 2^\sigma$. By the Corollary to 1, E is the continuous image of the product space $\mathbf{A}_\sigma^{2^\sigma}$. Since the Hewitt—Pondiczery—Marczewski Theorem ([10] p. 48) implies that the latter space has density σ, $d(E)\leqq\sigma$, too. □

6. For a topological space R $d\chi(R)$ denotes the minimal cardinal λ s.t. the set $\{x\in R;\ \chi(x,R)\leqq\lambda\}$ is dense in R.

In an analogous manner can be defined $d\pi\chi(R)$.

PROPOSITION. $d\chi(E)=\pi\chi(E)=d\pi\chi(E)=t(E)$.

PROOF. $t(E)\leqq\lambda$ is equivalent to $t(E)<\lambda^+$ so to the negation of condition c) in (3.9) for $\varkappa=\lambda^+$. By (3.9) this is equivalent to the negation of condition e), that is to $\pi\chi(E)<\lambda^+$ so to $\pi\chi(E)\leqq\lambda$. Since by (3.8) the set $\{x\in E;\ \pi\chi(x,E)<\lambda^+\}$ is closed in E, $d\pi\chi(E)<\lambda^+$ is equivalent to the negation of c). Finally, condition l) of (3.13) is not satisfied exactly if $d\chi(E)<\lambda^+$, i.e., if $d\chi(E)\leqq\lambda$. □

7. *Character*

The relation $\chi(E)=w(E)$ was proved in [7].

8. *Depth*

For a topological space R, the system $\mathcal{G}=\{G_\xi;\ \xi<\mu\}$ of open sets is said to be a *strongly descending family* if for any $\xi<\eta<\mu$ $\overline{G}_\eta \subsetneqq G_\xi$ holds.

The *depth* of R is defined by $k(R)=\sup\{|\mathcal{G}|;\ \mathcal{G}$ is a strongly descending family in $R\}$ [10].

Proposition. $k(E)=\omega$.

Proof. Assume $\{G_\xi;\ \xi<\omega_1\}$ is a strongly descending family in E. Put $C=\bigcap\{G_\xi;\ \xi<\omega_1\}=\bigcap\{\overline{G}_\xi;\ \xi<\omega_1\}$; the set C is closed in E. Selecting a point $x_\xi\in G_\xi-\overline{G}_{\xi+1}$ for $\xi<\omega_1$, a complete accumulation point x of $\{x_\xi;\ \xi<\omega_1\}$ lies in $\mathrm{Fr}\,(C)\subset C$.

By a theorem of S. Mrówka [13] if G is an open set in a polyadic compactum and $x\in\overline{G}$ then there exist a sequence $\langle x_n;\ n<\omega\rangle\subset G$ converging to x. Choose a sequence $\langle x_n;\ n<\omega\rangle\subset E-C$ converging to the point $x\in C$. Picking a $\xi<\omega_1$ with $\{x_n;\ n<\omega\}\cap G_\xi=\emptyset$, we get a contradiction. □

9. *The cardinality of the underlying set*

Lemma. *Let τ, σ denote infinite cardinals, $\vartheta\in\mathbf{A}_\tau^\tau$; then $|\sum(\sigma,\vartheta)|\leqq\tau^\sigma$.*

The easy proof is left to the reader. □

Let now $c(E)=\lambda$, $t(E)=\varkappa$. $\mathbf{D}^\varkappa\subset E$ denotes that E contains a copy of $\mathbf{D}^\varkappa$, its negation is denoted by $\mathbf{D}^\varkappa\not\subset E$.

Proposition.

a) *If $\lambda\leqq 2^\varkappa$ and $\mathbf{D}^\varkappa\subset E$ then $|E|=2^\varkappa$;*

b) *If $\lambda<2^\varkappa$ and $\mathbf{D}^\varkappa\not\subset E$ then $|E|=2^{\underset{\smile}{\varkappa}}$;*

c) *If $\lambda\geqq\varkappa$ and $\mathbf{D}^\varkappa\subset E$ then $\lambda\leqq|E|\leqq\lambda^\varkappa$;*

d) *If $\lambda\geqq\varkappa$ and $\mathbf{D}^\varkappa\not\subset E$ then $\lambda\leqq|E|\leqq\lambda^{\underset{\smile}{\varkappa}}$.*

Proof. We prove only b) and d) in case $\varkappa=\omega$; the proof of the other cases are similar (and simpler).

b) By $\lambda<2^{\underset{\smile}{\varkappa}}=\sup\{2^\sigma;\ \sigma<\varkappa\}$, there exists a cardinal $\sigma<\varkappa$ with $\lambda<2^\sigma$. Note that since $t(E)=\varkappa$ and $\mathbf{D}^\varkappa\not\subset E$, $\varkappa$ is a limit cardinal. By the Lemma in 4, $|E|\geqq 2^{\underset{\smile}{\varkappa}}$ hence we must only prove $|E|\leqq 2^{\underset{\smile}{\varkappa}}$.

Suppose first that $\lambda\geqq\varkappa$, then E is a continuous image of $\mathbf{A}_\lambda^\lambda$. If $\vartheta\in\mathbf{A}_\lambda^\lambda$ is arbitrary, $E=S(\varkappa,\vartheta)$ by (3.9). Hence $|E|\leqq|S(\varkappa,\vartheta)|\leqq|\sum(\varkappa,\vartheta)|\leqq\lambda^{\underset{\smile}{\varkappa}}$ using also the Lemma. But $\lambda^{\underset{\smile}{\varkappa}}\leqq(2^\sigma)^{\underset{\smile}{\varkappa}}=2^{\underset{\smile}{\varkappa}}$.

If $\lambda<\varkappa$ then, again by (3.9), $E=S(\varkappa)=\bigcup\{S(\sigma);\ \sigma$ is regular, $\lambda<\sigma<\varkappa\}$. It is proved in [9], Theorem 2 that for such a σ $w(S(\sigma))<\sigma$ so $|S(\sigma)|\leqq 2^\sigma$. This shows that $|E|\leqq\sum\{2^\sigma;\ \sigma<\varkappa\}=2^{\underset{\smile}{\sigma}}$.

d) If $\varkappa\geqq\omega_1$ then the proof is very similar to that given, so let $\varkappa=\omega$, $\mathbf{D}^\omega\not\subset E$, $c(E)=\lambda$; we must prove $|E|\leqq\lambda^{\underset{\smile}{\omega}}=\lambda$. By (3.16) E is dispersed, i.e., each subset of E contains an isolated point. But $w(E)\leqq\lambda\cdot\omega=\lambda$ according to 1, and so by a classical and well-known theorem $|E|\leqq\lambda$. □

Note that if $\lambda<\mathrm{cf}\,(\varkappa)$ then, by Theorem B and A, $\mathbf{D}^\varkappa\subset E$ hence $|E|=2^\varkappa$. □

REFERENCES

[1] ARHANGELSKIĬ, A. V., An approximation of the theory of dyadic bicompacta, *Dokl. Akad. Nauk SSSR* **184** (1969), 767—770 (in Russian).

[2] ARHANGELSKIĬ, A. V., On bicompacta satisfying Suslin condition hereditarily. Tightness and free sequences, *Dokl. Akad. Nauk SSSR* **199** (1971), 1227—1230 (in Russian).

[3] COMFORT, W., Compactness-like properties for generalized weak topological sum, *Pacific J. Math.* **60** (1975), 31—38.

[4] EFIMOV, B. A., Dyadic bicompacta, *Trudy Mosk. Mat. Obshch.* **14** (1965), 211—247 (in Russian).

[5] EFIMOV, B. A., On subspaces of dyadic bicompacta, *Dokl. Akad. Nauk SSSR* **185** (1969), 987—990 (in Russian).

[6] FEDORCHUK, V. V., The compatibility between some theorems of the general topology and certain axioms of the theory of sets, *Dokl. Akad. Nauk SSSR* **220** (1975), 786—789 (in Russian).

[7] GERLITS, J., On a problem of S. Mrówka, *Period. Math. Hungar.* **4** (1973), 71—80.

[8] GERLITS, J., Continuous functions on products of topological spaces, *Fund. Math.* **106** (1980), 67—75.

[9] GERLITS, J., On subspaces of dyadic compacta, *Studia Sci. Math. Hungar.* **11** (1976), 115—120.

[10] JUHÁSZ, I., *Cardinal functions in topology,* Amsterdam, 1971.

[11] MALYHIN, V. I., On the tightness and the Suslin number in exp X and in product spaces, *Dokl. Akad. Nauk SSSR* **203** (1972), 1001—1003 (in Russian).

[12] MARTY, R., On m-adic spaces, *Acta Math. Acad. Sci. Hungar.* **22** (1971), 441—447.

[13] MRÓWKA, S., Mazur Theorem and m-adic spaces, *Bull. Acad. Polon. Sci. Sér. Sci. Math.* **18** (1970), 299—305.

[14] PELCZYNSKI, A.—SEMADENI, Z., Spaces of continuous functions III, *Studia Math.* **18** (1959), 211—222.

[15] SHAPIROVSKIĬ, B., Canonical sets and character, density and weight in bicompacta, *Dokl. Akad. Nauk SSSR* **218** (1974), 58—61 (in Russian).

Mathematical Institute of the Hungarian Academy of Sciences
H—1053 Budapest, Reáltanoda u. 13—15.

(Received July 8, 1977)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 19—27.

ON THE THINNEST NON-SEPARABLE LATTICE OF CONVEX BODIES

by

E. MAKAI JR.

§ 1

G. FEJES TÓTH [4] posed the following problem: Find the thinnest n-dimensional lattice of spheres such that every k-dimensional subspace $(0 \leqq k \leqq n-1)$ intersects some closed sphere of the lattice. For $k=0$ we have a well-known problem: the problem of the thinnest lattice covering with spheres. We shall show that for $k=n-1$ the problem is equivalent to the problem of the densest lattice packing of spheres.

In fact we shall prove a more general theorem showing that the problem of the thinnest non-separable lattice of translates of a convex body D is equivalent to the problem of the densest lattice packing of another convex body E (the term *non-separable* refers to the property that every $(n-1)$-dimensional plane intersects the closure of some body of the lattice). This will enable us to give a lower (resp. upper) bound for the density of the thinnest non-separable lattice of an arbitrary (centrosymmetric) convex body. Earlier, L. FEJES TÓTH—E. MAKAI, JR [5] found in the plane the thinnest non-separable lattice of circles and the minimum of the densities of the thinnest non-separable lattices of arbitrary convex plates.

K. MAHLER [10] considered a problem which, as we shall show, is equivalent to our one. For centrosymmetric bodies he obtained implicitly a great part of our Theorem 1, explicitly he obtained in an equivalent formulation our Corollary 5, especially for $n=2$ our Theorem 3 (without the cases of equality).

For the case $0<k<n-1$ of G. Fejes Tóth's original question we also have lower estimates for the density (for $n=4, k=1$ it is presumably sharp), but they will be treated in another paper.

After completion of this paper I have been kindly informed by Prof. R. P. BAMBAH that two of his students, V. C. DUMIR and R. J. HANS-GILL also obtained our Theorem 1 and some related results.

§ 2

NOTATIONS. Let D be a convex body in R^n. Denote with D_c the convex body obtained from D by central symmetrization, i.e. $(D+(-D))/2$. Denote with D^* the polar of D with respect to the unit sphere centred at the centre of D, provided D is centrosymmetric.

If L is a lattice of points, $\Lambda(L, D)$ denotes the corresponding lattice of translates of D, i.e. $\{D+v, v \in L\}$. The lattice of points with integer coordinates will be denoted by L_0.

For a matrix A denote A^* the transpose of A.

Let $\varrho(D)$ denote the density of the thinnest non-separable lattice (more exactly the infimum of the densities of non-separable lattices) of translates of D, while $\delta(D)$ the density of the densest lattice packing of translates of D. Denote by $V(D)$ and $d(D)$ the volume, resp. the diameter of D.

§ 3

The main result of the paper is the following

THEOREM 1. *Let D be a convex body in R^n. Let A be a linear transformation (i.e., an $n\times n$ matrix) with $|A|\neq 0$. Denote by E the body $(D_c)^*/4$. Then $\Lambda(AL_0, D)$ is a (locally) thinnest non-separable lattice of translates of D if and only if $\Lambda(A^{*-1}L_0, E)$ is a (locally) densest lattice packing of translates of E. The densities $\varrho(D)$ and $\delta(E)$ satisfy the equality*

$$\varrho(D)=V(E)V(D)/\delta(E). \tag{1}$$

Observe that $\Lambda(AL_0, D)$ is the lattice of translates of D by vectors which are integer linear combinations of the column vectors of A. The lattice $A^{*-1}L_0$ is the so called *polar lattice* of the lattice AL_0 (cf. [2], § I.5).

REMARK. Earlier, Mahler [8] considered for centrosymmetric D for the lattice $\Lambda(L_0, D)$ the lattice $\Lambda(L_0, D^*)$. Essentially assuming $\Lambda(L_0, D^*)$ is a packing, but $\Lambda(L_0, (1+\varepsilon)D^*)$ is not (for any $\varepsilon>0$) he obtained lower and upper estimates for $\min\{r, \Lambda(L_0, rD)$ is a covering$\}$. However, [8] did not find the property of $\Lambda(L_0, D)$ exactly corresponding to the fact that $\Lambda(L_0, D^*)$ is a packing. Mahler [10] essentially showed the following (partly implicitly). Let D be a convex body with centre of symmetry at $\mathbf{0}$, and $V(D)/\|A\|\leqq 2^n V(E)V(D)/\delta(E)$. Then he asserts: there is a hyperplane with equation $(A^{-1}\mathbf{x}, \mathbf{p})=1$, where the components of $\mathbf{p}$ are integers, which does not intersect the interior of D. Further he states that this assertion is equivalent to the fact that D^* (centred at $\mathbf{0}$) does not contain in its interior non-zero points of the lattice $A^{*-1}L_0$. Later we shall see that his assertion is equivalent to non-separability of $\Lambda(AL_0, D/2)$ (Corollary 2).

The proof of Theorem 1 is based on the following

LEMMA 1. *For a convex body D $\Lambda(L_0, D)$ is non-separable if and only if $\Lambda(L_0, E)$ is a packing.*

PROOF. First we show that a hyperplane $(\mathbf{u}, \mathbf{x})=c$, where the coordinates of $\mathbf{u}$ are incommensurable, intersects some translate of D in $\Lambda(L_0, D)$. We can suppose D contains $\mathbf{0}$ in its interior. The distances of the vectors (k_i) of L_0 from the hyperplane $(\mathbf{u}, \mathbf{x})=c$ are given by $\left|\sum^n u_i k_i - c\right|/\|\mathbf{u}\|$. Since the coordinates of $\mathbf{u}$ are incommensurable, the numbers $\sum^n u_i k_i$ are dense on the line. So there will be lattice points of L_0 arbitrarily near to the hyperplane $(\mathbf{u}, \mathbf{x})=c$, which contradicts the fact that the spheres about the lattice points of L_0 of radius some $\varepsilon>0$ do not intersect the hyperplane $(\mathbf{u}, \mathbf{x})=c$.

Hence non-separability of $\Lambda(L_0, D)$ is equivalent to the fact that any hyperplane $(\mathbf{u}, \mathbf{x})=c$, where the coordinates of $\mathbf{u}$ are commensurable, intersects some closed translate of D in $\Lambda(L_0, D)$. Let us consider a hyperplane $(\mathbf{p}, \mathbf{x})=c$, where the coordinates of $\mathbf{p}$ are commensurable, which does not intersect the closure of any translate of D in $\Lambda(L_0, D)$. We can suppose the normal vector $\mathbf{p}$ of the hyperplane has integer coordinates p_i with greatest common divisor 1.

Let the width of D in the direction of $\mathbf{p}$ be $w_{\mathbf{p}}(D)$. The hyperplanes with normal vector $\mathbf{p}$ containing some lattice points of L_0 are equidistant, with distances $1/\|\mathbf{p}\|$. (In fact, the two neighbouring ones given by $(\mathbf{p}, \mathbf{x})=0$ and $(\mathbf{p}, \mathbf{x})=1$ contain $\mathbf{0}$ and $\mathbf{p}/\|\mathbf{p}\|^2$). Thus some hyperplane with normal vector (p_i), where the p_i-s are integers with greatest common divisor 1, does not intersect the closure of any translate of D in $\Lambda(L_0, D)$ if and only if

$$(2) \qquad 1/\|\mathbf{p}\| > w_{\mathbf{p}}(D) = w_{\mathbf{p}}(D_c) = 2r_{\mathbf{p}}^{-1}((D_c)^*),$$

where $r_{\mathbf{p}}((D_c)^*)$ is the length of the radius vector of $(D_c)^*$ from its centre in the direction of $\mathbf{p}$. Hence non-separability of $\Lambda(L_0, D)$ is equivalent to

$$(3) \qquad \|\mathbf{p}\| \geqq r_{\mathbf{p}}((D_c)^*)/2 = 2r_{\mathbf{p}}(E)$$

for every vector $\mathbf{p}\neq 0$ with integer components, having the greatest common divisor 1. Further this is equivalent to the fact, that $\Lambda(L_0, E)$ is a packing.

PROOF of the theorem.

We shall show that the following statements are equivalent: (1) $\Lambda(AL_0, D)$ is non-separable; (2) $\Lambda(L_0, A^{-1}D)$ is non-separable; (3) $\Lambda(L_0, A^*(D_c)^*/4)$ is a packing; (4) $\Lambda(A^{*-1}L_0, (D_c)^*/4)=\Lambda(A^{*-1}L_0, E)$ is a packing. The equivalence of (1) and (2), resp. (3) and (4) follow from the fact that the properties of being a packing, resp. being non-separable are affine invariant. The equivalence of (2) and (3) follows from Lemma 1 and the fact that $(A^{-1}D_c)^*=A^*(D_c)^*$ by [2], § IV.3.3, Corollary 3.

The product of the densities of $\Lambda(AL_0, D)$ and $\Lambda(A^{*-1}L_0, E)$ is $\frac{V(D)}{\|A\|}\cdot\frac{V(E)}{\|A^{*-1}\|}=$ $=V(D)V(E)$. Hence keeping in mind the equivalence of (1) and (4) we see that $\Lambda(AL_0, D)$ is a (locally) thinnest non-separable lattice if and only if $\Lambda(A^{*-1}L_0, E)$ is a (locally) densest lattice packing. The same considerations prove (1).

We have the following corollaries analogous to well-known results concerning packings.

COROLLARY 1. *$\Lambda(AL_0, D)$ is non-separable if and only if $\Lambda(AL_0, D_c)$ is non-separable.*

COROLLARY 2. *$\Lambda(AL_0, D)$ is non-separable if and only if each hyperplane with equation $(A^{-1}\mathbf{x}, \mathbf{p})=1$, where the components of $\mathbf{p}$ are integers, intersects the closure of the body $D+(-D)$, centred at $\mathbf{0}$.*

PROOF. It suffices to deal with $A=$identity, and in this case the proof of Lemma 1 proves our statement.

COROLLARY 3. *If $\Lambda(AL_0, D)$ is a locally thinnest non-separable lattice of translates of D, then there are at least $\binom{n}{2}$ non-parallel hyperplanes not containing interior points of any body of this lattice.*

Proof by [2], § V.8, Theorem VIII.

COROLLARY 4. *If $\Lambda(AL_0, D)$ is a non-separable lattice of translates of D, then there are at most $(3^n-1)/2$ non-parallel hyperplanes not containing interior points of any body of this lattice.*

Proof by [2], § V.8, Theorem IX.

REMARK. Since the densest lattice packing of spheres is known for $n\leqq 8$ (cf. [12] pp. 2—3 or in more details [7], § 39.5), the same can be said about the thinnest non-separable lattice of spheres. E.g., in R^3 for the unit sphere S the thinnest non-separable lattice is (up to a congruence) $\Lambda(AL_0, S)$ with

$$A=\frac{2}{\sqrt{3}}\begin{pmatrix}-1 & 1 & 1\\ 1 & -1 & 1\\ 1 & 1 & -1\end{pmatrix} \tag{4}$$

(which is eventually similar to the lattice of the thinnest lattice covering by spheres, i.e., a space-centred cubic lattice).

§ 4

Now we shall apply Theorem 1 to the case of a variable body to obtain estimates for $\min\limits_D \varrho(D)$ and $\max\limits_D \varrho(D)$. First we give the following simple

COROLLARY 5. *We have*

$$\begin{aligned}(5)\quad & \varrho(D)\geqq\min\{V(E)V(D),\ D \text{ is a convex body in } R^n\}\geqq\\ & \geqq\min\{V(E)V(D),\ D \text{ is a centrosymmetric convex body in } R^n\}\cdot 2^n\Big/\binom{2n}{n}\end{aligned}$$

while for centrosymmetric D

$$\varrho(D)\geqq\min\{V(E)V(D),\ D \text{ is a centrosymmetric convex body in } R^n\}. \tag{6}$$

In the centrosymmetric case this result was obtained in an equivalent formulation in [10].

PROOF. By [12], Theorem 2.4

$$V(E)V(D)=V(E)V(D_c)[V(D)/V(D_c)]\geqq V(E)V(D_c)2^n\Big/\binom{2n}{n}. \tag{7}$$

REMARK. Presumably (6) and the first inequality in (5) are sharp (see § 5); however, best known lower estimates of the right-hand side of (6) ([7], § 14.2 (11)) furnish for $\varrho(D)$ in both of the cases only a relatively weak lower bound. A better lower bound for large n will be given by our Theorem 2.

Lemma 2. *For any convex body D in R^n there is an affinity Φ with*

$$d^n(\Phi D) \leqq \lambda_n V(\Phi D) \tag{8}$$

where λ_n is a constant independent of D. For the centrosymmetric case the same holds with λ_n' instead of λ_n. We have

$$\lambda_n' \leqq 2^n n^{n/2}/\varkappa_n', \quad \lambda_n \leqq \binom{2n}{n} \lambda_n'/2^n \leqq \binom{2n}{n} n^{n/2}/\varkappa_n' \tag{9}$$

where $\varkappa_n'$ is the volume of the convex hull of the unit sphere about $\mathbf{0}$ *and* $(\pm\sqrt{n}, 0, \dots, 0)$.

Proof. An ellipsoid D_1 of minimal volume containing a centrosymmetric convex body D has a volume $\leqq V(D) n^{n/2} \varkappa_n/\varkappa_n'$ by John [6], where $\varkappa_n$ is the volume of the unit sphere. (He proves in fact that if the centre of D_1 is $\mathbf{0}$, the convex hull of $D_1/\sqrt{n}$ and two opposite points of the surface of D_1 belongs to D. Let $\mu \leqq \sqrt{n}$ be the minimal number for which this holds instead of $\sqrt{n}$; then

$$V(D_1) \leqq V(D)\varkappa_n/f(\mu) \leqq V(D) n^{n/2} \varkappa_n/\varkappa_n',$$

where $f(\mu)$ is the volume of the convex hull of the sphere about $\mathbf{0}$ of radius $1/\mu$ and $(\pm 1, 0, \dots, 0)$.) Let Φ transform D_1 into a sphere.

For non-centrosymmetric D $d^n(\Phi D) = d^n(\Phi D_c) \leqq \lambda_n' V(\Phi D_c) \leqq \lambda_n' V(\Phi D) \binom{2n}{n} \Big/ 2^n$ for the same Φ as for D_c by [12], Theorem 2.4.

By Lemma 2 we prove the following analogue for arbitrary n of Theorem 2 in [5], essentially with the same proof.

Theorem 2. *The density $\varrho(D)$ of the thinnest non-separable lattice of translates of D satisfies*

$$\varrho(D) \geqq \frac{\varkappa_n}{2^n \min\limits_{\Phi} [d^n(\Phi D)/V(\Phi D)] \delta_n} \geqq \frac{\varkappa_n}{2^n \lambda_n \delta_n} \geqq \frac{\varkappa_n \varkappa_n'}{2^n \binom{2n}{n} n^{n/2} \delta_n} \tag{10}$$

where Φ runs over all affinities, λ_n, $\varkappa_n'$ are from Lemma 2, $\varkappa_n$ denotes the volume of the unit sphere and δ_n the density of the densest lattice packing of spheres in R^n. For centrosymmetric bodies

$$\varrho(D) \geqq \frac{\varkappa_n}{2^n \lambda_n' \delta_n} \geqq \frac{\varkappa_n \varkappa_n'}{4^n n^{n/2} \delta_n}. \tag{11}$$

Proof. By Corollary 1 non-separability of $\Lambda(AL_0, D)$ implies that of $\Lambda(\Phi AL_0, S)$, where S is a sphere of diameter $d(\Phi D)$. Apply Theorem 1 to this lattice of spheres.

Remark. Besides $n=2$ the lower bound $\varkappa_n/2^n \lambda_n \delta_n$ for $\varrho(D)$ in (10) will be sharp for $n=3$, provided the value of λ_3 is the conjectured one (see § 5). The same cannot be told for any $n \geqq 4$, since no $n \geqq 4$ is known for which the densest lattice packing of spheres is generated by the edge-vectors of a regular simplex (compare § 5, Proposition 1).

REMARK. One obtains a lower estimate for $\varrho(D)$ analogous to the first inequality in (10) if one uses instead of spheres any fixed convex body D_0. This will be

$$\varrho(D) \geqq \frac{V(D)V(E_0)}{\min_{\Phi} \{\|\Phi\|, D_c \subset \Phi(D_0)_c\}\delta(E_0)} \tag{12}$$

where $E_0=((D_0)_c)^*/4$ and Φ runs over all affinities. (In fact by Corollary 1 non-separability of $\Lambda(AL_0, D)$ implies that of $\Lambda(AL_0, \Phi D_0)$.)

For $n=2$ both the non-centrosymmetric and the centrosymmetric cases are completely settled by the following

THEOREM 3. *For $n=2$ we have for any convex domain*

$$\varrho(D) \geqq \frac{3}{8} \tag{13}$$

and for any centrosymmetric convex domain

$$\varrho(D) \geqq \frac{1}{2}. \tag{14}$$

Equality holds in the first case only for the triangle, in the second case only for the parallelogram.

This result was obtained, by a different proof, in the non-centrosymmetric case in [5], Theorem 2. In the centrosymmetric case this result was obtained in an equivalent formulation in [10], excepting the cases of equality. In both cases all the thinnest non-separable lattices are easily obtained, see § 5.

PROOF. Using Theorem 1 $\varrho(D)\geqq V(E)V(D)=V(D)V((D_c)^*)/16\geqq 3/8$, and in the last inequality equality holds only for the triangle, while in the centrosymmetric case $\varrho(D)\geqq V(E)V(D)=V(D)V(D^*)/16\geqq 1/2$, and in the last inequality equality holds only for the parallelogram, by E. MAKAI JR. [11].

Now we turn to the estimation of $\varrho(D)$ from above, i.e., to the "minimax" problem, corresponding to the theorem of MINKOWSKI—HLAWKA (cf. e.g. [12], p. 8) in case of packings.

THEOREM 4. *We have*

$$\varrho(D) \leqq \varrho(D_c) \leqq \frac{\varkappa_n^2}{4^n \min_{D'} \delta(D')} \leqq \frac{\varkappa_n^2}{(n \log 2 - \text{const})\, 2^{n-1}}, \tag{15}$$

where the minimum is extended over all centrosymmetric convex bodies in R^n, and the last inequality holds if n is sufficiently large. Here $\varkappa_n$ and $\delta(D')$ denote the volume of the unit sphere and the density of the densest lattice packing of translates of D'.

PROOF. We have $V(D)\leqq V(D_c)$ (cf. e.g. [7], § 1.5, Theorem 7) and by [7], § 14.2, (12) $V(D_c)V(E)\leqq \varkappa_n^2/4^n$. Hence by Theorem 1

$$\varrho(D) = \frac{V(D)V(E)}{\delta(E)} \leqq \varrho(D_c) = \frac{V(D_c)V(E)}{\delta(E)} \leqq \frac{\varkappa_n^2}{4^n\delta(E)}. \tag{16}$$

The last inequality in (15) follows from [7], § 19.5, Theorem 8.

THEOREM 5. *For $n=2$ we have*

$$\varrho(D) \leqq \varrho(D_c) \leqq \frac{\pi^2}{16 \min\limits_{D'} \delta(D')} \leqq \frac{\pi^2}{4(3\sqrt{2}+\sqrt{3}-\sqrt{6})} = 0{,}6999 \ldots . \tag{17}$$

PROOF. Like at Theorem 4, referring to ENNOLA [3].

REMARK. Using the result announced by Ennola [3] the right-hand side of (17) can be diminished to 0,691....

§ 5

In this final paragraph we give some illustrative examples and some conjectures related to the theorems in § 4.

As to the quantities λ_n and λ_n' we state the following

CONJECTURE. The exact values of λ_n and λ_n' are equal to the reciprocal of the volume of the regular simplex of unit edge, i.e., $n!\,2^{n/2}/\sqrt{n+1}$, and the reciprocal of the volume of the regular cross-polytope of unit diameter, i.e., $n!$ (and the only extremal domains are the simplices and the cross-polytopes).

For $n=2$ these are proved (including the cases of equality) in BEHREND [1], p. 716, formula (II$_3$) and p. 715, formula (I), both in case of $\nu=5$. The question about λ_n' is equivalent to a question in JOHN [6] about the upper bound of the volume of a circumscribed ellipsoid of minimal volume of a centrosymmetric convex body of unit volume (since such an ellipsoid is known to have the same centre as the body).

About $\min \varrho(D)$ we give the following examples and conjectures.

PROPOSITION 1. *For $D=$simplex*

$$\varrho(D) = V(E)V(D) = \frac{n+1}{2^n n!}. \tag{18}$$

In case of a regular simplex of unit edge one of the thinnest non-separable lattices is the polar of the lattice Λ generated by the edge-vectors of the given regular simplex of unit edge.

PROOF. In (18) we have the second equality by [11]. The first equality follows by Theorem 1 from $\delta(E)=1$. $\delta(E)=1$ and the remainder of the proposition also follow from [11] (keeping in mind the reference to [2] in the proof of Theorem 1), where it is shown that if D is a regular simplex of unit edge, then E is the Dirichlet-cell of a lattice-point in the lattice Λ.

REMARK. Presumably this is the only thinnest non-separable lattice; presumably even the only space-filling with translates of E is obtained by the Dirichlet-cells of the lattice-points of Λ. For $n=2$ this is evident.

PROPOSITION 2. *For D=cross-polytope*

$$\varrho(D) = V(E)V(D) = \frac{1}{n!}. \tag{19}$$

In case of the regular cross-polytope given by $\sum^{n} |x_i| \leqq 1/2$ *the thinnest non-separable lattices are just the lattices* $\Lambda(AL_0, D)$, *where after a permutation of the coordinates* A *has only* 0-*s under the diagonal and* 1-*s in the diagonal.*

PROOF. E is a parallelotope, hence $\delta(E)=1$, so $\varrho(D)=V(E)V(D)$. $\varrho(D)$ is affine invariant, and for the regular case $V(E)V(D)=1/n!$. If D is given by $\sum^{n} |x_i| \leqq 1/2$ then E is given by $\max |x_i| \leqq 1/2$. So by [2], § IX, 1.3, Theorem IV all space-filling lattices of translates of E are given by $\Lambda(BL_0, E)$, where B is of the form given in the proposition (for A) — which holds if and only if $A=B^{*-1}$ is of this form.

CONJECTURE. For general convex bodies $\min \varrho(D)=(n+1)/2^n n!$ while for the centrosymmetric case $\min \varrho(D)=1/n!$ and the only domains for which the minima are attained are the simplices and the cross-polytopes.

REMARK. The general case would follow, including the case of equality, if the conjecture in [11], $V(D)V((D_c)^*) \geqq 2^n(n+1)/n!$, with equality only for simplices, would hold. The centrosymmetric case, not including the case of equality, would follow if the conjecture in [8], $V(D)V(D^*) \geqq 4^n/n!$ would hold. However, in this conjectured inequality equality holds not only for cross-polytopes (and parallelotopes) — see [11] for a conjecture about all cases of equality. However, most likely for the other cases of equality, i.e., when E is not a parallelotope, there is no space-filling lattice of translates of E.

Now we turn to the minimax problem. For large n this is unlikely to have a simple extremal domain. For $n=2$ we have for D = circle $\varrho(D)=\sqrt{3}\,\pi/8 = 0{,}6802\ldots$. The value given in Theorem 5 (and even more the value in the remark after it) is rather close to this value. For D = regular octagon $\delta(E)=4(3-\sqrt{2})/7 = 0{,}9062\ldots$ which is somewhat less than in case of the circle, i.e., 0,9069... (cf. MAHLER [9]); however, $\varrho(D)=1-\sqrt{2}/4=0{,}6464\ldots<0{,}6802\ldots$. Hence it is possible that there holds the following

CONJECTURE. For $n=2$ $\max\limits_{D} \varrho(D)$ is attained for an ellipse, i.e., it is equal to $\sqrt{3}\pi/8$.

REFERENCES

[1] BEHREND, F., Über einige Affininvarianten konvexer Bereiche, *Math. Ann.* **113** (1937), 713—747. *Zbl* **15**, 367.

[2] CASSELS, J. W. S., *An introduction to the geometry of numbers,* Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Band 99, Springer-Verlag, Berlin—Göttingen—Heidelberg, 1959. *MR* **28** # 1175.

[3] ENNOLA, V., On the lattice constant of a symmetric convex domain, *J. London Math. Soc.* **36** (1961), 135—138. *MR* **28** # 1177.

[4] Fejes Tóth, G., Research problem 18, *Period. Math. Hungar.* **7** (1976), 89—90.

[5] Fejes Tóth, L.—Makai, E., Jr., On the thinnest non-separable lattice of convex plates, *Studia Sci. Math. Hungar.* **9** (1974), 191—193. *MR* **51** # 6596.

[6] John, F., Extremum problems with inequalities as subsidiary conditions, Studies and essays presented to R. Courant, New York, 1948, 187—204. *MR* **10**, 719.

[7] Lekkerkerker, C. G., *Geometry of numbers,* Wolters—Noordhoff Publishing Co., Groningen, North-Holland Publishing Co., Amsterdam—London, 1969. *MR* **42** # 5915.

[8] Mahler, K., Ein Übertragungsprinzip für konvexe Körper, *Časopis Pěst. Mat. Fys.* **68** (1939), 93—102. *MR* **1**, 202.

[9] Mahler, K., On the minimum determinant and the circumscribed hexagons of a convex domain, *Proc. Kon. Ned. Akad. Wet.* **50** (1947), 692—703 (=Indag. Math. **9**, 326—337). *MR* **9**, 10.

[10] Mahler, K., Polar analogues of two theorems by Minkowski, *Bull. Austr. Math. Soc.* **11** (1974), 121—129. *MR* **50** # 7039.

[11] Makai, E., Jr., Areas of polar reciprocal plates (to appear).

[12] Rogers, C. A., *Packing and covering,* Cambridge University Press, 1964. *MR* **30** # 2405.

Mathematical Institute of the Hungarian Academy of Sciences
H—1053 Budapest, Reáltanoda u. 13—15

(Received July 20, 1977)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 29—42.

ON THE REMAINDER TERM OF THE PRIME NUMBER FORMULA IV. SIGN CHANGES OF $\pi(x)-\mathrm{li}\,x$

by

J. PINTZ

1. Riemann asserted (without proof) in his memoir [12] that for every $x>2$

$$\pi(x)<\mathrm{li}\,x. \tag{1.1}$$

His assertion was disproved by LITTLEWOOD [8] in 1914, who proved that the difference $\pi(x)-\mathrm{li}\,x$ changes sign infinitely many times. However, his proof was ineffective, so it was impossible to give any upper bound for the first sign change. After many attempts this problem was solved in 1955 by SKEWES [13] who proved the first explicit bound $e_4(7{,}705)$[1]. Another interesting problem was — for which Littlewood's original work had no answer — how often has $\pi(x)-\mathrm{li}\,x$ sign changes. Let us denote the number of sign changes of $\pi(x)-\mathrm{li}\,x$ in $[2, Y]$ by $V_1(Y)$. So the problem would be to give lower estimate for $V_1(Y)$. Such a result was first achieved by INGHAM [2], under an unproved and very deep condition however. Let us denote by θ the least upper bound of the real parts of the ζ-zeros. Then Ingham's theorem says, that if there exists a zero on the line $\sigma=\theta$, then $\pi(x)-\mathrm{li}\,x$ has a sign change in every interval

$$[Y, c_0Y] \tag{1.2}$$

where c_0 is an absolute, however, not effectively computable constant. From this one gets easily that

$$V_1(Y)>c_1\log Y \quad \text{for} \quad Y>Y_1 \tag{1.3}$$

with ineffective absolute constants c_1 [2] and Y_1.

The first unconditional lower bound for $V_1(Y)$ was proved in 1961—62 by S. KNAPOWSKI [3], [4]. He proved

$$V_1(Y)>c_2\log_2 Y \quad \text{for} \quad Y>Y_2 \tag{1.4}$$

with ineffective Y_2, further the weaker effective inequality

$$V_1(Y)>c_3\log_4 Y \quad \text{for} \quad Y>Y_3 \tag{1.5}$$

where both c_3 and Y_3 are explicitly calculable.

[1] We use the notations $e_1(x)=\exp(x)=e^x$, $e_{\nu+1}(x)=\exp\left(e_\nu(x)\right)$ and analogously with $\log x$.

[2] All the constants c_i are positive.

In 1974—1976 S. KNAPOWSKI and P. TURÁN [5], [6] showed the improvements of (1.4)—(1.5), namely

$$V_1(Y) > c_4 \frac{(\log Y)^{1/4}}{(\log_2 Y)^4} \quad \text{for} \quad Y > Y_4 \tag{1.6}$$

with ineffective Y_4 and

$$V_1(Y) > c_5 \log_3 Y \quad \text{for} \quad Y > Y_5 \tag{1.7}$$

with effective absolute constants c_5 and Y_5.

(1.7) was improved by the author [9] in 1976 to

$$V_1(Y) > c_6 (\log_2 Y)^{c_7} \quad \text{for} \quad Y > Y_6 \tag{1.8}$$

where c_6, c_7 and Y_6 are effective. In part III of this series [10] the effective lower bound

$$V_1(Y) > c_8 \frac{\sqrt{\log Y}}{\log_2 Y} \quad \text{for} \quad Y > Y_7 \tag{1.9}$$

was proved, which is better than (1.8) and the former best ineffective lower estimate (1.6).

Finally we mention that LEVINSON [7] showed in 1975

$$\overline{\lim_{Y \to \infty}} \frac{V_1(Y)}{\log Y} > 0 \tag{1.10}$$

which is the best result in this direction.

Let

$$\begin{aligned} \Delta_1(x) &\overset{\text{def}}{=} \pi(x) - \operatorname{li} x \overset{\text{def}}{=} \sum_{p \leqq x} 1 - \int_0^x \frac{dt}{\log t}, \\ \Delta_2(x) &\overset{\text{def}}{=} \Pi(x) - \operatorname{li} x \overset{\text{def}}{=} \sum_{p^m \leqq x} \frac{1}{m} - \operatorname{li} x, \\ \Delta_3(x) &\overset{\text{def}}{=} \Theta(x) - x \overset{\text{def}}{=} \sum_{p \leqq x} \log p - x, \\ \Delta_4(x) &\overset{\text{def}}{=} \psi(x) - x \overset{\text{def}}{=} \sum_{p^m \leqq x} \log p - x, \end{aligned} \tag{1.11}$$

and let $V_i(Y)$ denote (for $1 \leqq i \leqq 4$) the number of sign changes of $\Delta_i(x)$ in the interval $[2, Y]$.

With these notations we shall prove the partially ineffective

THEOREM. *There are absolute constants Y_i ($1 \leqq i \leqq 4$) such that for $Y > Y_i$ the inequality*

$$V_i(Y) > \frac{1}{10^{11}} \cdot \frac{\log Y}{(\log_2 Y)^3} \qquad (1 \leqq i \leqq 4) \tag{1.12}$$

holds, where the constants Y_i are effectively computable for $i = 2, 4$ and they are ineffective for $i = 1, 3$.

This is already very near to Ingham's conditional (and also ineffective) lower bound (1.3), however, here we cannot give a corresponding localization of a sign change. The inequality (1.12) is also not far from Levinson's result (1.10).

2. In the course of proof we shall use the following "kernel function":

$$\mathcal{I}_{k,\mu}(u)=\frac{1}{2\pi i}\int_{(2)}\left(\frac{e^s-e^{-s}}{2s}\right)^k e^{us+\frac{s^2}{\mu}}\,ds \tag{2.1}$$

where $k\geqq 1$ integer, $\mu\geqq 1$ real, and u is real.

First we state some properties of the $\mathcal{I}_{k,\mu}(u)$ functions as

LEMMA 2.1. *For the $\mathcal{I}_{k,\mu}(u)$ function defined by* (2.1) *we have*

$$\mathcal{I}_{k,\mu}(u)=\mathcal{I}_{k,\mu}(-u); \tag{2.2}$$

$$\mathcal{I}_{k,\mu}(u)\geqq 0; \tag{2.3}$$

$$\textit{if}\quad |u|\geqq k+2\quad \textit{then}\quad |\mathcal{I}_{k,\mu}(u)|\leqq\frac{1}{e^{(|u|-k-1)\mu}}. \tag{2.4}$$

For the proof of (2.2) we note that shifting the line of integration to $\sigma=0$ we get

$$\mathcal{I}_{k,\mu}(u)=\frac{1}{\pi}\int_0^\infty\left(\frac{\sin t}{t}\right)^k e^{-\frac{t^2}{\mu}}\cos(ut)\,dt \tag{2.5}$$

from which (2.2) follows.

Now it is enough to prove (2.4) for $u\geqq k+2$. Then shifting the line of integration to $\sigma=-\mu$ we get

$$|\mathcal{I}_{k,\mu}(u)|=\left|\frac{1}{2\pi i}\int_{(-\mu)}\left(\frac{e^s-e^{-s}}{2s}\right)^k e^{us+\frac{s^2}{\mu}}\,ds\right|=$$

$$=\left|\frac{1}{2^k}\sum_{d=0}^{k}(-1)^d\binom{k}{d}\cdot\frac{1}{2\pi i}\int_{(-\mu)}\frac{\exp\left\{(k-2d+u)s+\frac{s^2}{\mu}\right\}}{s^k}\,ds\right|\leqq$$

$$\leqq\max_{0\leqq d\leqq k}\left|\frac{1}{2\pi i}\int_{(-\mu)}\frac{\exp\left\{(k-2d+u)s+\frac{s^2}{\mu}\right\}}{s^k}\,ds\right|\leqq \tag{2.6}$$

$$\leqq\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\exp\left\{(k-2k+u)(-\mu)+\frac{\mu^2-t^2}{\mu}\right\}}{\mu^k}\,dt\leqq$$

$$\leqq\frac{\exp\{-\mu(u-k-1)\}}{\mu^k}\cdot\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-\frac{t^2}{\mu}}\,dt<\frac{1}{e^{\mu(u-k-1)}}.$$

Thus we have also

$$\lim_{u\to\infty}\mathcal{I}_{k,\mu}(u)=0. \tag{2.7}$$

So (2.3) will be proved considering (2.7) and (2.2) if we show that $\mathscr{I}_{k,\mu}(u)$ is monotonically decreasing for $u \geqq 0$.

This will be proved by induction with respect to k. Using the well-known formula

$$\frac{1}{2\pi i} \int_{(2)} e^{As^2+Bs}\, ds = \frac{1}{2\sqrt{\pi A}} \exp\left(-\frac{B^2}{4A}\right) \tag{2.8}$$

valid for real positive A and arbitrary complex B, we get for $k=1$

$$\begin{aligned} \frac{d\mathscr{I}_{1,\mu}(u)}{du} &= \frac{1}{2\pi i} \int_{(2)} \frac{e^s-e^{-s}}{2s} \cdot e^{\frac{s^2}{\mu}} \cdot s \cdot e^{us}\, ds = \\ &= \frac{1}{2}\left\{\frac{1}{2\pi i} \int_{(2)} e^{\frac{s^2}{\mu}+(u+1)s}\, ds - \frac{1}{2\pi i} \int_{(2)} e^{\frac{s^2}{\mu}+(u-1)s}\, ds\right\} = \\ &= \frac{\sqrt{\mu}}{4\sqrt{\pi}}\left\{\exp\left(-\frac{\mu}{4}(u+1)^2\right) - \exp\left(-\frac{\mu}{4}(u-1)^2\right)\right\} \leqq 0. \end{aligned} \tag{2.9}$$

If it is proved already for $k=k_0-1$ then we have

$$\begin{aligned} \frac{d\mathscr{I}_{k_0,\mu}(u)}{du} &= \frac{1}{\pi} \int_0^\infty \left(\frac{\sin t}{t}\right)^{k_0} (-t \sin(ut)) e^{-\frac{t^2}{\mu}}\, dt = \\ &= \frac{1}{\pi} \int_0^\infty \left(\frac{\sin t}{t}\right)^{k_0-1} \cdot \frac{1}{2}\{\cos((u+1)t) - \cos((u-1)t)\} e^{-\frac{t^2}{\mu}}\, dt = \\ &= \frac{1}{2}(\mathscr{I}_{k_0-1,\mu}(u+1) - \mathscr{I}_{k_0-1,\mu}(u-1)) \leqq 0 \end{aligned} \tag{2.10}$$

or alternately

$$= \frac{1}{2}(\mathscr{I}_{k_0-1,\mu}(1+u) - \mathscr{I}_{k_0-1,\mu}(1-u)) \leqq 0. \tag{2.11}$$

Thus (2.10) proves the assertion in case of $u \geqq 1$ and (2.11) in case of $0 \leqq u \leqq 1$, and so Lemma 1 is completely proved.

Our main tool in the proof will be Turán's method. Here we shall use the so called second main theorem in the special case when all the coefficients are equal to 1.

Lemma 2 (T. Sós—Turán). *For arbitrary complex numbers* z_j

$$\max_{m<\nu\leqq m+n} \frac{\left|\sum_{j=1}^{n} z_j^\nu\right|}{|z_1|^\nu} \geqq \left(\frac{1}{8e\left(\frac{m}{n}+1\right)}\right)^n. \tag{2.12}$$

The proof is contained in Vera T. Sós—P. Turán [14].

If we have only an upper bound N for the number of z_j's then we can define

$$z_j = 0 \quad \text{for} \quad n < j \leqq [N]. \tag{2.13}$$

This implies immediately the modified form of (2.12), namely the inequality

$$\max_{m < \nu \leqq m+N} \frac{\left|\sum\limits_{j=1}^{n} z_j^\nu\right|}{|z_1|^\nu} \geqq \left(\frac{1}{8e\left(\frac{m}{[N]}+1\right)}\right)^N. \tag{2.14}$$

We shall use Lemma 2 in this form.

Further we shall use some known properties of the zeta function which we state here.

The number of zeros with imaginary part between T and $T+1$

$$N(T+1)-N(T) < c\log T \quad \text{where} \quad c = 15 \quad \text{for} \quad T > T_0 \tag{2.15}$$

(see W. J. ELLISON—M. MENDÈS FRANCE [1] p. 165).

If $\zeta(s)\neq 0$ in the domain

$$\sigma > \beta, \quad |t| \leqq T+1 \tag{2.16}$$

then for

$$2 \geqq \sigma \geqq \beta+\eta, \quad 2 \leqq |t| \leqq T \tag{2.17}$$

one has

$$\left|\frac{\zeta'}{\zeta}(s)\right| = O\left(\frac{\log t}{\eta}\right). \tag{2.18}$$

(This follows easily from Satz 4.1 of PRACHAR [11], p. 225, in the special case $k=1$.)

Finally we shall use the standard estimate

$$\zeta(s) = O(\sqrt{t}) \quad \text{for} \quad \sigma \geqq \frac{1}{2}, \quad t \geqq 1. \tag{2.19}$$

3. First we shall treat the (ineffective) case $i=1$. If the Riemann hypothesis is true then the quoted theorem of Ingham (see (1.3)) already settles the problem. For the sake of completeness we mention that Ingham's theorem with essentially unchanged proof is valid for $2\leqq i\leqq 4$, too.

Thus we shall suppose the Riemann hypothesis to be false.

So let $\varrho_0=\beta_0+i\gamma_0$ be a zero with $\beta_0>\frac{1}{2}$ and with minimal $\gamma_0>0$. If there are more such zeros then let $\varrho_1'=\beta_1'+i\gamma_1'$ be the zero among those with maximal real part. If there is only one such zero then let $\varrho_1'=\varrho_0$.

Let denote successively $\varrho_{n+1}'=\beta_{n+1}'+i\gamma_{n+1}'$ the zero with maximal real part among those satisfying

$$\gamma_n' < \gamma \leqq \gamma_n'+2\log Y, \quad \beta \geqq \beta_n'+\frac{1}{\log Y} \tag{3.1}$$

if such a zero exists.

Thus we get after at most $\left[\frac{\log Y}{2}\right]$ steps a zero $\varrho'_N=\beta'_N+i\gamma'_N\overset{\text{def}}{=}\varrho_1=\beta_1+i\gamma_1$ with

(3.2) $$\beta_1>\frac{1}{2},\quad 0<\gamma_1<2\log^2 Y$$

(because $\log^2 Y+\gamma_1'<2\log^2 Y$ if $Y>Y_0$ ineffective constant), such that the domains

(3.3) $$0\leqq t\leqq \gamma_1,\quad \sigma>\beta_1$$

and

(3.4) $$|t-\gamma_1|\leqq 2\log Y,\quad \sigma\geqq\beta_1+\frac{1}{\log Y}$$

are zero-free.

4. Let us introduce the following notations. Let

(4.1) $$\mu\overset{\text{def}}{=}\log Y,\quad L\overset{\text{def}}{=}\log_2 Y.$$

Let k be any positive integer to be chosen later, for which

(4.2) $$4000L\leqq k\leqq 4400L.$$

Let λ be any real number satisfying

(4.3) $$\frac{\mu}{10^4L}\leqq\lambda\leqq\frac{2\mu}{10^4L}.$$

Let further

(4.4) $$A\overset{\text{def}}{=}\exp\{k(\lambda-2)\},$$

(4.5) $$B\overset{\text{def}}{=}\exp\{k(\lambda+2)\},$$

(4.6) $$g_{k,D}(u,s)\overset{\text{def}}{=}\left(\frac{e^s-e^{-s}}{2s}\right)^k e^{us+\frac{s^2}{D}},$$

(4.7) $$f(x)\overset{\text{def}}{=}\Pi(x)-\lg x\pm\sqrt{x}\overset{\text{def}}{=}\Pi(x)-\sum_{2\leqq n\leqq x}\frac{1}{\log n}\pm\sqrt{x},$$

(4.8) $$H(s)\overset{\text{def}}{=}\frac{\zeta'}{\zeta}(s)+\zeta(s)-1\mp\frac{1}{2\left(s-\frac{1}{2}\right)^2},$$

where both in (4.7) and in (4.8) the upper or in both the lower signs are meant.

Further we choose

(4.9) $$\mu'=\frac{\mu\cdot 4400L}{k}.$$

Thus by (4.2)

(4.10) $$\mu\leqq\mu'\leqq\frac{11}{10}\mu.$$

We shall prove that to every real λ satisfying (4.3) there exists an integer k satisfying (4.2) such that $f(x)$ has a sign change in $[A, B]$.

Let λ be fixed in (4.3) and let us assume in contrary that with any k in (4.2) $f(x)$ does not change his sign in $[A, B]$.

5. We shall start with the formula (valid for $\sigma > 1$)

$$\int_1^\infty f(x) \frac{d}{dx} (x^{-s} \log x)\, dx = H(s). \tag{5.1}$$

Replacing s by $s+i\gamma_1$ in (5.1), multiplying by $g_{k,\mu'}(k\lambda, s)$ and integrating with respect to s along the line $\sigma=2$, using (2.10) we get

$$\begin{aligned}
U &= \frac{1}{2\pi i} \int_{(2)} H(s+i\gamma_1) g_{k,\mu'}(k\lambda, s)\, ds = \\
&= \frac{1}{2\pi i} \int_{(2)} \int_1^\infty f(x) \frac{d}{dx} \left(x^{-s-i\gamma_1} \log x\, g_{k,\mu'}(k\lambda, s)\right) dx\, ds = \\
&= \int_1^\infty f(x) \frac{d}{dx} \left\{x^{-i\gamma_1} \log x \cdot \frac{1}{2\pi i} \int_{(2)} g_{k,\mu'}(k\lambda - \log x, s)\, ds\right\} dx = \\
&= \int_1^\infty f(x) \frac{d}{dx} \{x^{-i\gamma_1} \log x \mathscr{J}_{k,\mu'}(k\lambda - \log x)\}\, dx = \\
&= \int_1^\infty f(x) \left\{-i\gamma_1 \cdot \frac{x^{-i\gamma_1}}{x} \log x \cdot \mathscr{J}_{k,\mu'}(\log x - k\lambda) + \frac{x^{-i\gamma_1}}{x} \mathscr{J}_{k,\mu'}(\log x - k\lambda) + \right. \\
&\qquad \left. + x^{-i\gamma_1} \log x \cdot \mathscr{J}'_{k,\mu'}(\log x - k\lambda) \cdot \frac{1}{x}\right\} dx = \\
&= \int_1^\infty \frac{f(x) \log x \cdot x^{-i\gamma_1}}{x} \left\{\mathscr{J}_{k,\mu'}(\log x - k\lambda) \left(-i\gamma_1 + \frac{1}{\log x}\right) + \right. \\
&\qquad \left. + \frac{1}{2} \mathscr{J}_{k-1,\mu'}(\log x - k\lambda + 1) - \frac{1}{2} \mathscr{J}_{k-1,\mu'}(\log x - k\lambda - 1)\right\} dx.
\end{aligned} \tag{5.2}$$

Now we shall give an upper bound for the right side of (5.2) using the proved properties of the kernel-function $\mathscr{J}_{k,\mu}(u)$ and the fact that $f(x)$ does not change its sign in $[A, B]$ (defined by (4.4)—(4.5)). On the other hand we shall show that the left side can be reduced essentially to a finite powersum, for which we can give a non-trivial lower estimate by suitable choice of k within (4.2) using the second main theorem, which contradicts to the upper estimate sketched above.

6. To estimate U from above we shall split the integral U into three parts

$$U = U_1 + U_2 + U_3 \tag{6.1}$$

where

$$U_1 = \int_1^A, \quad U_2 = \int_A^B, \quad U_3 = \int_B^\infty. \tag{6.2}$$

Considering (3.2), (5.2), (2.3), (4.2)—(4.5) and that $f(x)$ does not change its sign in $[A, B]$ we get

$$|U_2| \leqq \int_A^B \frac{|f(x)|\,k(\lambda+2)}{x} \left\{ \frac{11}{10} \gamma_1 \mathscr{I}_{k,\mu'}(\log x - k\lambda) + \right.$$

$$\left. + \frac{1}{2} \mathscr{I}_{k-1,\mu'}(\log x - k\lambda + 1) + \frac{1}{2} \mathscr{I}_{k-1,\mu'}(\log x - k\lambda - 1) \right\} dx \leqq$$

$$\leqq \mu^3 \int_A^B \frac{|f(x)|}{x} \{ \mathscr{I}_{k,\mu'}(\log x - k\lambda) + \mathscr{I}_{k-1,\mu'}(\log x - k\lambda + 1) + \tag{6.3}$$

$$+ \mathscr{I}_{k-1,\mu'}(\log x - k\lambda - 1)\} \, dx =$$

$$= \mu^3 \left| \int_A^B \frac{f(x)}{x} \{ \mathscr{I}_{k,\mu'}(\log x - k\lambda) + \mathscr{I}_{k-1,\mu'}(\log x - k\lambda + 1) + \right.$$

$$\left. + \mathscr{I}_{k-1,\mu'}(\log x - k\lambda - 1)\} \, dx \right| .$$

Similarly we get owing to $f(x) \log x = O(x)$ introducing the new variable $u = \log x - k\lambda$, using (4.2)—(4.3) and (2.4)

$$|U_3| \leqq \mu^3 \int_B^\infty \{ \mathscr{I}_{k,\mu'}(\log x - k\lambda) + \mathscr{I}_{k-1,\mu'}(\log x - k\lambda + 1) +$$

$$+ \mathscr{I}_{k-1,\mu'}(\log x - k\lambda - 1)\} \, dx =$$

$$= \mu^3 \int_{2k}^\infty \{ \mathscr{I}_{k,\mu'}(u) + \mathscr{I}_{k-1,\mu'}(u+1) + \mathscr{I}_{k-1,\mu'}(u-1)\} e^{u+k\lambda} \, du \leqq \tag{6.4}$$

$$\leqq e^{\mu} \int_{2k}^\infty \{ e^{-(u-k-1)\mu'} + e^{-(u+1-k-1)\mu'} + e^{-(u-1-k-1)\mu'} \} e^u \, du \leqq$$

$$\leqq 3 \int_{2k}^\infty e^{-(u-k-2)\mu' + u} \, du = o(1)$$

and analogously

$$|U_1| = o(1). \tag{6.5}$$

Further, mutatis mutandis, we have

$$U_4 \overset{\text{def}}{=} \mu^3 \int_B^\infty \frac{f(x)}{x} \{ \mathscr{I}_{k,\mu'}(\log x - k\lambda) + \mathscr{I}_{k-1,\mu'}(\log x - k\lambda + 1) + \tag{6.6}$$

$$+ \mathscr{I}_{k-1,\mu'}(\log x - k\lambda - 1)\} \, dx = o(1)$$

and

$$U_5 \overset{\text{def}}{=} \mu^3 \int_1^A \frac{f(x)}{x} \{ \mathscr{I}_{k,\mu'}(\log x - k\lambda) + \mathscr{I}_{k-1,\mu'}(\log x - k\lambda + 1) + \tag{6.7}$$

$$+ \mathscr{I}_{k-1,\mu'}(\log x - k\lambda - 1)\} \, dx = o(1).$$

Now (6.3)—(6.8) give immediately with the notation

$$K \overset{\text{def}}{=} \int\limits_1^{\infty} \frac{f(x)}{x} \{\mathscr{I}_{k,\mu'}(\log x - k\lambda) + \mathscr{I}_{k-1,\mu'}(\log x - k\lambda + 1) + \\ + \mathscr{I}_{k-1,\mu'}(\log x - k\lambda - 1)\}\, dx \tag{6.8}$$

the relation

$$|U| = |U_2| + o(1) \leqq \mu^3 |K| + o(1). \tag{6.9}$$

7. In the following we shall estimate K from above using the formula (valid for $\sigma > 1$ with a constant h)

$$\int\limits_1^{\infty} \frac{f(x)}{x^{s+1}}\, dx = \frac{1}{s}\left\{\int\limits_2^{s}\left(\frac{\zeta'}{\zeta}(z) + \zeta(z)\right)dz + h\right\} \pm \frac{1}{s-\dfrac{1}{2}} = \\ \overset{\text{def}}{=} \varphi(s) \pm \frac{1}{s-\dfrac{1}{2}}, \tag{7.1}$$

which can be proved easily by partial integration.

Multiplying on both sides by $\dfrac{1}{2\pi i} G(s)$, where

$$G(s) \overset{\text{def}}{=} g_{k,\mu'}(k\lambda, s) + g_{k-1,\mu'}(k\lambda - 1, s) + g_{k-1,\mu'}(k\lambda + 1, s) \tag{7.2}$$

and integrating along the line $\sigma = 2$ one gets easily by (4.6) and (6.8) the formula

$$K = \pm \frac{1}{2\pi i} \int\limits_{(2)} \frac{G(s)}{s - \dfrac{1}{2}}\, ds + \frac{1}{2\pi i} \int\limits_{(2)} G(s)\varphi(s)\, ds = \\ \overset{\text{def}}{=} \pm \qquad K_1 \qquad + \qquad K_2. \tag{7.3}$$

Shifting the line of integration in K_1 to $\sigma = -\mu'$ we get by easy computation

$$K_1 = G\left(\frac{1}{2}\right) + \frac{1}{2\pi i} \int\limits_{(-\mu')} \frac{G(s)}{s - \dfrac{1}{2}}\, ds = \\ = O\left(e^{k + \frac{k\lambda}{2}}\right) + O\left(\int\limits_{-\infty}^{+\infty} \frac{\exp\left\{\mu' k - (k\lambda - 1)\mu' + \mu' - \dfrac{t^2}{\mu'}\right\}}{\mu'^k}\, dt\right) = \\ = O\left(e^{k + \frac{k\lambda}{2}}\right) + O\left(e^{-\mu'(k\lambda - k - 2)}\right) = O\left(e^{\frac{1}{2}k\lambda + k}\right). \tag{7.4}$$

In order to estimate the integral K_2 we transform it on the broken line l defined for $t \geqq 0$ by

$$(7.5)\quad \begin{aligned} &I_1\colon \sigma = \frac{5}{4} && \text{for} \quad t \geqq 2\mu \\ &I_2\colon \beta_1 + \frac{2}{\mu} \leqq \sigma \leqq \frac{5}{4} && \text{for} \quad t = 2\mu \\ &I_3\colon \sigma = \beta_1 + \frac{2}{\mu} && \text{for} \quad 10 \leqq t \leqq 2\mu \\ &I_4\colon \frac{1}{4} \leqq \sigma \leqq \beta_1 + \frac{2}{\mu} && \text{for} \quad t = 10 \\ &I_5\colon \sigma = \frac{1}{4} && \text{for} \quad 0 \leqq t \leqq 10 \end{aligned}$$

and for $t \leqq 0$ by reflection on the real axis, so that

$$(7.6)\quad K_2 = \frac{1}{2\pi i} \int_{(l)} G(s)\varphi(s)\,ds$$

because owing to (3.3)—(3.4) $\varphi(s)$ is regular right of l and on l.

Now using already mentioned well-known properties of $\zeta(s)$ (see (2.16)—(2.19)) and (3.2)—(3.4) (further the definitions (4.6), (7.2)) we have the following estimates for the integrals on I_ν $(1 \leqq \nu \leqq 5)$:

$$(7.7)\quad \begin{aligned} |\mathcal{I}_1| &= O\left(\left(\frac{2}{\mu}\right)^{k-2} \exp\left(\frac{5}{4}k\lambda - \frac{4\mu^2}{\mu'}\right)\right) = O(1), \\ |\mathcal{I}_2| &= O\left(\mu \log \mu \left(\frac{2}{\mu}\right)^{k-1} \exp\left(\frac{5}{4}k\lambda - \frac{4\mu^2}{\mu'}\right)\right) = O(1), \\ |\mathcal{I}_3| &= O\left(\mu^2 \log \mu \left(\frac{1}{5}\right)^{k-1} \exp\left(\left(\beta_1 + \frac{2}{\mu}\right) k\lambda\right)\right) = O(e^{\beta_1 k\lambda - k}), \\ |\mathcal{I}_4| &= O\left(\left(\frac{1}{5}\right)^{k-1} \exp\left(\left(\beta_1 + \frac{2}{\mu}\right) k\lambda\right)\right) = O(e^{\beta_1 k\lambda - k}), \\ |\mathcal{I}_5| &= O\left(e^{\frac{1}{4}k\lambda + 2k}\right). \end{aligned}$$

Hence

$$(7.8)\quad |K_2| = O(e^{(\beta_1\lambda - 1)k}) + O\left(e^{\left(\frac{\lambda}{4} + 2\right)k}\right).$$

Further using (7.3)—(7.4) we get

$$(7.9)\quad |K| = O(e^{(\beta_1\lambda - 1)k}) + O\left(e^{\left(\frac{\lambda}{2} + 1\right)k}\right).$$

Now with an ineffective absolute constant Y_1, by the definition of $\varrho_0=\beta_0+i\gamma_0$ and $\varrho_1=\beta_1+i\gamma_1$ (in **3**) for $Y>Y_1$ we have (owing to (4.1) and (4.3))

$$\beta_1-\frac{1}{2}\geqq\beta_0-\frac{1}{2}\geqq\frac{2\cdot 10^4\log_2 Y}{\log Y}=\frac{2\cdot 10^4 L}{\mu}\geqq\frac{2}{\lambda}. \tag{7.10}$$

Thus

$$e^{(\beta_1\lambda-1)k}\geqq e^{\left(\frac{\lambda}{2}+1\right)k} \tag{7.11}$$

and so

$$|K|=O(e^{\beta_1\lambda k-k}). \tag{7.12}$$

Combining this with (6.9) and (4.1) we get already the required upper estimate for the integral U in (5.2), namely we have

$$|U|\leqq\mu^3|K|+o(1)=O(e^{\beta_1\lambda k-k+3L}). \tag{7.13}$$

8. Now we can start with the lower estimate of the left side of U in (5.2) (with the choice of a suitable k). Shifting the line of integration to $\sigma=-\frac{1}{2}$ we get

$$U=\sum_{\varrho} g_{k,\mu'}(k\lambda,\varrho-i\gamma_1)\mp \tag{8.1}$$

$$\mp\frac{1}{2}\frac{d}{ds}\left(g_{k,\mu'}(k\lambda,s)\right)_{s=\frac{1}{2}-i\gamma_1}+\frac{1}{2\pi i}\int\limits_{(-\frac{1}{2})} H(s+i\gamma_1)\,g_{k,\mu'}(k\lambda,s)\,ds.$$

Easy computation shows that the last integral is $O(1)$ further the second term is

$$\mp\frac{1}{2}\left(\frac{e^{\frac{1}{2}-i\gamma_1}-e^{-\frac{1}{2}+i\gamma_1}}{1-2i\gamma_1}\right)^k e^{k\lambda\left(\frac{1}{2}-i\gamma_1\right)+\left(\frac{1}{2}-i\gamma_1\right)^2\cdot\frac{1}{\mu}}\cdot$$

$$\cdot\left\{k\frac{\frac{d}{ds}\left(\frac{e^s-e^{-s}}{2s}\right)_{s=\frac{1}{2}-i\gamma_1}}{\frac{e^{\frac{1}{2}-i\gamma_1}-e^{-\frac{1}{2}+i\gamma_1}}{1-2i\gamma_1}}+k\lambda+\frac{1-2i\gamma_1}{\mu}\right\}= \tag{8.2}$$

$$=O\left(k\lambda e^{\frac{k\lambda}{2}}\right)=O\left(e^{\frac{k\lambda}{2}+k}\right)=O(e^{\beta_1\lambda k-k})$$

if we use the inequality (7.11), too.

In the first term (in (8.1)) we can trivially estimate the sum containing the infinitely many zeros with

$$|\gamma-\gamma_1|\geqq 2\mu. \tag{8.3}$$

Namely, by (2.15), we have for the contribution of these zeros the upper estimate

$$2\sum_{n\geqq[2\mu]} c\log(\gamma_1+n)\left(\frac{2}{n}\right)^k e^{k\lambda+\frac{1-n^2}{\mu'}}=O(1). \tag{8.4}$$

Similarly we can easily estimate the sum corresponding to the zeros with

$$6\leqq|\gamma-\gamma_1|<2\mu. \tag{8.5}$$

The number of these zeros is owing to (2.15) and (3.2) at most

$$4\mu c \log(\gamma_1+2\mu) \leqq 4c\mu \log(\mu^2+2\mu) = O(\mu^2). \tag{8.6}$$

Further, by (3.4), we have for the zeros with (8.5)

$$\beta \leqq \beta_1+\frac{1}{\mu} \tag{8.7}$$

and so for these zeros

$$|g_{k,\mu'}(k\lambda,\, \varrho-i\gamma_1)| \leqq \left(\frac{e+1}{2\cdot 6}\right)^k e^{k\lambda\left(\beta_1+\frac{1}{\mu}\right)+\frac{1}{\mu'}} < e^{k\lambda\beta_1-k}. \tag{8.8}$$

Thus we get for the contributions of zeros with (8.5) to the sum (8.1) the upper bound

$$O(\mu^2 e^{k\lambda\beta_1-k}) = O(e^{k\lambda\beta_1-k+2L}). \tag{8.9}$$

9. These estimates were naturally independent from the choice of k in (4.2). So the essential part of U is the finite powersum, containing the zeros with

$$|\gamma-\gamma_1| < 6. \tag{9.1}$$

The number n of such zeros is owing to (2.15) and (3.2)

$$1 \leqq n \leqq 2\cdot 6\cdot c \log(\gamma_1+6) \leqq 180 \log(2\log^2 Y+6) \leqq 400 \log_2 Y = 400L. \tag{9.2}$$

So for the zeros with (9.1) we can use Lemma 2 in the form given in (2.14). Thus choosing

$$m = 4000L \tag{9.3}$$

we get a positive integer k satisfying (4.2) for which

$$\begin{aligned} |W| &= \Big|\sum_{|\gamma-\gamma_1|<6} g_{k,\mu'}(k\lambda,\, \varrho-i\gamma_1)\Big| = \\ &= \left|\sum_{|\gamma-\gamma_1|<6} \left\{\frac{e^{\varrho-i\gamma_1}-e^{-(\varrho-i\gamma_1)}}{2(\varrho-i\gamma_1)}\cdot e^{\lambda(\varrho-i\gamma_1)}+\frac{(\varrho-i\gamma_1)^2}{4400L\mu}\right\}^k\right| \geqq \\ &\geqq \left(\frac{1}{8e\cdot 12}\right)^{400L} \left(\frac{e^{\beta_1}-e^{-\beta_1}}{2\beta_1}\right)^k e^{k\lambda\beta_1+\frac{\beta_1^2 k}{4400L\mu}} > \\ &> \left(\frac{1}{e^6}\right)^{400L}\cdot e^{k\lambda\beta_1} = e^{k\lambda\beta_1-2400L} \end{aligned} \tag{9.4}$$

(because for real $x>0$ one has $e^x-e^{-x}>2x$).

Now the estimate $O(1)$ for the integral, the upper bound (8.2) for the residue in (8.1), further the inequalities (8.4) and (8.9) concerning the zeros with $|\gamma-\gamma_1|\geqq 6$ give together with the lower bound (9.4) that owing to (4.2) we have for $|U|$ the lower estimate

$$|U| \geqq \frac{1}{2}\, e^{k\lambda\beta_1-2400L} \tag{9.5}$$

which contradicts to (7.13) (again owing to (4.2)). Thus we got that to every λ with (4.3) there exists an integer k with (4.2), such that $f(x)$ in (4.7) and thus also $\Delta_1(x)$ and naturally $\Delta_2(x)$, too, has a sign change in the interval

$$[A, B] = [e^{k\lambda-2k}, e^{k\lambda+2k}]. \tag{9.6}$$

This in itself does not give still the required inequality (1.12), i.e. the assertion of the theorem.

10. However, we can notice that as the total Lebesgue measure of λ's in (4.3) is

$$\frac{\mu}{10^4 L} \tag{10.1}$$

and k can take at most $[400L]+1$ values, there must exist a fixed k_0 with (4.2) for which there are λ's with Lebesgue measure at least

$$\frac{\mu}{10^4 L}\cdot\frac{1}{401L} = \frac{\mu}{4.01\cdot 10^6 L^2} \tag{10.2}$$

such that $\Delta_1(x)$ has a sign change in

$$[A_\lambda, B_\lambda] = [e^{k_0\lambda-2k_0}, e^{k_0\lambda+2k_0}] \subset (e^{k_0\lambda-10^4 L}, e^{k_0\lambda+10^4 L}). \tag{10.3}$$

But as the Lebesgue measure of the λ's belonging to this fixed k_0 is at least the quantity given by (10.2) we can choose among them at least

$$N = \left[\frac{\mu}{4.01\cdot 10^6 L^2}\cdot\frac{1}{2\cdot 10^4 L}\right] > \frac{\mu}{10^{11} L^3} \tag{10.4}$$

λ_j's $(1 \leqq j \leqq N)$, such that the difference of any two of them would be

$$|\lambda_j - \lambda_\nu| \geqq 2\cdot 10^4 L \quad (1 \leqq \nu < j \leqq N) \tag{10.5}$$

and so for the corresponding intervals $[A_{\lambda_j}, B_{\lambda_j}]$ we have by (10.3)

$$[A_{\lambda_j}, B_{\lambda_j}] \cap [A_{\lambda_\nu}, B_{\lambda_\nu}] = \emptyset \quad (1 \leqq \nu < j \leqq N). \tag{10.6}$$

Owing to (4.2)—(4.5) to every k and λ satisfying (4.2) and (4.3) resp. the corresponding interval

$$[A, B] = [e^{k(\lambda-2)}, e^{k(\lambda+2)}] \subset [e^{0,3\mu}, e^{0,9\mu}] = [Y^{0,3}, Y^{0,9}] \subset [2, Y]. \tag{10.7}$$

So, considering (10.4), (10.6) and (10.7), we get at least

$$\frac{\mu}{10^{11} L^3} = \frac{\log Y}{10^{11}(\log_2 Y)^3} \tag{10.8}$$

disjoint intervals, contained in $[2, Y]$ such that $\Delta_1(x)$ changes its sign in each of these intervals, and thus we finished the proof of Theorem 1.

11. In the case $i=2$ the following slight changes are necessary in the course of proof to get the inequality (1.12) for $Y > Y_2$ effective constant.

Here we do not make any difference whether the Riemann hypothesis is supposed to be true or not, and choose $\varrho_0 = \frac{1}{2} + i\gamma_0$ as the zero with the minimal imaginary part ($\gamma_0 \approx 14.13$). Thus we get a zero $\varrho_1=\beta_1+i\gamma_1$ with the properties described in (3.2)—(3.4) with the only change that instead of $\beta_1>\frac{1}{2}$ we have only $\beta_1\geqq\frac{1}{2}$. In (4.7) we modify the exponent $\frac{1}{2}$ of x to $\frac{1}{4}$ and correspondingly we define $H(s)$ in (4.8) with $\frac{1}{4}\left(s-\frac{1}{4}\right)^{-2}$ in the last term instead of $\frac{1}{2}\left(s-\frac{1}{2}\right)^{-2}$. Thus we get for the K_1 in (7.4) the upper bound

$$|K_1| = O\left(e^{k+\frac{k\lambda}{4}}\right) \tag{11.1}$$

and so we get without (7.10) and (7.11) immediately (7.12), i.e.

$$|K| = O(e^{\beta_1\lambda k-k}). \tag{11.2}$$

Further we get for the residue in (8.1) in the point $s=\frac{1}{4}-i\gamma_1$ the upper estimate $O\left(e^{\frac{k\lambda}{4}+k}\right)=O(e^{\beta_1\lambda k-k})$ and the other parts of the proof remain again valid without any change.

The cases $i=3$ and $i=4$ can be treated similarly to the cases $i=1$ and $i=2$, they are even easier, so we do not work them out.

REFERENCES

[1] ELLISON, W. J. and MENDÈS FRANCE, M.: *Les nombres premiers,* Paris, Hermann, 1975.

[2] INGHAM, A. E.: A note on the distribution of primes, *Acta Arith.* **1** (2), (1936), 201—211.

[3] KNAPOWSKI, S.: On the sign changes in the remainder term in the prime number formula, *Journ. Lond. Math. Soc.* **36** (1961), 451—460.

[4] KNAPOWSKI, S.: On the sign changes of the difference $(\pi(x)-\text{li}\, x)$, *Acta Arith.* **7** (2) (1962), 107—120.

[5] KNAPOWSKI, S. and TURÁN, P.: On the sign changes of $(\pi(x)-\text{li}\, x)$, I. *Topics in Number Theory, Coll. Math. Soc. János Bolyai* **13.**, North-Holland P. C., Amsterdam—Oxford—New York, 1976, pp. 153—169.

[6] KNAPOWSKI, S. and TURÁN, P.: On the sign changes of $(\pi(x)-\text{li}\, x)$, II., *Monatshefte für Math.* **82** (1976), 163—175.

[7] LEVINSON, N.: On the number of sign changes of $\pi(x)-\text{li}\, x$, *Topics in Number Theory, Coll. Math. Soc. János Bolyai* **13.**, North-Holland P. C., Amsterdam—Oxford—New York 1976, pp. 171—177.

[8] LITTLEWOOD, J. E.: Sur la distribution des nombres premiers, *C. R. Acad. Sci. Paris* **158** (1914), 1869—1872.

[9] PINTZ, J.: Bemerkungen zur Arbeit von S. Knapowski und P. Turán, *Monatshefte für Math.* **82** (1976), 199—206.

[10] PINTZ, J.: On the remainder term of the prime number formula III. Sign changes of $\pi(x)-\text{li}\, x$, *Studia Sci. Math. Hungar.* **12** (1977), 345—369.

[11] PRACHAR, K.: *Primzahlverteilung,* Berlin—Göttingen—Heidelberg, 1957.

[12] RIEMANN, B.: Über die Anzahl der Primzahlen unter einer gegebenen Grösse, *Monatsh. Preuss. Akad. Wiss.,* Berlin, 1859, pp. 671—680.

[13] SKEWES, S.: On the difference $\pi(x)-\text{li}\, x$, II, *Proc. London Math. Soc.,* **5** (1955), 48—70.

[14] T. SÓS, VERA and TURÁN, P.: On some new theorems in the theory of diophantine approximation, *Acta Math. Acad. Sci. Hungar.* **6** (1955), 241—255.

Mathematical Institute of the Hungarian Academy of Sciences,
Budapest, Reáltanoda u. 13—15, Hungary 1053

(Received January 10, 1978)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 43—45.

ON AN ISOPERIMETRIC PROBLEM

by

J. PACH

Scatter in the plane a finite number of line segments. The region *enclosed* by the segments is defined as the set of those points which cannot be connected by a Jordan arc with the exterior of the convex hull of the segments, without intersecting at least one segment. How should the segments be arranged so as to maximize the area of the region enclosed by them? Specializing a more general problem of L. FEJES TÓTH [1], G. HAJÓS raised the following question. Prove or disprove the intuitively obvious conjecture that, in an extremal arrangement, the region enclosed by the segments is a simple polygon.

Fejes Tóth [2] claimed to have proved the above conjecture in the special case when only polygonal arrangements are compared, i.e., arrangements in which each endpoint of a segment is the endpoint of exactly one other segment, with the additional condition that no subset of the segments has this property. Soon after the publication of [2], Fejes Tóth observed that his proof was wrong and he called my attention to the problem of finding a correct proof.

Our result is contained in the following

THEOREM 1. *Among all polygons of given side lengths the convex polygon inscribed into a circle has the greatest area.*

Here the area of a polygon is defined as the area enclosed by the sides of the polygon.

The convex hull of the vertices of a polygon P will be denoted by conv P. Our Theorem 1 is an immediate consequence of the following stronger result:

THEOREM 2. *Let P and P_c be two polygons with the same lengths of sides, and suppose that P_c is a convex polygon inscribed into a circle. Then the area of* conv P *is at most that of P_c.*

Equality holds here if and only if P is also a convex polygon inscribed into a circle.

REMARK. For polygons not intersecting themselves the proof of Theorem 2 is easy and well-known (see e.g. [3]).

The proof of Theorem 2 is based on the following simple

LEMMA. *Let Q be a strictly convex polygon. Suppose that the area of the triangle xyz is minimum among the triangles spanned by the vertices of Q. Then two sides of Δxyz lie on the boundary of Q.*

AMS (MOS) subject classifications (1970). Primary 52A40; Secondary 50B15.
Key words and phrases. Isoperimetric problem, polygon, convex hull.

PROOF of the Lemma. Let h_{xy} denote the altitude belonging to the side xy in Δxyz. Let S_{xy} be the parallel-strip consisting of those points which have a distance less than h_{xy} from the line xy. Obviously, any vertex of Q (different from x and y) is outside S_{xy}. The forbidden strips S_{xz}, S_{yz} can be defined similarly. Thus the vertices of Q, different from x, y and z, can be placed only in the six remaining angular regions $T_1, T_2, \ldots, T_6$ (see Figure 1). There are no vertices of Q in T_4, T_5 and T_6. For supposing the contrary, there is a vertex x', say, in T_4. This means that x is inside the triangle $x'yz$, which contradicts the fact that x is a vertex of Q. A similar argument shows that there are vertices in at most one of T_1, T_2, T_3 (say in T_1). In this case the segments xy and yz lie on the boundary of Q.

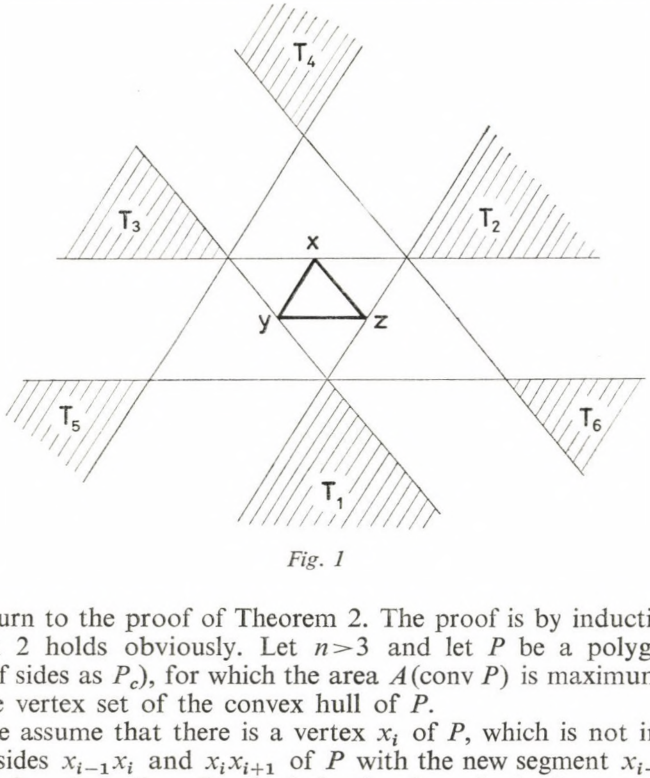


Fig. 1

Now we turn to the proof of Theorem 2. The proof is by induction on n. For $n=3$ Theorem 2 holds obviously. Let $n>3$ and let P be a polygon (with the same lengths of sides as P_c), for which the area $A(\operatorname{conv} P)$ is maximum. $V(\operatorname{conv} P)$ will denote the vertex set of the convex hull of P.

1. First we assume that there is a vertex x_i of P, which is not in $V(\operatorname{conv} P)$. Replacing the sides $x_{i-1}x_i$ and x_ix_{i+1} of P with the new segment $x_{i-1}x_{i+1}$ we get a polygon P' of $n-1$ vertices. By the induction hypothesis, for a convex polygon P'_c inscribed into a circle, having the same lengths of sides as P', we have $A(P'_c) \geqq A(\operatorname{conv} P') = A(\operatorname{conv} P)$. Replacing in P'_c the side corresponding to $x_{i-1}x_{i+1}$ with an arc congruent to $x_{i-1}x_ix_{i+1}$ we get a polygon P'' for which

$$(1) \qquad A(\operatorname{conv} P'') \geqq A(P'_c) + A(x_{i-1}x_ix_{i+1}) \geqq A(\operatorname{conv} P) + A(x_{i-1}x_ix_{i+1})$$

holds, where $A(x_{i-1}x_ix_{i+1})$ denotes the area of the triangle $x_{i-1}x_ix_{i+1}$. By the maximality of $A(\operatorname{conv} P)$ we have $A(\operatorname{conv} P'') \leqq A(\operatorname{conv} P)$. Comparing this with (1) we obtain

$$(2) \qquad A(x_{i-1}x_ix_{i+1}) = 0, \quad A(\operatorname{conv} P'') = A(\operatorname{conv} P).$$

But P'' does not intersect itself, thus taking our Remark into account, P'' ought to be a convex polygon inscribed into a circle. This is, however, impossible because the first part of (2) shows that three points of P'' lie on a straight line.

2. Suppose now that every vertex of P belongs to V (conv P). Using our Lemma for conv P we can choose a triangle $x_i x_j x_k$ of minimum area, whose sides $x_i x_j$ and $x_j x_k$ lie on the boundary of conv P. Let $x_{j-1}x_j$ and $x_j x_{j+1}$ be those sides of P which meet in the vertex x_j. Replacing the arc $x_{j-1}x_j x_{j+1}$ with the new segment $x_{j-1}x_{j+1}$ we get a polygon P' of $n-1$ vertices. Let us consider a convex polygon P'_c inscribed into a circle, having the same lengths of sides as P'. Using again the induction hypothesis we get

$$A(P'_c) \geqq A(\operatorname{conv} P') = A(\operatorname{conv} P) - A(x_i x_j x_k). \tag{3}$$

Replacing in P'_c the side corresponding to $x_{j-1}x_{j+1}$ with an arc congruent to $x_{j-1}x_j x_{j+1}$ we get a polygon P'' for which

$$A(\operatorname{conv} P'') \geqq A(P'_c) + A(x_{j-1}x_j x_{j+1}) \tag{4}$$

holds. Using (3) and the minimality of $A(x_i x_j x_k)$, we can write (4) in the form

$$A(\operatorname{conv} P'') \geqq A(\operatorname{conv} P) + A(x_{j-1}x_j x_{j+1}) - A(x_i x_j x_k) \geqq A(\operatorname{conv} P). \tag{5}$$

The polygon P'' does not intersect itself, hence, by the Remark, we have

$$A(\operatorname{conv} P'') \leqq A(P_c), \tag{6}$$

where P_c is a convex polygon inscribed into a circle, having the same lengths of sides as P and P''. Putting (5) and (6) together we obtain A (conv $P) \leqq A(P_c)$. It can easily be seen that equality holds here if and only if P is a convex polygon which can be inscribed into the same circle as P_c. This completes the proof of Theorem 2.

REFERENCES

[1] Fejes Tóth, L., Über das Didosche Problem, *Elemente der Mathematik* **23** (1968), 97—101.
[2] Fejes Tóth, L., Research problem No. 6, *Period. Math. Hungar.* **4** (1973), 231—232.
[3] Fejes Tóth, L., On the isoperimetric problem, *Mat. Lapok* **1** (1950), 363—383 (in Hungarian).

Mathematical Institute of the Hungarian Academy of Sciences
H—1053, Budapest, Reáltanoda u. 13—15

(Received March 10, 1978)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 47—51.

NEAT REDUCTS OF VARIETIES

by

H. ANDRÉKA and I. NÉMETI

Neat reducts play an important role in algebraic logic, cf. [8], [1], [2]. Neat reducts of cylindric algebras and related structures are a central tool in the representation theory of cylindric algebras and related structures.

The problem of finding a recursive enumeration of the quasiequations valid in all neat reducts of a given variety is frequently investigated in algebraic logic, cf. as above (also in [9], [7]).

Here we give a universal algebraic definition of neat reducts such that all the above quoted neat reduct concepts are special cases of the present one. We shall see that the quasiequations defining the class of all neat reducts of an arbitrary variety are recursively enumerable. The recursive enumerability of the equational theory of representable cylindric algebras is a corollary of the present statement. Other corollaries are the analogous results concerning neat reducts in other versions of algebraic first order logic, e.g. Modal—Cylindric algebras [6], and "systems of varieties definable by schemes of equations", cf. [3]. Also T.3.3, T.3.5, T.3.15, C.3.16 of [1] are corollaries of the present statement. Let t be a similarity type and let $t_1 \subseteq t$ be a part of t. If $\mathfrak{A}$ is an algebra of type t then $\mathbf{Rd}_{t_1}\mathfrak{A}$ denotes the t_1-type reduct of $\mathfrak{A}$. Similarly for any class K of algebras of type t:

$$\mathbf{Rd}_{t_1} K \overset{\mathrm{d}}{=} \{\mathbf{Rd}_{t_1}\mathfrak{A} : \mathfrak{A} \in K\}.$$

S denotes the formation of subalgebras, i.e. $\mathbf{SRd}_{t_1} K$ consists of all t_1-type subreducts of K. Clearly, if V is a quasivariety of type t then $\mathbf{SRd}_{t_1} V$ is a quasivariety again.

Let C be a set of equations of type t. Then for any algebra $\mathfrak{A}$ and $B \subseteq A$ the following property is meaningful: C is valid in $\mathfrak{A}$ relative to B, cf. [5] p. 242 and p. 509:

We shall say that C is valid in $\mathfrak{A}$ relative to B iff any valuation of the variables of C into B satisfies C. In other words: any substitution (or valuation) of the variables by elements of B is a *solution* of the set C of equations in the algebra $\mathfrak{A}$.

For a subreduct $\mathfrak{B}$ of $\mathfrak{A}$ we say that $\mathfrak{B}$ satisfies the equations C in $\mathfrak{A}$ iff C is valid in $\mathfrak{A}$ relative to B (where B is the universe of $\mathfrak{B}$). Such a subreduct $\mathfrak{B}$ is called a *C-neat subreduct* of $\mathfrak{A}$.

Recall that $t_1 \subseteq t$. Let V be a variety of type t, let C be a set of equations of type t (*not* necessarily valid in V). The collection of t_1-type subreducts of elements of V satisfying the equations C is denoted by $\mathbf{Sr}^C_{t_1} V$ and is called the class of *C-neat subreducts* of V. Formally:

$$\mathbf{Sr}^C_{t_1} V \overset{\mathrm{d}}{=} \{\mathfrak{B} \in \mathbf{SRd}_{t_1} V : C \text{ is valid in } \mathfrak{A} \text{ relative to } B \text{ for some } \mathfrak{A} \in V \text{ and } \mathfrak{B} \in \mathbf{Rd}_{t_1}\mathfrak{A}\}.$$

EXAMPLES. 1. [8] p. 401 Def. 2.6.28.

Let t be the similarity type of $\alpha+\omega$ dimensional cylindric algebras. ($\mathbf{CA}_{\alpha+\omega}$ denotes the variety of these cylindric algebras.) Let t_1 be the similarity type of α dimensional cylindric algebras.

Define C as $C\overset{\mathrm{d}}{=}\{c_i x=x\colon\ \alpha\leqq i<\alpha+\omega\}$. $\mathbf{Sr}^C_{t_1}\mathbf{CA}_{\alpha+\omega}$ is the well-known class "$\mathbf{SNr}_\alpha\mathbf{CA}_{\alpha+\omega}$ of *neat* subreducts of cylindric algebras".

2. [1] p. 28.

Let t be the similarity type of $\alpha+\omega$ dimensional substitution algebras. ($\mathbf{SA}_{\alpha+\omega}$ is the variety of these algebras.) Let t_1 be the similarity type of α dimensional substitution algebras and let C be as in Example 1. Now $\mathbf{Sr}^C_{t_1}\mathbf{SA}_{\alpha+\omega}$ is the class $\mathbf{SNr}_\alpha\mathbf{SA}_{\alpha+\omega}$ of representable substitution algebras.

3. [6], [3], [1] Cor. 3.17.

The above example applied to Modal—Cylindric algebras yields their neat subreducts; and more generally in case of an arbitrary system $\langle V_\alpha\rangle_{\alpha\in\mathrm{Ord}}$ of varieties definable by a scheme of equations it yields the quasivariety of neat subreducts of $V_{\alpha+\omega}$.

4. Let t consist of two binary operation symbols $+$ and $\dot{-}$. Let

$$E\overset{\mathrm{d}}{=}\{(x+y)+z=x+(y+z)\},\quad C\overset{\mathrm{d}}{=}\{(x+y)\dot{-}x=y\}.$$

Now, if t_1 consists of $+$ only, then $\mathbf{Sr}^C_+\mathbf{Md}(E)$ is the quasivariety of left cancellative semigroups. Where $\mathbf{Md}(E)$ is the t-type variety defined by the equations E.

5. Let R be the variety of rings with a derived operation d. (E.g. $d(x,y)=$ $=(x+-y)^2$).

Let $C=\{x\cdot x=x\}$ and let t_1 consist of 0, 1 and d. Now, $\mathbf{Sr}^C_{t_1}R$ is a quasivariety of algebras consisting of idempotent elements of rings.

6. Let R be the variety of rings with the derived operation $p(x,y)=$ $=d(d(x,y),(x\cdot y))$ where d is as above. Let $C=\{x\cdot x=x\}$ and t_1 consist of "$\cdot$" and p. Now $\mathbf{Sr}^C_{t_1}R$ is a quasivariety containing the variety of Boolean algebras (when $\cdot$ and p are interpreted as meet and join, respectively).

THEOREM 1. *Let $t_1\subseteq t$ be two similarity types, let V be a variety of type t defined by a set E of equations. Let C be a set of equations of type t. Now* (i) *and* (ii) *below hold.*

(i): $\mathbf{Sr}^C_{t_1}V$ *is a quasivariety.*

(ii): *If E and C are recursively enumerable then a recursive enumeration of the quasiequations defining $\mathbf{Sr}^C_{t_1}V$ can be obtained from the enumerations of E and C.*

PROOF. Part (ii) of the theorem will be proved as a corollary of Theorem 2. Here we prove only part (i).

To show that $\mathbf{Sr}^C_{t_1}V$ is a quasivariety, it is enough to prove that $\mathbf{Sr}^C_{t_1}V$ is closed w.r.t. reduced products and subalgebras.

By definition it is closed w.r.t. subalgebras: $\mathfrak{N}\subseteq\mathfrak{B}\subseteq\mathbf{Rd}_{t_1}\mathfrak{A}$ such that C is valid in $\mathfrak{A}$ relative to B implies that C is also valid in $\mathfrak{A}$ relative to N since $N\subseteq B$.

Let $\mathop{\mathbf{P}}_{j\in J}\mathfrak{B}_j/D$ be a reduced product and $\mathfrak{B}_j\in\mathbf{Sr}^C_{t_1}\mathfrak{A}_j$, $\mathfrak{A}_j\in V$ for every $j\in J$. Take the canonical embedding:

$$\eta\colon \mathop{\mathbf{P}}_{j\in J}\mathfrak{B}_j/D \overset{\sim}{\rightarrowtail} \mathop{\mathbf{P}}_{j\in J}\mathfrak{A}_j/D\in V.$$

Since reduced products preserve equations, C is valid in $\mathop{\mathbf{P}}_{j\in J}\mathfrak{A}_j/D$ relative to the image of $(\mathop{\mathbf{P}}_{j\in J}B_j/D)$ by η. *QED*

Let Y denote the (infinite) set of variables (of E, C etc.). Let I be a completely new infinite set of constant symbols. t' denotes the expansions of the type t by the new constants I. Similarly t_1' is the expansion of t_1 by I.

C' denotes the set of equations obtainable from C by substituting all the variables by constant t_1'-terms. More precisely:

Let Fr_{It_1} denote the set of terms without variables in type t_1'.

Any $\xi\colon Y\to Fr_{It_1}$ induces a substitution, such that if e is an equation (with all its variables in Y) then $\xi(e)$ is another equation obtained from e by replacing each variable $x\in Y$ in it by $\xi(x)$. Obviously, $\xi(e)$ contains no variables and belongs to the type t'. Now,

$$C' \overset{\mathrm{d}}{=} \{\xi(e)\colon e\in C \text{ and } \xi\in {}^{Y}Fr_{It_1}\}.$$

Notation. $\xi\colon Y\rightarrowtail I$ denotes that ξ is a one-to-one map of Y into the set I.

THEOREM 2. *For any quasiequation* $(\bigwedge_{i\leqq n} e_i\to e)$ *of type* t_1: $\mathbf{Sr}^C_{t_1}V\models(\bigwedge_{i\leqq n} e_i\to e)$ *iff* $E\cup C'\cup\{\xi(e_i)\}_{i\leqq n}\models\xi(e)$ *for some* $\xi\colon Y\rightarrowtail I$.

PROOF. Let $\xi\colon Y\rightarrowtail I$ be arbitrary.

$E\cup C'\cup\{\xi(e_i)\}_{i\leqq n}\models\xi(e)$ iff: For any homomorphism f from the word-algebra $\mathfrak{Fr}_{It}$ generated by I (in similarity type t) into V i.e. for any $f\colon\mathfrak{Fr}_{It}\to\mathfrak{A}\in V$, it is true that $\langle\mathfrak{A}, f(r)\rangle_{r\in I}\models(\bigwedge_{i\leqq n}\xi(e_i)\to\xi(e))$ if f satisfies C' i.e. if $\ker f\supseteqq C'$.

1. Now we prove $\mathbf{Sr}^C_{t_1}V\models(\bigwedge_{i\leqq n} e_i\to e)$ implies $E\cup C'\cup\{\xi(e_i)\}_{i\leqq n}\models\xi(e)$ for every $\xi\colon Y\rightarrowtail I$.

Let $\xi\colon Y\to I$ be arbitrary. Let $f\colon\mathfrak{Fr}_{It}\to\mathfrak{A}\in V$ be such that $\ker f\supseteqq C'$. Since $Fr_{It}\supseteqq Fr_{It_1}$ we can define $B\overset{\mathrm{d}}{=}f(Fr_{It_1})$. This is a subuniverse of $\mathbf{Rd}_{t_1}\mathfrak{A}$ and thus defines $\mathfrak{B}\subseteqq\mathbf{Rd}_{t_1}\mathfrak{A}$. Now, $\mathfrak{B}\in\mathbf{Sr}^C_{t_1}V$ since $\ker f\supseteqq C'$ and $\mathfrak{A}\in V$. Therefore if $\mathbf{Sr}^C_{t_1}V\models(\bigwedge_{i\leqq n} e_i\to e)$, then $\mathfrak{B}\models(\bigwedge_{i\leqq n} e_i\to e)$ then further $\langle\mathfrak{B}, f(r)\rangle_{r\in I}\models(\bigwedge_{i\leqq n}\xi(e_i)\to\xi(e))$ and therefore $\langle\mathfrak{A}, f(r)\rangle_{r\in I}\models(\bigwedge_{i\leqq n}\xi(e_i)\to\xi(e))$, which by the above implies $E\cup C'\cup$ $\cup\{\xi(e_i)\}_{i\leqq}\models_n\xi(e)$.

2. Now we assume that $E\cup C'\cup\{\xi(e_i)\}_{i\leqq n}\models\xi(e)$ for some $\xi\colon Y\rightarrowtail I$. We prove that this implies $\mathbf{Sr}^C_{t_1}V\models(\wedge e_i\to e)$.

Let $\mathfrak{B}\in\mathbf{Sr}^C_{t_1}V$ be arbitrary. Let $g\colon Y\to B$ be a valuation of the variables. We have to show $\mathfrak{B}\models(\bigwedge_{i\leqq n} e_i\to e)[g]$.

There exists a $g'\colon I\to B$ such that $g=\xi\circ g'$, since ξ is one-to-one. ($\xi\circ g$ denotes the composition of ξ and g.)

There is an $\mathfrak{A}\in V$ such that $\mathfrak{B}$ is a C-neat subreduct of $\mathfrak{A}$. Let f denote the unique homomorphic extension of $g'\colon I\to A$ to $\mathfrak{Fr}_{It}\to A$. Clearly $g=\xi\circ f$. Now,

f satisfies C' since $f(Fr_{It_1}) \subseteq B$ and $\mathfrak{B}$ is a C-neat subreduct of $\mathfrak{A}$. Now $E \cup C' \cup \cup \{\xi(e_i)\}_{i \leqq n} \models \xi(e)$ implies

$$\langle \mathfrak{A}, f(r) \rangle_{r \in I} \models (\bigwedge_{i \leqq n} \xi(e_i) \to \xi(e))$$

which by $g = \xi \circ f$ implies $\mathfrak{A} \models (\bigwedge_{i \leqq n} e_i \to e)[g]$ which is equivalent to $\mathfrak{B} \models (\bigwedge_{i \leqq n} e_i \to e)[g]$. Since $\mathfrak{B}$ and g were chosen arbitrarily, this implies $\mathbf{Sr}^C_{t_1} V \models (\bigwedge_{i \leqq n} e_i \to e)$. *QED*

By this we have proved Theorem 1, too.

The following corollary uses the notations of [8].

COROLLARY.

$$\mathbf{Cr}_I \mathbf{SNr}_\alpha \mathbf{CA}_{\alpha+\omega} = \mathbf{Cr}_I \mathbf{R}_\alpha = \mathbf{Cr}_I \mathbf{Lf}_\alpha = (\mathbf{Cr}_I^{I \times \{\alpha\}} \mathbf{CA}_{\alpha+\omega} \cap {}^2\mathbf{Fr}_{It}).$$

This was used in [1] to prove completeness of type-free logics, cf. C.4.4 and T.6.3 there.

A decidable axiomatization of the first order theory of the class $\mathbf{Lf}_\alpha$ was given in [12] by a single scheme of axioms.

PROBLEM. Find a "nice sufficient condition for $\mathbf{Sr}^C_{t_1} V$ to be a variety. (This is the case in Examples 1 and 2 but not in Example 4.) Compare the problem on p. 28 in [1]. About this problem: It was proved in [10], [11] that in the case of cylindric algebras the class $\mathbf{Nr}_\alpha \mathbf{CA}_\beta$ is not a variety, which solves Problem 2.11 of [8]. Hence we *cannot* replace neat *sub*reducts by neat reducts in our investigations.

Questions related to the subject of the present paper were investigated in BURRIS [4]. This subject is strongly related to "theory morphisms" in the sense of Lawvere, Burstall, and Goguen, which can be treated rather naturally in the frame of Cylindric Algebra Theory, see, e.g. [13] or [15].

REFERENCES

[1] ANDRÉKA, H.—GERGELY, T.—NÉMETI, I.: On universal algebraic construction of logics, *Studia Logica* **36** (1977), 9—47.

[2] ANDRÉKA, H.—NÉMETI, I.: On universal algebraic logic and cylindric algebras, *Bulletin of the Section of Logic* **7** No. 4, Wroclaw, 1978, pp. 152—158.

[3] ANDRÉKA, H., NÉMETI, I.: On systems of varieties definable by schemes of equations, *Algebra Universalis* **11** (1980), 105—116.

[4] BURRIS, S.: Remarks of reducts of varieties, *Proc. Coll. Univ. Alg. Esztergom.* (to appear).

[5] CHANG, C. C.—KEISLER, H. J.: *Model Theory,* North Holland, 1973.

[6] FREEMAN, J. B.: Algebraic Semantics for Modal Predicate Logic. *Zeitschr. f. Math. Log.* **22**, (1976), 523—552.

[7] HENKIN, L.—MONK, J. D.—TARSKI, A.: Cylindric Set Algebras and related structures I. Lecture Notes in Mathematics, Springer-Verlag, Berlin—Heidelberg—New York (to appear).

[8] HENKIN, L.—MONK, J. D.—TARSKI, A.: *Cylindric Algebras, Part I.,* North-Holland, 1971.

[9] MONK, J. D.: Nonfinitizability of classes of representable cylindric algebras, *J. Symbolic Logic* **34** N. 3 (1969), 331—343.

[10] NÉMETI, I.—ANDRÉKA, H.: Not all representable cylindric algebras are neat reducts, *Bulletin of the Section of Logic* **8** No. 3, Wroclaw, (1979), 145—147.

[11] NÉMETI, I.—ANDRÉKA, H.: The class of neat-reducts of cylindric algebras is not a variety but is closed w.r.t. HP, *Preprint, Math. Inst. Hung. Acad. Sci., February 1978, No. 14* (1979).

[12] NÉMETI, I.—ANDRÉKA, H.: Dimension complemented and locally finite dimensional cylindric algebras are elementarily equivalent, *Algebra Universalis* (to appear).

[13] NÉMETI, I.—SAIN, I.: Connections between Algebraic Logic and Initial Algebra Semantics of CF languages, Mathematical Logic in Computer Science (*Proc. Coll. Salgótarján 1978*), *Colloq. Math. Soc. J. Bolyai Vol.* 26, *North-Holland Publ. Co.*, Amsterdam—New York, 1981, 511—556.

[14] PRATT, V. R.: *Dynamic Algebras,* Preprint, MIT, 1979.

[15] SAIN, I.: Theories, Theory Morphisms, and Cylindric Algebras, *Preprint, Math. Inst. Hung. Acad. Sci.*

Mathematical Institute of the Hungarian Academy of Sciences,
Budapest, Reáltanoda u. 13—15, Hungary 1053

(Received August 24, 1978)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 53—78.

EXTREMUM PROBLEMS FOR THE MOTIONS OF A BILLIARD BALL III. THE MULTI-DIMENSIONAL CASE OF KÖNIG AND SZŰCS

by

I. J. SCHOENBERG

Contents

1. Introduction and main results

This is the third paper on the subject, but can be read independently of the first two ([3], [4]). Let

$$U_n\colon\ 0 \leqq x_\nu \leqq 1, \quad (\nu = 1, \dots, n) \tag{1.1}$$

be the unit cube in $\mathbf{R}^n$. Let (a_ν) be a point interior to U_n and

$$L_n^1\colon\ x_\nu = \lambda_\nu u + a_\nu, \qquad (\nu = 1, \dots, n;\ -\infty < u < \infty) \tag{1.2}$$

a rectilinear and uniform motion, where $u=t$ denotes the time. We interpret (1.2) as the motion of a billiard ball (b.b.); as we wish to reflect the b.b. in the usual way on striking the $2n$ facets $x_\nu=0$ or 1 of U_n, we use the function $\langle x\rangle$ defined by

$$\langle x\rangle = \begin{cases} x & \text{if } \ 0 \leqq x \leqq 1, \\ 2-x & \text{if } \ 1 \leqq x \leqq 2 \end{cases} \quad \text{and} \quad \langle x+2\rangle = \langle x\rangle \quad \text{for all } x. \tag{1.3}$$

We have used this function in [3] and [4] in a slightly different normalization. The reflected path of the b.b. within U_n may be described by the equations

$$\Pi_n^1\colon\ x_\nu = \langle \lambda_\nu u + a_\nu\rangle, \qquad (\nu = 1, \dots, n;\ -\infty < u < \infty). \tag{1.4}$$

[1] Sponsored by the United States Army under Contract No. DAAG29—75—C—0024.

A classical theorem of KRONECKER (see [2]), and its generalization (see [1]), show the following: If the n components (λ_ν) are arithmetically linearly independent, then the motion (1.4) is ergodic, i.e., the path Π_n^1 is dense in U_n. If $1 \leqq k \leqq n-1$, while the (λ_ν) admit precisely $n-k$ linearly independent linear homogeneous relations with integer coefficients, then the path Π_n^1 is contained in and is dense in a finite k-dimensional skew polytope Π_n^k. This was shown by KÖNIG and SZŰCS in [2] for $k=2$ and $n=3$.

This result shows that the b.b. motions generalize naturally as follows: Let

$$\lambda^i = (\lambda_1^i, \ldots, \lambda_n^i), \qquad (i = 1, \ldots, k) \quad (1 \leqq k \leqq n-1) \tag{1.5}$$

be k *linearly independent vectors.* We replace (1.2) by

$$L_n^k\colon\ x_\nu = \sum_1^k \lambda_\nu^i u_i + a_\nu, \qquad (\nu = 1, \ldots, n;\ -\infty < u_i < \infty), \tag{1.6}$$

which we interpret as a k-dimensional optical signal starting from the point (a_ν) inside U_n at the time $t=0$, and spreading uniformly within the k-flat L_n^k. As we now think of the $2n$ facets of U_n as mirrors, the reflected path of the signal is a finite or infinite k-dimensional skew polytope Π_n^k. The function $\langle x \rangle$ may again be used and shows that the reflected path is parametrically represented by the equations

$$\Pi_n^k\colon\ x_\nu = \Big\langle \sum_1^k \lambda_\nu^i u_i + a_\nu \Big\rangle, \qquad (\nu = 1, \ldots, n;\ -\infty < u_i < \infty). \tag{1.7}$$

In order to avoid degenerate lower-dimensional problems we shall assume that the original signal (1.6) is in a *general position.*

DEFINITION 1. We say that the signal (1.6) is in *general position* (G.P.), provided that

(1.8) the n by k matrix $\|\lambda_\nu^i\|$ has no vanishing minor of order k.

Equivalently: If $1 \leqq \nu_1 < \nu_2 < \ldots < \nu_k \leqq n$, then the k linear functions

$$x_{\nu_1}, x_{\nu_2}, \ldots, x_{\nu_k},$$

of (1.6), may assume arbitrarily prescribed values for appropriate u_i.

Let $0 < \varrho < \dfrac{1}{2}$, $x = (x_\nu)$, $c = \left(\dfrac{1}{2}, \dfrac{1}{2}, \ldots, \dfrac{1}{2}\right)$, and consider the cube

$$C_\varrho^n\colon\ \|x - c\|_\infty < \varrho, \tag{1.9}$$

where $\|x\|_\infty = \max\limits_\nu (|x_\nu|)$.

DEFINITION 2. We say that the path (1.7) is *ϱ-admissible,* and denote it by $\Pi_n^k(\varrho)$, provided that it is in G.P., and that Π_n^k never penetrates into the cube (1.9), hence that

$$\Pi_n^k \cap C_\varrho^n = \emptyset. \tag{1.10}$$

As an extreme opposite of the ergodic case, we study the following

PROBLEM 1. To determine, or to estimate, the quantity

$$\varrho_{k,n} = \text{supremum } \varrho, \tag{1.11}$$

the supremum being taken for all ϱ having ϱ-admissible path $\Pi^k_n(\varrho)$.

Our main result is an estimate.

THEOREM 1. *We have the inequality*

$$\varrho_{k,n} \geqq \frac{1}{2} - \frac{k}{2n}, \quad (1 \leqq k \leqq n-1). \tag{1.12}$$

In § 9 we establish Theorem 1 by constructing a path $\Pi^k_n(\varrho)$ for values of ϱ which are as close to $\frac{1}{2}-\frac{k}{2n}$ as we wish.

In [4] I have shown that the equality sign holds in (1.12) for the case when $k=1$. We can now do the same for the other extreme case when $k=n-1$.

THEOREM 2. *We have that*

$$\varrho_{n-1,n} = \frac{1}{2} - \frac{n-1}{2n} = \frac{1}{2n}, \qquad (n \geqq 2). \tag{1.13}$$

The simplest case when $n=3$, and therefore

$$\varrho_{2,3} = \frac{1}{6}, \tag{1.14}$$

leads to what I call *Kepler's tetrahedron.* J. KEPLER was the first to notice that four appropriate vertices of the cube U_3 are the vertices of a *regular* tetrahedron T. As any two facets of T intersect in a facet of U_3 forming equal angles with that facet, it should be clear that the surface of T carries a reflected signal Π^2_3. It carries, of course, many such, but let us single out one of them and denote it by $\tilde{\Pi}^2_3$. Actually, this signal $\tilde{\Pi}^2_3$ is readily found to be $\frac{1}{6}$-admissible, and it is essentially the only Π^2_3 which is $\frac{1}{6}$-admissible. This is an apparently new characteristic extremum property of Kepler's tetrahedron: *Any other signal Π^2_3 in general position, must penetrate into the cube C^3_ϱ, with $\varrho=\frac{1}{6}$.*

Theorem 2 allows us to generalize this extremum property of T: *There is an essentially unique signal $\tilde{\Pi}^{n-1}_n$ which is in general position and is $\frac{1}{2n}$-admissible. It is explicitly given by*

$$\tilde{\Pi}^{n-1}_n\colon \begin{array}{ll} x_\nu = \langle u_\nu \rangle, & (\nu = 1, \dots, n-1), \\ x_n = \left\langle u_1 + u_2 + \dots + u_{n-1} + \dfrac{n-1}{2} \right\rangle, & (-\infty < u_i < \infty). \end{array} \tag{1.15}$$

In our elementary paper [5] we considered the case of $n=2$, when the path of $\tilde{\Pi}_2^1$ is the square with vertices in the midpoints of U_2.

THEOREM 3. *We construct explicitly the signal* $\tilde{\Pi}_n^k\left(\frac{1}{2}-\frac{k}{2n}\right)$ *for the two cases*

$$(1.16) \qquad (k, n) = (2, 4) \quad \textit{and} \quad (k, n) = (2, 6).$$

Notice that $(k, n)=(2, 5)$ is missing: I could not do it.

In view of Theorems 1, 2 and 3, I wish to state

CONJECTURE 1. *The value of* (1.11) *is*

$$(1.17) \qquad \varrho_{k,n} = \frac{1}{2}-\frac{k}{2n}, \qquad (1 \leqq k \leqq n-1).$$

The remainder of this paper is in two parts and an Appendix. Part I deals with monochromos and n-chromos in $\mathbf{R}^k$ already used in [4] for $k=1$. We derive Theorems 1′, 2′, and 3′; in Part II it will be shown that these theorems are equivalent to the above Theorems 1, 2, and 3, respectively.

There are three outstanding problems that we leave unresolved:

1°. A proof of Conjecture 1.

2°. A general arithmetic-analytic construction of a signal

$$(1.18) \qquad \Pi_n^k\left(\frac{1}{2}-\frac{k}{2n}\right),$$

as done by Theorem 3 in two very special cases.

3°. To show that the number of signals (1.18) is finite, as shown in [4] for $k=1$.

Problems 1° and 2° are probably related and all three difficult.

In the Appendix (§ § 10 and 11) we study the same extremum problems, where the rectilinear reflected k-flats are replaced by k-dimensional Lissajous manifolds. Again, Theorems 1′, 2′, 3′ on n-chromos, allow us to derive immediately three Theorems $1^L, 2^L, 3^L$, concerning the new situation.

I. n-CHROMOS

2. Monochromes

We consider the function

$$(2.1) \qquad \{x\} = \min |x-m| \quad \text{for} \quad m \in \mathbf{Z},$$

which is related to the function (1.3), in fact $\{x\}=\langle 2x\rangle/2$. It seems tailor made for dealing with systems of parallel and equidistant planes in $\mathbf{R}^k=\{u=(u_1, \dots, u_k)\}$. For if $\sum_1^k \lambda^i u_i+a$ is a *non-constant* linear function of the variables u_i, then the equation

$$(2.2) \qquad \{\sum_1^k \lambda^i u_i+a\} = 0$$

represents such a system of planes, it being equivalent with the system of equations

$$\sum_1^k \lambda^i u_i + a = j \qquad (j \in \mathbf{Z}). \tag{2.3}$$

Let

$$0 < \delta < 1, \tag{2.4}$$

and let us replace (2.2) by the inequality

$$M^k(\delta)\colon \{\sum_1^k \lambda^i u_i + a\} \leqq \frac{\delta}{2}. \tag{2.5}$$

This represents a system of congruent, parallel, and equidistant *slabs of space*. We call the point-set $M^k(\delta)$ *a monochrome* (M.C.) *of* $\mathbf{R}^k$, because we like to think of its points as carrying a certain color γ. The most familiar case is $k=2$, when $M^2(\delta)$ assumes the aspect of an *awning*, of the kind used to provide shade to storefronts.

We shall refer to the planes (2.3) as the *central planes* of the monochrome (2.5) (*central lines* if $k=2$).

The distance between two consecutive central planes (2.3) is found to be $p=1/\sum (\lambda^i)^2$, while the width of a slab of (2.5) is seen to be $w=\delta/\sum (\lambda^i)^2$. Therefore

$$\delta = \frac{w}{p},$$

and for this reason we call δ *the density of the monochrome* $M^k(\delta)$. Clearly δ represents the *density of the color* γ in the space $\mathbf{R}^k$ containing $M^k(\delta)$.

3. *n*-chromos

Let

$$n > k, \tag{3.1}$$

and let us have in $\mathbf{R}^k$ n monochromes

$$M_1^k(\delta), M_2^k(\delta), \dots, M_n^k(\delta), \tag{3.2}$$

all of the same density δ. To make matters more picturesque, we think of $M_\nu^k(\delta)$ as carrying the color γ_ν.

Definition 3. We say that the n monochromes (3.2) define an n-chromo $\chi_n^k(\delta)$, provided that

$$\bigcup_{\nu=1}^n M_\nu^k(\delta) = \mathbf{R}^k. \tag{3.3}$$

The characteristic property of an n-chromo is therefore that every point (u_i) of $\mathbf{R}^k$ is covered by one or more of the colors γ_ν. Using (2.5) we may represent our monochromes by

$$M_\nu^k(\delta)\colon \{\sum_1^k \lambda_\nu^i u_i + a_\nu\} \leqq \frac{\delta}{2}, \qquad (\nu = 1, \dots, n). \tag{3.4}$$

DEFINITION 4. We say that the n-chromo $\chi_n^k(\delta)$ is admissible, provided that the set of n vectors

$$\vec{\lambda}_\nu = (\lambda_\nu^1, \lambda_\nu^2, \dots, \lambda_\nu^k), \qquad (\nu = 1, \dots, n), \tag{3.5}$$

which are the normal vectors of our monochromes, have the following property: Every subset of k vectors $\vec{\lambda}_{\nu_1}, \vec{\lambda}_{\nu_2}, \dots, \vec{\lambda}_{\nu_k}$ $(\nu_1 < \dots < \nu_k)$, spans the space $\mathbf{R}^k$.

Equivalently: All $\binom{n}{k}$ kth order minors of the matrix

$$\Lambda = \|\lambda_\nu^i\| \tag{3.6}$$

are different from zero.[2]

The following lemma seems evident and requires no proof.

LEMMA 1. *A non-singular affine transformation of* $\mathbf{R}^k$ *into itself maps monochromes and n-chromos into like objects of the same density.*

4. An extremum problem for admissible n-chromos

Let

$$\chi_n^k(\delta) = \{M_1^k(\delta), M_2^k(\delta), \dots, M_n^k(\delta)\} \tag{4.1}$$

denote the n-chromo defined by (3.4). If we keep everything fixed in (3.4), except that we replace the density δ by $\delta' > \delta$, then it is clear that $\chi_n^k(\delta')$ is a fortiori an n-chromo. This is no longer true if we try to diminish the density δ. In fact, keeping only k and n fixed, it will be our main concern to find an admissible n-chromo $\chi_n^k(\delta)$ having as small a density δ as possible. Evidently, δ cannot be too small. It is trivial that we must have

$$\delta \geqq \frac{1}{n}, \tag{4.2}$$

for if $\delta < \frac{1}{n}$, then our monochromos (3.2) are clearly unable to cover $\mathbf{R}^k$, as required by (3.3): There just isn't enough paint around!

As mentioned above we are interested in

PROBLEM 1'. To determine, or to estimate, the quantity

$$\delta_{k,n} = \text{infimum } \delta \tag{4.3}$$

for all densities δ of admissible n-chromos $\chi_n^k(\delta)$.

The main result of Part I is

[2] To see examples of n-chromos in $\mathbf{R}^2$, the reader is invited to inspect the 5-chromo $\chi_5^2(2/5)$ of Figure 1 (§5), and the 4-chromo $\chi_4^2(1/2)$ of Figure 2 (§7). The first is *not* admissible, because its monochromes M_3, M_4, and M_5, are parallel; the second is *admissible*, since no two of its monochromes are parallel.

THEOREM 1'. *We have the inequality*

$$\delta_{k,n} \leqq \frac{k}{n}, \qquad (1 \leqq k \leqq n-1). \tag{4.4}$$

REMARK. The result (4.4) is rather trivial if $k=1$, in fact

$$\delta_{1,n} = \frac{1}{n}. \tag{4.5}$$

PROOF of (4.5). By (4.2) it suffices to exhibit an admissible $\chi_n^1\left(\frac{1}{n}\right)$ of density $\frac{1}{n}$. Observe first that the requirement that χ_n^1 be admissible drops out because it is automatically fulfilled for $k=1$. The relations (3.4) reduce to

$$M_\nu^1(\delta)\colon \{\lambda_\nu^1 u_1 + a_\nu\} \leqq \frac{\delta}{2}, \qquad (\nu = 1, \ldots, n), \tag{4.6}$$

where we implicitly assume that $\lambda_\nu^1 \neq 0$ for all ν, or else we could not speak of monochromes. *The matrix* (3.6) *reduces to a column of non-vanishing elements.* Secondly, it is clear that the monochromes of $\mathbf{R}^1$ of density $\frac{1}{n}$

$$M_\nu^1\left(\frac{1}{n}\right)\colon \left\{u_1 + \frac{\nu-1}{n}\right\} \leqq \frac{1}{2n}, \qquad (\nu = 1, \ldots, n) \tag{4.7}$$

do not overlap and cover the real axis $\mathbf{R}^1$. Therefore $\delta_{1,n} \leqq \frac{1}{n}$ and this established (4.5).

5. A proof of Theorem 1'

We shall proceed as follows: We shall exhibit an admissible $\chi_n^k(\delta)$ having a density which is as close to $\frac{k}{n}$ as we wish, thereby establishing the inequality (4.4). This is done in two stages.

A. *Construction of a certain non-admissible* $\chi_n^k(\delta)$ *of density* $\delta = \frac{k}{n}$.
Let

$$\delta = \frac{k}{n}, \quad q = n-k. \tag{5.1}$$

We use the freedom afforded by Lemma 1 and may, without loss of generality, assume the central planes of the first k monochromes to be the planes $u_\nu - \frac{1}{2} = j$, hence

$$M_\nu^k(\delta)\colon \left\{u_\nu - \frac{1}{2}\right\} \leqq \frac{\delta}{2}, \qquad (\nu = 1, \ldots, k). \tag{5.2}$$

In Figure 1 we exhibit the case $k=2$ and $n=5$ of our construction, but the same construction holds for any k and $n(>k)$.

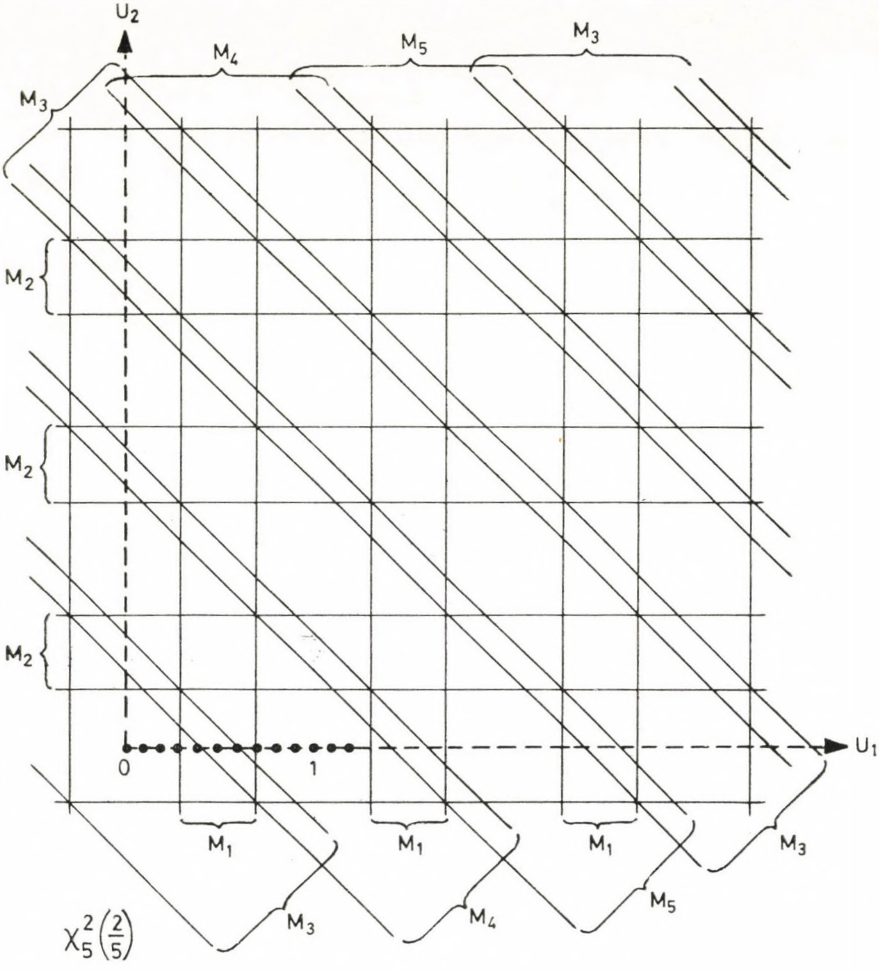


Fig. 1

Notice that monochromes (5.2) already cover all of $\mathbf{R}^k$, with the exception of the lattice of cubes having sides $= 1-\delta$

$$C(m_1, \ldots, m_k)\colon\ \|x-\vec{m}\|_\infty < \frac{1-\delta}{2}, \quad \vec{m} = (m_i) \in \mathbf{Z}^k. \tag{5.3}$$

The remaining $q=m-k$ monochromes are to cover all those cubes.

Claim. *The monochrome*

$$M^k_{k+1}(\delta)\colon \left\{\frac{\sum\limits_1^k u_i}{q}\right\} \leqq \frac{\delta}{2} \tag{5.4}$$

just covers all cubes

$$C(m_1, \dots, m_k) \quad \textit{such that} \quad \sum_1^k m_i \equiv 0 \pmod{q}. \tag{5.5}$$

PROOF of claim. We look at the cube $C(0, \dots, 0)$ and let

$$A = \left(-\frac{1-\delta}{2}, \dots, -\frac{1-\delta}{2}\right), \quad B = \left(\frac{1-\delta}{2}, \dots, \frac{1-\delta}{2}\right) \tag{5.6}$$

be its vertices such that $\overrightarrow{AB}$ has direction numbers $(1, 1, \dots, 1)$. The slab of (5.4) containing the origin is defined by

$$-\frac{\delta}{2} \leqq \frac{\sum_1^k u_i}{q} \leqq \frac{\delta}{2}. \tag{5.7}$$

Notice that the right bounding plane $\sum u_i/q = \delta/2$ contains the point B, because at B we have by (5.6)

$$\frac{\sum u_i}{q} = \frac{k}{q}\,\frac{1-\delta}{2} = \frac{k}{q}\,\frac{1-\frac{k}{n}}{2} = \frac{k}{2n}\,\frac{n-k}{q} = \frac{k}{2n} = \frac{\delta}{2}$$

by (5.1). Similarly, the left bounding plane $-(\delta/2) = (\sum u_i)/q$ passes through A. The normal to the monochrome (5.4) being the vector $(1, \dots, 1)$, it is clear that (5.4) contains the set

$$\bigcup_{\sum m_i = 0} C(m_1, \dots, m_k).$$

However, the central planes of (5.4) are

$$\frac{\sum_1^k u_i}{q} = j \quad (j \in \mathbf{Z}),$$

and these pass through the centers of all cubes $C(m_1, \dots, m_n)$ such that $\sum_1^k m_i = qj$. This proves our claim.

By parallel translation we now define

$$M_{k+r}^k(\delta) \colon \left\{\frac{\sum_1^k u_i + r - 1}{q}\right\} \leqq \frac{\delta}{2}, \qquad (r = 1, 2, \dots, q), \tag{5.8}$$

and this covers all cubes

$$C(m_1, \dots, m_k) \quad \text{such that} \quad \sum_1^k m_i \equiv -r+1 \pmod{q},$$

because the central planes of (5.8) pass through their centers. The n-chromo

$$\chi_n^k(\delta) = \{M_1^k(\delta), \dots, M_n^k(\delta)\} \tag{5.9}$$

defined by (5.2) and (5.8) is the inadmissible n-chromo of density $\delta=k/n$ we wished to construct. (5.9) is *not* admissible because its last $n-k=q$ monochromes are pairwise parallel.

B. *Construction of an admissible n-chromo of density δ close to k/n.*

This will be achieved by an appropriate slight perturbation of (5.9). We start by selecting a fixed matrix

$$(5.10) \qquad A=\|a_{ri}\| \qquad (r=1,\dots,q;\ i=1,\dots,k)$$

having the following properties:

(5.11) *The elements a_{ri} are integers,*

(5.12) *All minors of A, hence of orders from* 1 *to* $\min(q,k)$ *are* $\neq 0$.

From the known *total positivity* properties of the binomial coefficients, both conditions are verified if we select

$$(5.13) \qquad a_{vi}=\binom{k+v}{i}.$$

We are now going to modify the n-chromo (5.9) as follows. We will select for it a density $\tilde{\delta}$ to be determined later. We replace the first monochromes (5.2) by

$$(5.14) \qquad M_v^k(\tilde{\delta}):\ \left\{u_v-\frac{1}{2}\right\}\leqq\frac{\tilde{\delta}}{2} \qquad (v=1,\dots,k).$$

For the last $n-k=q$ monochromes we prescribe their central planes to be

$$(5.15) \qquad \pi_{r,j}:\ N\frac{\sum_1^k u_i+(r-1)}{q}+a_{r1}u_1+a_{r2}u_2+\dots+a_{rk}u_k=j,\ (j\in\mathbf{Z}),\ (r=1,\dots,q).$$

Here N is a positive integer to be made large later. We claim that *every lattice point $(m_i)\in\mathbf{Z}^k$ is in one of these planes $\pi_{r,j}$.*

For if $(m_i)\in\mathbf{Z}^k$ is given, we determine the unique r such that

$$(5.16) \qquad r\equiv-\sum_1^k m_i+1 \pmod q$$

and then

$$(5.17) \qquad (m_i)\in\pi_{r,j'} \quad \text{for some} \quad j'\in\mathbf{Z}.$$

Thus

$$(5.18) \qquad \mathbf{Z}^k\subset\bigcup_{r=1}^{q}\ \bigcup_{j=-\infty}^{\infty}\pi_{r,j}.$$

Let us now look at the geometric aspect of the planes $\pi_{r,j}$. (5.15) may be written as

$$(5.19) \qquad \pi_{r,j}:\ \sum_1^k u_i+\frac{q}{N}(a_{r1}u_1+a_{r2}u_2+\dots+a_{rk}u_k)=\frac{q}{N}j-r+1,$$

and this shows that

(5.20) *all the planes* $\pi_{r,j}$ *are nearly parallel to the plane* $\sum_1^k u_i=0$, *provided that N is sufficiently large.*

Let $\overrightarrow{A(m_i)B(m_i)}$ be the diagonal of the cube $C(m_i)$, which is parallel to the old diagonal $\overrightarrow{AB}$ of $C(0, \dots, 0)$. Let r be fixed such that (5.17) holds. We construct a monochrome $M_{k+r}^k(\delta_r)$, parallel to $\pi_{r,j}$, which *just covers* the cube $C(m_i)$. It is obtained by bounding its slab of color (containing $C(m_i)$) by the two planes parallel to $\pi_{r,j}$, and passing through the points $A(m_i)$ and $B(m_i)$, respectively. This monochrome will also cover all cubes $C(m_i)$ such that (5.16) holds, or

$$\sum_1^k m_i \equiv -r+1 \pmod q. \tag{5.21}$$

We may write

$$M_{k+r}^k(\delta_r)\colon \left\{N\,\frac{\sum_1^k u_i+r-1}{q}+\sum_1^k a_{ri}u_i\right\} \leqq \frac{\delta_r}{2}. \tag{5.22}$$

In view of (5.19) we conclude that its density δ_r will be as close as we wish to the old density k/n of (5.9).

For the final selection of our monochrome $\tilde{M}_\nu^k$, we keep the inequalities (5.14) and (5.22), only modifying the density, by selecting for both groups the common density $\tilde\delta$ defined by

$$\tilde\delta = \max\left(\frac{k}{n}, \delta_1, \delta_2, \dots, \delta_q\right). \tag{5.23}$$

Thus $\tilde\delta \geqq \frac{k}{n}$. If $\tilde\delta > \frac{k}{n}$, then (5.14) shows that our old cubes $C(m_i)$ have shrunk, and are therefore a fortiori covered by the $M_{k+r}^k(\tilde\delta)$.

Since $\tilde\delta \to \frac{k}{n}$ as $N\to\infty$, the n-chromo

$$\tilde\chi_n^k(\tilde\delta) = \{\tilde M_1(\tilde\delta), \dots, \tilde M_n(\tilde\delta)\} \tag{5.24}$$

will have a density $\tilde\delta$ as close to $\frac{k}{n}$, provided that we select N sufficiently large.

The question: *Is the n-chromo* (5.24) *admissible?* By (5.14) and (5.19) we see that the matrix (3.6) for its central planes is

$$\Lambda = \|\lambda_\nu^i\| = \left\|\begin{matrix} 1 & 0 & & 0 \\ 0 & 1 & & \\ 0 & & \ddots & 1 \\ 1+\frac{q}{N}a_{11}, & \dots, & & 1+\frac{q}{N}a_{1k} \\ \vdots & & & \\ 1+\frac{q}{N}a_{q1}, & \dots, & & 1+\frac{q}{N}a_{qk} \end{matrix}\right\| = \left\|\begin{matrix} I_k \\ 1+\frac{q}{N}a_{ri} \end{matrix}\right\|. \tag{5.25}$$

We claim that *all its* kth *order minors are* $\neq 0$ *if N is sufficiently large.*

This will be the case if and only if

(5.26) *for sufficiently large N the matrix* $\left\| 1+\frac{q}{N}a_{ri}\right\|$ *has no vanishing minor of any order from* 1 *to* $\min(q, k)$.

To verify this statement let us look at an sth order minor of the matrix of (5.26). We inspect the leading minor $\det\left|1+\frac{q}{N}a_{ri}\right|$ for $r=1, \ldots, s$, $i=1, \ldots, s$. [3] Splitting each of its columns into two columns, we find

$$\det\left|1+\frac{q}{N}a_{ri}\right|_{1,s}=\left(\frac{q}{N}\right)^{s}\det|a_{ri}|_{1,s}+\left(\frac{q}{N}\right)^{s-1}\cdot S, \tag{5.27}$$

where S is the sum of s determinants obtained from $\det|a_{ri}|_{1,s}$ by replacing each of its columns successively by a column of 1's. We distinguish two cases:

1. If $S\neq 0$, then the right-hand side of (5.27) will surely be $\neq 0$ if N is sufficiently large.

2. If $S=0$, we reach the same conclusion in view of the property (5.12) which implies that $\det|a_{ri}|_{1,s}\neq 0$. We have shown that the n-chromo (5.24) is admissible, which completes our proof of Theorem 1′.

6. Solution of Problem 1′ if $k=n-1$

Among the n-chromos (5.9) for $k=2, 3, \ldots, n-1$ we single out the case

$$k=n-1, \tag{6.1}$$

this being the only one which is *admissible*. Its density is

$$\delta=\frac{n-1}{n}. \tag{6.2}$$

By (5.2) and (5.8), its monochromes are described by

$$M_\nu\left(\frac{n-1}{n}\right):\left\{u_\nu-\frac{1}{2}\right\}\leqq\frac{n-1}{2n}\qquad(\nu=1, \ldots, n-1) \tag{6.3}$$

and

$$M_n\left(\frac{n-1}{n}\right):\left\{\sum_1^k u_i\right\}\leqq\frac{n-1}{2n}, \tag{6.4}$$

since $q=1$.

We wish to prove the

THEOREM 2′. *We have that*

$$\delta_{n-1,n}=\frac{n-1}{n},\quad(n\geqq 2). \tag{6.5}$$

[3] The same reasoning will apply to any other minor.

PROOF. We know from § 5 that the monochromes (6.3) cover all of R^{n-1} with the exception of the lattice of cubes

$$C(m_1, \dots, m_{n-1}),\ (m_1, \dots, m_{n-1}) \in \mathbf{Z}^{n-1}, \tag{6.6}$$

centered at the lattice points and having $\text{sides} = 2 \cdot \frac{1-\delta}{2} = 1-\delta = 1 - \frac{n-1}{n} = \frac{1}{n}$. We also know that the last monochrome (6.4) just covers all these cubes.

For convenience we say that a monochrome $M^{n-1}(\delta')$ of $\mathbf{R}^{n-1}$ is *slanting*, provided that all $n-1$ components of its normal vector are positive.

LEMMA 2. *If the slanting monochrome*

$$M^{n-1}(\delta'): \{u_1+\gamma_2 u_2+\dots+\gamma_{n-1}u_{n-1}+b\} \leqq \frac{\delta'}{2}, \tag{6.7}$$

where

$$\gamma_2 > 0, \dots, \gamma_{n-1} > 0, \tag{6.8}$$

covers the set

$$\Gamma = \bigcup_{(m_\nu)\in \mathbf{Z}^{n-1}} C(m_1, \dots, m_{n-1}), \tag{6.9}$$

then we must have that

$$\gamma_2 = \gamma_3 = \dots = \gamma_{n-1} = 1. \tag{6.10}$$

PROOF of Lemma 2. Let S denote the set of planes π which are parallel to (6.7) and intersect the set Γ, hence

$$S = \Big\{\pi\colon u_1 + \sum_2^{n-1} \gamma_i u_i = \text{const.};\ \pi \cap \Gamma \neq \emptyset\Big\}. \tag{6.11}$$

Crucial in our discussion is the nature of the set

$$\Omega = S \cap \mathbf{R}^1, \tag{6.12}$$

where

$$\mathbf{R}^1 = \{(u_i);\ u_2 = \dots = u_{n-1} = 0\} \tag{6.13}$$

is the u_1-axis.

CLAIM. *If*

$$(\gamma_2, \gamma_3, \dots, \gamma_{n-1}) \neq (1, 1, \dots, 1), \tag{6.14}$$

then

$$\Omega = \mathbf{R}^1. \tag{6.15}$$

PROOF of claim. The set S is the union of those planes π which intersect the individual cubes $C(m_1, \dots, m_{n-1})$. At this point it is more convenient to shift the origin of $\mathbf{R}^{n-1}$ to the "lower left-hand corner" of the cube $C(0, \dots, 0)$. This is the point A of (5.6), for $k=n-1$. Since $1-\delta=\frac{1}{n}$, we see that after this shift of origin

$$C(m_1, \dots, m_{n-1}) = \Big\{(u_\nu)\colon m_i \leqq u_i \leqq m_i + \frac{1}{n}, \quad (i = 1, \dots, n-1)\Big\}. \tag{6.16}$$

Let us project this cube onto $\mathbf{R}^1$ by planes parallel to our monochrome. The two extreme planes are two planes of support of C and their equations are

$$(u_1-m_1)+\sum_2^{n-1}\gamma_i(u_i-m_i)=0 \quad\text{and}\quad \left(u_1-m_1-\frac{1}{n}\right)+\sum_2^{n-1}\gamma_i\left(u_i-m_i-\frac{1}{n}\right)=0,$$

respectively. To intersect them with $\mathbf{R}^1$, we set $u_2=\ldots=u_{n-1}=0$ in these equations and solve them for u_1. In this way we find that the cube (6.16) is projected into the interval

$$I(m_1,\ldots,m_{n-1})=\left[m_1+\sum_2^{n-1}\gamma_i m_i,\; m_1+\sum_2^{n-1}\gamma_i m_i+\frac{1}{n}(1+\gamma_2+\ldots+\gamma_{n-1})\right]. \tag{6.17}$$

For the set (6.12) we now find that

$$\Omega=\cup\, I(m_1,\ldots,m_{n-1}) \quad\text{for}\quad (m_1,\ldots,m_{n-1})\in\mathbf{Z}^{n-1}. \tag{6.18}$$

We distinguish two cases.

1. *Among the γ_i there is an irrational one, γ_2 say.* Setting $m_i=0$ for $i>2$, we find the lower endpoint of $I(m_1,m_2,0,\ldots,0)$ to be

$$m_1+\gamma_2 m_2 \qquad (\gamma_2 \text{ is irrational}) \tag{6.19}$$

and Kronecker's theorem shows that these lower endpoints are dense in $\mathbf{R}^1$. By (6.18) our conclusion (6.15) clearly follows.

2. *All γ_i are rational.* Writing them in simplest terms with a common denominator we have

$$\gamma_\nu=\frac{a_\nu}{b}, \qquad (\nu=2,\ldots,n-1), \quad (b,a_2,\ldots,a_{n-1})=1. \tag{6.20}$$

As our assumption (6.14) excludes the case when $b=a_2=\ldots=a_{n-1}=1$, we have $b+a_2+\ldots+a_{n-1}\geqq n$ and therefore

$$\frac{1}{n}(1+\gamma_2+\ldots+\gamma_{n-1})\geqq\frac{1}{b}. \tag{6.21}$$

However, the lower endpoints of the intervals (6.17) form the arithmetic progression j/b ($j\in\mathbf{Z}$). Since (6.21) shows that the common length of our intervals (6.17) is $\geqq 1/b$, again we have by (6.18) that (6.15) holds.

Completing a proof of Theorem 2'. By Lemma 2 we learn that a monochrome (6.7) covering the set (6.9), must be of the form

$$M^{n-1}(\delta')\colon \left\{\sum_1^{n-1}u_i+b\right\}\leqq\frac{\delta'}{2}. \tag{6.22}$$

As this must also cover (6.4), we conclude that $\delta'\geqq\dfrac{n-1}{n}$. This establishes Theorem 2': For if we *diminish* the common density of the (6.3), then this would *increase* the common side of the cubes (6.6), and then these could only be covered by a slanting monochrome of density $>\frac{n-1}{n}$, as we have seen.

In view of Theorems 1′, 2′, and the examples of Theorem 3′, I wish to state

CONJECTURE 1′. *The value of* (4.3) *is*

$$\delta_{k,n}=\frac{k}{n} \qquad (1\leqq k\leqq n-1). \tag{6.23}$$

REMARK. Just a comment on the monochromes (4.7) of $\mathbf{R}^1$, of density $\delta=\frac{1}{n}$. Clearly, the inequalities

$$M_\nu^k\left(\frac{1}{n}\right)\colon \left\{u_1+\frac{\nu-1}{n}\right\}\leqq\frac{1}{2n}, \qquad (\nu=1,\ldots,n)$$

also define an n-chromo $\chi_n^k\left(\frac{1}{n}\right)$ in $\mathbf{R}^k$, having the density $\frac{1}{n}$. This does not contradict the above conjectured relation (6.23): The quantity $\delta_{k,m}$ was defined as the infimum of δ for n-chromos $\chi_n^k(\delta)$ in $\mathbf{R}^k$, which are *admissible,* while the above n-chromos $\chi_n^k\left(\frac{1}{n}\right)$ is far from satisfying that essential requirement. In fact all of its n monochromes are parallel.

7. Two special explicit n-chromos

Theorem 1′ was *not* established by exhibiting an n-chromo

$$\chi_n^k\left(\frac{k}{n}\right) \tag{7.1}$$

which is both admissible and of density k/n. Rather in § 5 we construct *admissible* $\chi_n^k(\delta)$, *with* δ *as close to* k/n *as we wished.* In view of our Conjecture 1′ of § 6, the construction of an admissible n-chromo (7.1), for prescribed k and n $(k<n)$, is a most desirable but as yet unsolved problem. Even for low values of k and n, the success depends, so far, on luck and visual inspection. Needed is a general arithmetic-analytic construction.

As a guide to the nature of this problem, the following two specific examples might be useful.

THEOREM 3′. *We give explicit constructions of the n-chromo* (7.1) *for the following two cases*

$$(k,n)=(2,4) \textit{ and } (k,n)=(2,6). \tag{7.2}$$

1. $\mathbf{k=2, n=4}$. Here the density is

$$\delta=\frac{1}{2}. \tag{7.3}$$

The four monochromes of $\chi_4^2\left(\frac{1}{2}\right)$ are

$$(7.4)\qquad \begin{aligned} &M_1^2\left(\frac{1}{2}\right):\left\{u_1-\frac{1}{2}\right\}\leqq\frac{1}{4}, \quad && M_2^2\left(\frac{1}{2}\right):\left\{u_2-\frac{1}{2}\right\}\leqq\frac{1}{4},\\ &M_3^2\left(\frac{1}{2}\right):\left\{\frac{u_1+u_2}{2}\right\}\leqq\frac{1}{4}, && M_4^2\left(\frac{1}{2}\right):\left\{\frac{u_1-u_2+1}{2}\right\}\leqq\frac{1}{4}.\end{aligned}$$

These are easily derived from Figure 2 which shows that we have an admissible 4-chromo of $\mathbf{R}^2$.

The first two monochromes (7.4) cover the plane with the exception of the lattice of squares $C(m_1, m_2)$ having sides $=1/2$. The third monochromo M_3 covers all those squares such that m_1+m_2 is even, while M_4 covers those with an odd sum m_1+m_2.

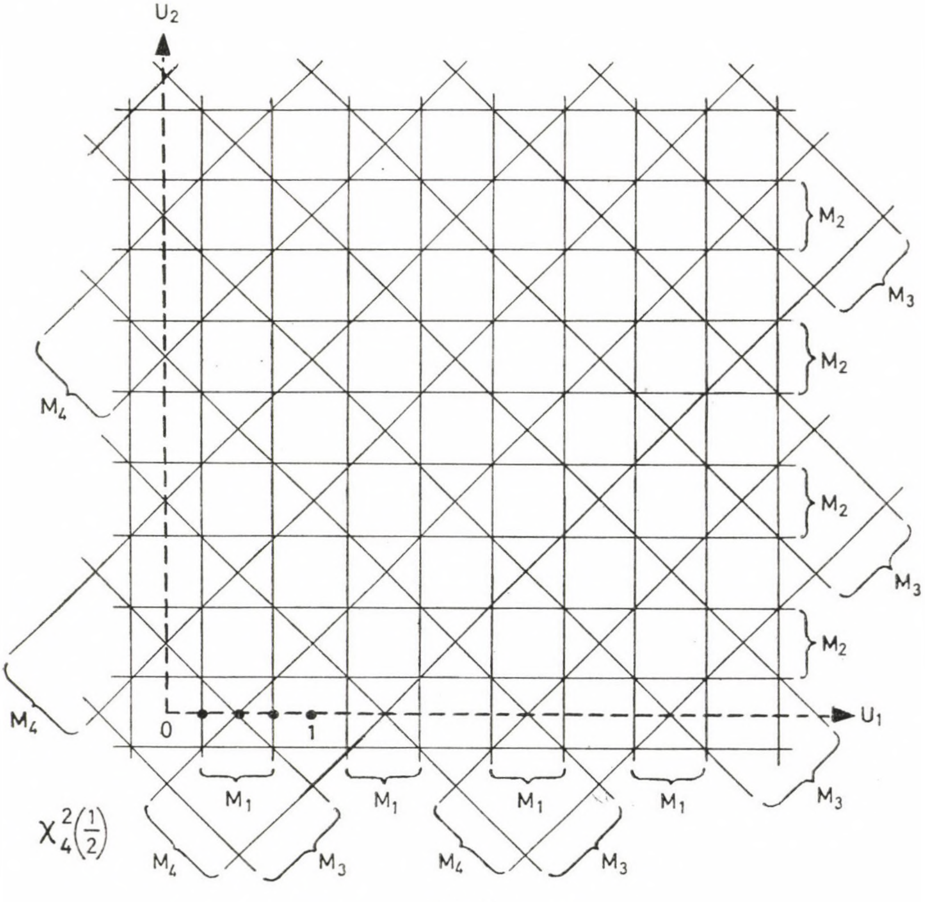


Fig. 2

2. **k**=**2**, **n**=**6**. Now

$$\delta = \frac{1}{3}. \tag{7.5}$$

The six monochromes of $\chi_6^2\left(\frac{1}{3}\right)$ are

$$M_1^2\left(\frac{1}{3}\right): \left\{u_1-\frac{1}{2}\right\} \leqq \frac{1}{6}, \qquad M_2^2\left(\frac{1}{3}\right): \left\{u_2-\frac{1}{2}\right\} \leqq \frac{1}{6}, \tag{7.6}$$

$$M_3^2\left(\frac{1}{3}\right): \{u_1+u_2\} \leqq \frac{1}{6}, \qquad M_4^2\left(\frac{1}{3}\right): \left\{\frac{u_1-u_2}{3}\right\} \leqq \frac{1}{6}, \tag{7.7}$$

$$M_5^2\left(\frac{1}{3}\right): \left\{\frac{2u_1+u_2}{3}+\frac{1}{2}\right\} \leqq \frac{1}{6}, \quad M_6^2\left(\frac{1}{3}\right): \left\{\frac{u_1+2u_2}{3}+\frac{1}{2}\right\} \leqq \frac{1}{6}. \tag{7.8}$$

These are easily derived from Figure 3 which shows that we have an admissible 6-chromo of $\mathbf{R}^2$.

A guiding word in this maze of lines seems appropriate. The two monochromes (7.6) cover $\mathbf{R}^2$, except for the lattice of squares $C(m_1, m_2)$ having sides $=2/3$. The monochrome M_3, having central lines $u_1+u_2=j$ ($j\in\mathbf{Z}$), is seen to slice each of the squares into two congruent isosceles triangles; we denote the lower one by $T_1(m_1, m_2)$ and the upper one by $T_2(m_1, m_2)$. The monochrome M_4, having central lines $u_1-u_2=3j$ ($j\in\mathbf{Z}$), is seen to cover all pairs

$$T_1(m_1, m_2), \quad T_2(m_1, m_2)$$

such that $m_1-m_2\equiv 0 \pmod 3$. The last two monochromos M_5 and M_6 are to cover all remaining triangles.

At this point we observe, by (7.6), (7.7), (7.8), that each of the six M_ν admits a (double) periodicity of period 3 in each of the variables u_1 and u_2. It follows that it suffices to inspect our Figure 3 only in the square

$$S: -\frac{1}{2} \leqq u_1 < 2+\frac{1}{2}, \quad -\frac{1}{2} \leqq u_2 < 2+\frac{1}{2},$$

which in Figure 3 is indicated by a solid frame. In that square we are only left with the following triangles as yet uncovered:

$$T_1(1, 0),\ T_2(1, 0),\ T_1(2, 1),\ T_2(2, 1),\ T_1(2, 0),\ T_2(2, 0),$$

and the symmetric set

$$T_1(0, 1),\ T_2(0, 1),\ T_1(1, 2),\ T_2(1, 2),\ T_1(0, 2),\ T_2(0, 2).$$

However, M_5 covers

$$T_1(1, 0),\ T_2(0, 1),\ T_1(0, 2) \quad\text{and}\quad T_2(2, 0),\ T_1(2, 1),\ T_2(1, 2),$$

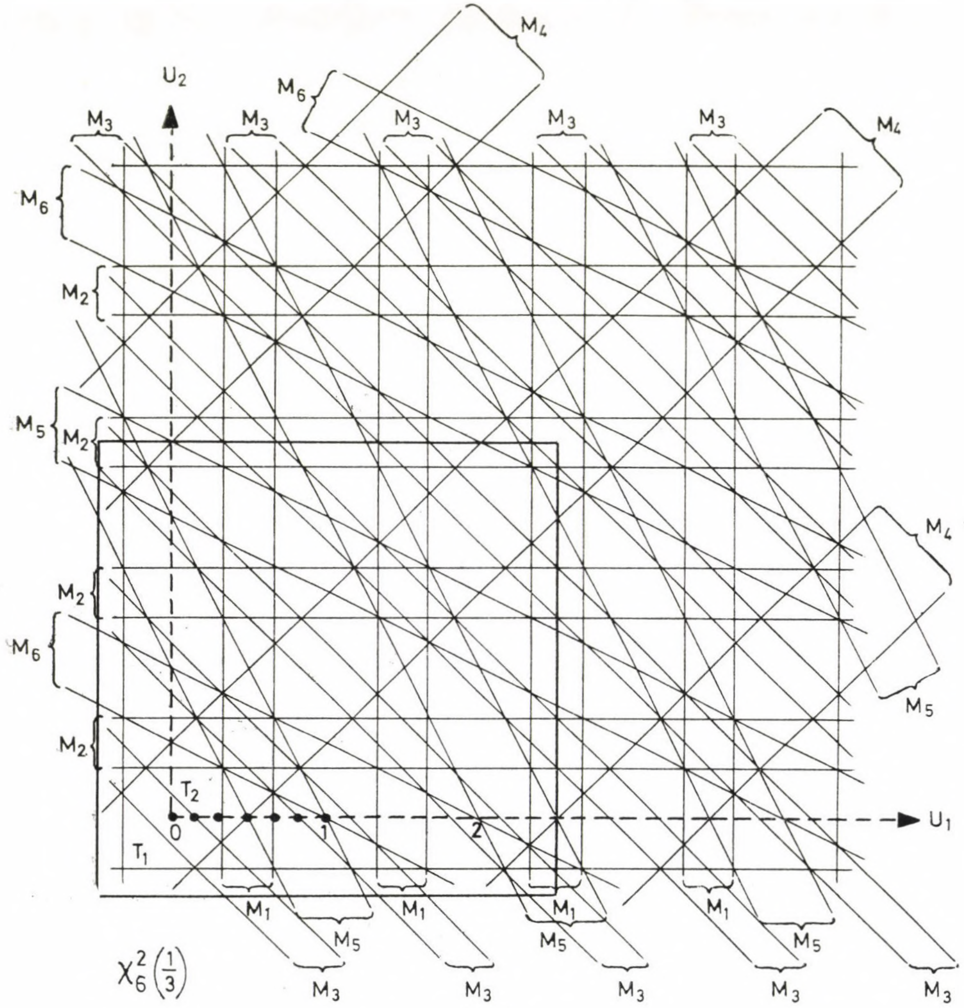


Fig. 3

while M_6 covers their symmetric images

$$T_1(0, 1),\ T_2(1, 0),\ T_1(2, 0) \quad \text{and} \quad T_2(0, 2),\ T_1(1, 2),\ T_2(2, 1).$$

This proves that we have a 6-chromo; it is admissible because no two monochromes are parallel.

Let me add that I could not discover a $\chi_5^2(2/5)$.

II. APPLICATIONS OF n-CHROMOS TO BILLIARD BALL MOTIONS

8. The equivalence of Problems 1 and 1′

This equivalence appears immediately as soon as we switch the problems (or "action") from the space $\mathbf{R}^n=\{(x_\nu)\}$ to the lower dimensional space $\mathbf{R}^k=\{(u_i)\}$. Indeed, let

$$\Pi_n^k\colon x_\nu=\left\langle\sum_{i=1}^k \lambda_\nu^i u_i+a_\nu\right\rangle, \qquad (\nu=1, \ldots, n), \tag{8.1}$$

be a ϱ-admissible reflected signal. Let ϱ and δ be related by

$$\varrho=\frac{1}{2}-\frac{\delta}{2} \quad \text{or} \quad \delta=1-2\varrho. \tag{8.2}$$

That (8.1) is ϱ-admissible means that it is contained in the cubical shell

$$B_\varrho^n=U_n-C_\varrho^n, \tag{8.3}$$

having the width $\frac{1}{2}-\varrho=\delta/2$. The structure of the function $\langle x\rangle$ implies the following: The point of $\mathbf{R}^n$

$$\left(\sum_1^k \lambda_\nu^i u_i+a_\nu\right) \qquad (\nu=1, \ldots, n)$$

has the property that for every (u_ν) and for some ν the number $\sum \lambda_\nu^i u_i+a_\nu$ differs from an integer by $\leqq\delta/2$. However, this last property can be expressed thus:

$$\text{(8.4)} \qquad \textit{For every } (u_\nu) \textit{ and for some } \nu \textit{ we have } \left\{\sum \lambda_\nu^i u_i+a_\nu\right\} \leqq \frac{\delta}{2}.$$

In terms of the monochromes

$$M_\nu^k(\delta)\colon \left\{\sum_1^k \lambda_\nu^i u_i+a_\nu\right\} \leqq \frac{\delta}{2}, \quad (\nu=1, \ldots, n), \tag{8.5}$$

the property (8.4) is equivalent to the set relation

$$\mathbf{R}^k=\bigcup_{\nu=1}^n M_\nu^k, \tag{8.6}$$

which is the definition (3.3) of an n-chromo. The steps can be reversed and establish

LEMMA 4. *Let the relations* (8.2) *hold. The reflected signal* (8.1) *is ϱ-admissible if and only if*

$$\chi_n^k(\delta)=\{M_1^k(\delta), \ldots, M_n^k(\delta)\}, \tag{8.7}$$

defined by (8.5), *is an n-chromo. That* (8.1) *is in general position if and only if* (8.5) *is admissible is obvious, because they are expressed by the same condition on the matrix $\Lambda=\|\lambda_\nu^i\|$.*

9. Applications of Lemma 4: Proofs of Theorems 1, 2, and 3

The relation (8.2), Lemma 4, and the definitions (1.11) of $\varrho_{k,n}$, and (4.3) of $\delta_{k,n}$, show that

$$\varrho_{k,n} = \frac{1}{2} - \frac{\delta_{k,n}}{2}, \tag{9.1}$$

or

$$\delta_{k,n} = 1 - 2\varrho_{k,n}. \tag{9.2}$$

By Theorem 1′ $\delta_{k,n} \leqq \frac{k}{n}$ and (9.1) implies that $\varrho_{k,n} \geqq \frac{1}{2} - \frac{k}{2n}$ and Theorem 1 is established.

By Theorem 2′ $\delta_{n-1,n} = \frac{n-1}{n}$ and (9.1) implies that $\varrho_{n-1,n} = \frac{1}{2} - \frac{n-1}{2n} = \frac{1}{2n}$ and Theorem 2 is proved.

Let us use Lemma 4 to derive for $k=n-1$ the equations for the signal $\tilde{\Pi}_n^{n-1}\left(\frac{1}{2n}\right)$. From the relations (6.3), (6.4), we find by Lemma 4 for this signal the equations

$$\tilde{\Pi}_n^{n-1}\left(\frac{1}{2n}\right): \begin{array}{ll} x_\nu = \langle u_\nu - \frac{1}{2}\rangle, & (\nu = 1, \dots, n-1). \\ x_n = \langle \sum_1^{n-1} u_i \rangle. & \end{array} \tag{9.3}$$

Replacing here u_ν by $u_\nu + \frac{1}{2}$, we obtain

$$\tilde{\Pi}_n^{n-1}\left(\frac{1}{2n}\right): \begin{array}{ll} x_\nu = \langle u_\nu \rangle, & (\nu = 1, \dots, n), \\ x_n = \left\langle \sum_1^{n-1} u_i + \frac{n-1}{2} \right\rangle, & \end{array} \tag{9.4}$$

which are identical with (1.15.) The essential unicity of the n-chromo (6.3), (6.4), established in § 6, implies the essential unicity of the signal (9.4).

In the special case that $n=3$, we obtain *Kepler's tetrahedron T* mentioned in connection with the relation (1.14). By (9.4) its parametric equations are

$$\tilde{\Pi}_3^2\left(\frac{1}{6}\right): x_1 = \langle u_1 \rangle, \quad x_2 = \langle u_2 \rangle, \quad x_3 = \langle u_1 + u_2 + 1 \rangle. \tag{9.5}$$

The vertices of $T = ABCD$ are

$$A = (0, 0, 1), \quad B = (1, 0, 0), \quad C = (0, 1, 0), \quad D = (1, 1, 1).$$

An even simpler case is $n=2$ when

$$\tilde{\Pi}_2^1\left(\frac{1}{4}\right): x_1 = \langle u_1 \rangle, \quad x_2 = \left\langle u_1 + \frac{1}{2} \right\rangle. \tag{9.6}$$

This is the square having as vertices the midpoints of the sides of U_2 (see [5]).

As a last application of Lemma 4 we use Theorem 3′ to give the explicit constructions of the two signals for the cases (1.16) of Theorem 3. From (7.4), and Lemma 4, we find immediately

$$\Pi_4^2\left(\frac{1}{4}\right): \begin{array}{ll} x_1 = \left\langle u_1 - \frac{1}{2}\right\rangle, & x_2 = \left\langle u_2 - \frac{1}{2}\right\rangle, \\ x_3 = \left\langle \frac{u_1+u_2}{2}\right\rangle, & x_4 = \left\langle \frac{u_1-u_2+1}{2}\right\rangle. \end{array}$$

Replacing u_i by $u_i + \frac{1}{2}$, we obtain

$$\Pi_4^2\left(\frac{1}{4}\right): \begin{array}{ll} x_1 = \langle u_1\rangle, & x_2 = \langle u_2\rangle, \\ x_3 = \left\langle \frac{u_1+u_2+1}{2}\right\rangle, & x_4 = \left\langle \frac{u_1-u_2+1}{2}\right\rangle. \end{array}$$

Likewise, (7.6), (7.7), (7.8), and Lemma 4, show that

$$\Pi_6^2\left(\frac{1}{3}\right): \begin{array}{ll} x_1 = \langle u_1\rangle, & x_2 = \langle u_2\rangle, \\ x_3 = \langle u_1+u_2+1\rangle, & x_4 = \left\langle \frac{u_1-u_2}{3}\right\rangle, \\ x_5 = \left\langle \frac{2u_1+u_2}{3}+1\right\rangle, & x_6 = \left\langle \frac{u_1+2u_2}{3}+1\right\rangle. \end{array}$$

It is to be expected that these explicit parametric equations, as well as (9.4), should reveal pertinent geometric aspects of the polytopes that they represent.

Our approach via n-chromos suggests that a promising attack on the three problems stated at the end of § 1, should be to solve the corresponding problems for n-chromos in $\mathbf{R}^k$. These are:

1′°. A proof of Conjecture 1′.

2′°. A general arithmetic-analytic construction of the admissible n-chromo

$$\chi_n^k\left(\frac{k}{n}\right). \tag{9.7}$$

3′°. A proof that the number of n-chromos (9.7), no two of which are affinely equivalent, is finite. This was done in [4] for $k=1$.

APPENDIX. EXTREMUM PROBLEMS FOR LISSAJOUS-TYPE MANIFOLDS

10. Applications of n-chromos to Lissajous-type manifolds

In [3, § 6] we discussed our extremum problem for Lissajous curves in the unit cube $-\frac{1}{2} \leqq x_\nu \leqq \frac{1}{2}$, $(\nu=1, \ldots, n)$, the underlying norm being the Euclidean one. Here two changes alter the situation:

1. We replace the above cube by our cube U_n of (1.1). This requires replacing the basic function $w(x)=\cos x$ of [3, § 6] by the function

$$L(x) = \sin^2 \frac{\pi x}{2}, \qquad \text{(see Figure 4).} \tag{10.1}$$

Observe that $L(x)$ interpolates at the integers the zigzag curve of $\langle x \rangle$ defined by (1.3). The absence of corners assures the smoothness of the resulting motions within U_n. However, the 1-dimensional Lissajous motions

$$x_\nu = L(\lambda_\nu t + a_\nu), \qquad (\nu = 1, \ldots, n), \tag{10.2}$$

of [3, relation (6.3)] again exhibit the ergodic (or denseness) property described for b.b. motions in the second paragraph of our Introduction. For this reason, and following again the lead of König and Szűcs, we replace the motion (10.2) by the k-dimensional Lissajous-type manifold

$$\Lambda_n^k\colon\ x_\nu = L\Big(\sum_{i=1}^{k} \lambda_\nu^i u_i + a_\nu\Big), \qquad (\nu = 1, \ldots, n). \tag{10.3}$$

2. We replace the Euclidean norm of [3] by the L_∞ norm of the present paper.

The Definitions 1 and 2, of § 1, concerning the reflected path (1.7) carry over without any changes to the L-manifold (10.3). We may therefore safely assume that we know what is meant by "a Λ_n^k in general position", and by "a Λ_n^k that is ϱ^L-admissible". The latter will again be denoted by $\Lambda_n^k(\varrho^L)$.

As in the Introduction we propose

PROBLEM 1^L. *To determine, or estimate, the quantity*

$$\varrho_{k,n}^L = \text{supremum } \varrho^L, \tag{10.4}$$

the supremum being taken for all ϱ^L having ϱ^L-admissible L-type manifolds $\Lambda_n^k(\varrho^L)$.

It does seem remarkable that our results of Part I, on n-chromos in $\mathbf{R}^k$, apply equally well to establish Theorems 1^L, 2^L, and 3^L, below, that correspond to Theorems 1, 2, and 3, on b.b. motions. In particular the $\delta_{k,n}$ below, is again the old constant (4.3) for n-chromos. These theorems are as follows.

THEOREM 1^L. *We have the inequality*

$$\varrho_{k,n}^L \geqq \frac{1}{2} - \sin^2\left(\frac{\pi}{2}\frac{k}{2n}\right), \quad (1 \leqq k \leqq n-1). \tag{10.5}$$

THEOREM 2^L. *We have that*

$$\varrho_{n-1,n}^L = \frac{1}{2} - \sin^2\left(\frac{\pi}{2}\frac{n-1}{2n}\right). \tag{10.6}$$

THEOREM 3^L. *We construct explicitly the L-type manifold*

$$\Lambda_n^k(\varrho^L), \quad \textit{where} \quad \varrho^L = \frac{1}{2} - \sin^2\left(\frac{\pi}{2}\frac{k}{2n}\right), \tag{10.7}$$

for the two cases

$$(k, n) = (2, 4) \quad \textit{and} \quad (k, n) = (2, 6). \tag{10.8}$$

At this point we need an analogue of Lemma 4, that we shall call Lemma 4^L, which will relate L-type manifolds to n-chromos. Let (10.3) be ϱ^L-admissible. This means that for every $(u_i)\in\mathbf{R}^k$, the point of $\mathbf{R}^n$

(10.9) $$L\Big(\sum_1^k \lambda_\nu^i u_i+a_\nu\Big), \qquad (\nu = 1, \dots, n),$$

should belong to the closed cubical shell

(10.10) $$B^n_{\varrho^L} = U_n - C^n_{\varrho^L}, \quad \text{where} \quad C^n_{\varrho^L} = \{\|x-c\|_\infty < \varrho^L\}.$$

Equivalently:

(10.11) *For some ν, the number $L(\sum \lambda_\nu^i u_i+a_\nu)$ should differ from an integer by $\le\delta^L/2$, where*

(10.12) $$\frac{\delta^L}{2} = \frac{1}{2}-\varrho^L.$$

How is this condition expressed in terms of $\sum \lambda_\nu^i u_i+a_\nu$? If we define $\delta/2$ as a solution of the equation

(10.13) $$\frac{\delta^L}{2} = L\left(\frac{\delta}{2}\right),$$

then the symmetries of the graph of $L(x)$ show (Figure 4) that (10.11) will hold if and only if

(10.14) $$\Big\{\sum_1^k \lambda_\nu^i u_i+a_\nu\Big\} \le \frac{\delta}{2}.$$

This establishes

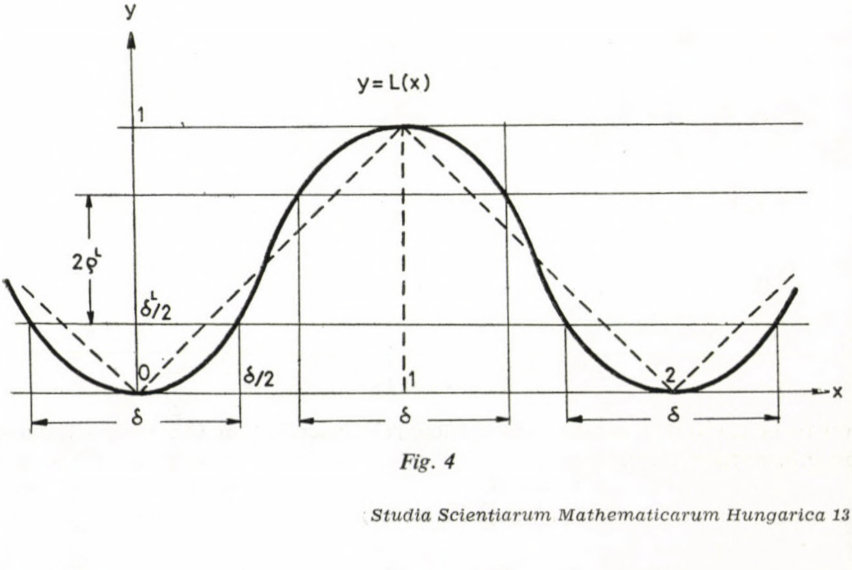


Fig. 4

LEMMA 4^L. *Let* $\varrho^L, 0<\varrho^L<\frac{1}{2}$, *be prescribed, then* δ^L *be defined by* (10.12), *and finally* δ *such that* (10.13) *holds. The L-manifold* (10.3) *is* ϱ^L*-admissible if and only if the n monochromes*

$$(10.15) \qquad M_\nu^k(\delta)\colon \Big\{\sum_1^k \lambda_\nu^i u_i + a_\nu\Big\} \leqq \frac{\delta}{2}, \qquad (\nu = 1, \dots, n),$$

define an n-chromo in $\mathbf{R}^k$.

Eliminating δ^L between (10.12) and (10.13), we find that

$$(10.16) \qquad \varrho^L = \frac{1}{2} - L\left(\frac{\delta}{2}\right).$$

If ϱ^L tends to its supremum $\varrho^L_{k,n}$, then δ tends to its infimum $\delta_{k,n}$, and we obtain

$$(10.17) \qquad \varrho^L_{k,n} = \frac{1}{2} - L\left(\frac{\delta_{k,n}}{2}\right),$$

which is the analogue of (9.1).

THEOREM 1′, hence that $\delta_{k,n} \leqq k/n$, and (10.17), immediately establishes (10.5), hence Theorem 1^L. Likewise Theorem 2′, hence that $\delta_{n-1,n} = (n-1)/n$, gives (10.6), hence Theorem 2^L, again in view of (10.17). Finally, Theorem 3′ implies Theorem 3^L.

11. Examples of extremal Lissajous manifolds

It was not mentioned above, but is evident by Lemma 4^L, that if (10.15) are the inequalities defining an n-chromo $\chi_n^k(\delta)$, then (10.3) defines an L-manifold which is ϱ^L-admissible, where ϱ^L is defined by (10.16). As an example, the n-chromo $\chi_n^{n-1}(\delta)$, defined by (6.2), (6.3), (6.4), give the L-manifold of Theorem 2^L

$$(11.1) \qquad \Lambda_n^{n-1}(\varrho^L_{n-1,n})\colon \begin{array}{l} x_\nu = \sin^2\left(\frac{\pi}{2} u_\nu\right), \quad (\nu = 1, \dots, n-1), \\ x_n = \sin^2 \frac{\pi}{2}\left(u_1 + \dots + u_{n-1} + \frac{n-1}{2}\right) \end{array}$$

where $\varrho^L_{n-1,n}$ is given by (10.6).

Let us look at this for the smallest values of n.

1. **k=1, n=2.** Here $\varrho^L_{1,2} = \frac{1}{2} - \sin^2(\pi/8) = (2\sqrt{2})^{-1}$. The extremizing L-motion

$$x_1 = \sin^2 \frac{\pi u_1}{2} = \frac{1}{2}(1 - \cos \pi u_1),$$

$$x_2 = \sin^2\left(\frac{\pi}{2} u_1 + \frac{\pi}{4}\right) = \frac{1}{2}(1 + \sin \pi u_1),$$

is seen to be a circular motion along the circle inscribed in U_2. This is the analogue of the b.b. motion (9.6).

2. **k=2, n=3.** The extremizing L-surface is found to be

$$x_1 = \sin^2\frac{\pi u_1}{2} = \frac{1}{2}(1-\cos \pi u_1),$$

$$(11.2) \qquad x_2 = \sin^2\frac{\pi u_2}{2} = \frac{1}{2}(1-\cos \pi u_2),$$

$$x_3 = \cos^2\frac{\pi}{2}(u_1+u_2) = \frac{1}{2}(1+\cos \pi(u_1+u_2)).$$

This is the L-analogue of Kepler's tetrahedron T parametrically given by (9.5). The largest cube inscribed in T was found to have its side $=\frac{1}{3}$. For our Λ_3^2 we find a *larger* cube $\|x-c\|_\infty<\frac{1}{4}$ of side $=\frac{1}{2}$, because

$$\varrho_{2,3}^L = \frac{1}{2}-\sin^2\frac{\pi}{6} = \frac{1}{4}.$$

The intersections of (11.2) with the planes $x_\nu=c$ $(0\leqq c\leqq 1)$, $(\nu=1, 2, 3)$ are ellipses, inscribed in the unit square, with axes parallel to the diagonals of the square. The surface is convex.

From (10.17), for $k=n-1$, we find that

$$(11.3) \qquad \lim_{n\to\infty} \varrho_{n-1,n}^L = 0.$$

Most likely the above Λ_3^2, given by (11.2), is the last Λ_n^{n-1} which is the boundary of a convex set in $\mathbf{R}^n$.

3. **k=1, general n.** With this last example we come close to the subject studied in [4]. The n-chromo (4.7) and Lemma 4^L show that

$$(11.4) \qquad x_\nu = \sin^2\frac{\pi}{2}\left(u_1+\frac{\nu-1}{n}\right), \qquad (\nu = 1, \dots, n;\ 0\leqq u_1\leqq 2),$$

describe an extremal curve Λ_n^1. From (10.17), for $k=1$, we obtain that

$$(11.5) \qquad \lim_{n\to\infty} \varrho_{1,n}^L = \frac{1}{2}.$$

The curve (11.4) is the Lissajous-analogue of the "lucky" billiard ball shot Γ_n^* of [4, relation (10.2) for $n=3$, and Figure 2].

REFERENCES

[1] LETTENMEYER, F., Neuer Beweis des allgemeinen Kroneckerschen Approximationssatzes, *Proc. London Math. Soc.* Ser. 2 **21** (1922), 306—314.

[2] KÖNIG, D.—SZŰCS, A., Mouvement d'un point abandonné à l'intérieur d'un cube, *Rendiconti del Circ. Mat. di Palermo* **38** (1913), 79—90.

[3] SCHOENBERG, I. J., Extremum problems for the motions of a billiard ball I. The L_p norm, $1 \leqq p < \infty$, *Indag. Math.* **38** (1976), 66—75.

[4] SCHOENBERG, I. J., Extremum problems for the motions of a billiard ball II. The L_∞ norm, *Indag. Math.* **38** (1976), 263—279

[5] SCHOENBERG, I. J., On the motion of a billiard ball in two dimensions, *Delta,* Madison, Wisconsin **5** (1975), 1—18.

United States Military Academy, West Point, New York
and
Mathematics Research Center, University of Wisconsin—Madison

(Received October 28, 1978)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 79—104.

ON MULTIPLICATIVE ARITHMETIC FUNCTIONS SATISFYING A LINEAR RECURSION

by

A. SÁRKÖZY

1. An arithmetic function $f(n)$ is said to be *multiplicative* if $(m, n)=1$ implies that

$$f(mn)=f(m)f(n) \tag{1}$$

and it is *strictly multiplicative* if (1) holds for all m, n. Furthermore, $f(n)$ is said to be *additive* if $(m, n)=1$ implies that $f(mn)=f(m)+f(n)$. Let m be a positive integer. An arithmetic function $\chi(n)$ is said to be a *character modulo m* if it is strictly multiplicative, $\chi(a)=\chi(b)$ for all a, b such that $a\equiv b \pmod m$, finally, $\chi(n)=0$ holds if and only if $(n, m)>1$. By the multiplicativity of the characters and since $\chi(n)\not\equiv 0$, we have $\chi(1)=1$ for all characters. If $\chi(n)=1$ for all n satisfying $(n, m)=1$ then $\chi(n)$ is called the *principal character* modulo m; the principal character is denoted by χ_0. If $(a, b)=1$ and $\chi(n)$ is a character modulo ab then there exist uniquely determined characters χ_1 and χ_2 modulo a and b, respectively, such that $\chi(n)=\chi_1(n)\chi_2(n)$ for all n. (For the theory of characters see, e.g., [1] or [4].) The number of the elements of a finite set S will be denoted by $|S|$. For a real number x, we write $\{x\}=x-[x]$ (where $[x]$ denotes the integer part of x).

In [3], L. Lovász, M. Simonovits and the author determined all the *additive* arithmetic functions $f(n)$ which satisfy a linear recursion of finite order, i.e., an equality of the form

$$a_0f(n)+a_1f(n+1)+\ldots+a_kf(n+k)=0, \quad n=1,2,\ldots \tag{2}$$

where k is a positive integer and $a_0, a_1, \ldots, a_k$ are any complex numbers such that

$$a_0\neq 0 \quad\text{and}\quad a_k\neq 0. \tag{3}$$

The aim of this paper is *to determine all the multiplicative arithmetic functions $f(n)$ which satisfy a linear recursion of finite order.* As Theorem 2 will show, we get a characterization of the (modulo m) characters in this way.

Let $\mathscr{A}$ denote the set of the multiplicative arithmetic functions which satisfy a linear recursion of form (1).

Theorem 1. *A multiplicative arithmetic function $f(n)$ belongs to $\mathscr{A}$ if and only if $f(n)\equiv 0$, or there exist a non-negative integer h, a positive integer m and a character $\chi(n)$ modulo m, satisfying the following conditions:*

(i) *If $(n, m)=1$ then $f(n)=n^h\chi(n)$.*

(ii) *If $m\geqq 2$ then let $m=p_1^{\alpha_1}p_2^{\alpha_2}\ldots p_r^{\alpha_r}$ (where $p_1, p_2, \ldots, p_r$ are distinct prime numbers, $\alpha_1, \alpha_2, \ldots, \alpha_r$ are positive integers). Let $\chi_1(n), \chi_2(n), \ldots, \chi_r(n)$ denote the uniquely determined characters modulo $p_1^{\alpha_1}, p_2^{\alpha_2}, \ldots, p_r^{\alpha_r}$, respectively, such that $\chi(n)=$*

$=\chi_1(n)\chi_2(n)\dots\chi_r(n)$ *for all* n. *Then for* $i=1,2,\dots,r$ *and* $j=0,1,2,\dots$ *we have*

$$f(p_i^{\alpha_i+j})=\begin{cases} p_i^{jh}f(p_i^{\alpha_i})\prod\limits_{\substack{1\leq l\leq r\\ l\neq i}}\chi_l(p_i^j) & \textit{if } \chi_i(n) \textit{ is the principal character modulo } p_i^{\alpha_i},\\ 0 & \textit{if } \chi_i(n) \textit{ is not the principal character modulo } p_i^{\alpha_i}.\end{cases}$$

Let $\mathcal{A}^*$ denote the set consisting of the *strictly* multiplicative arithmetic functions $f(n)$ which satisfy the following conditions:

(i) $f(n)$ satisfies a linear recursion of finite order;

(ii) $f(n)\not\equiv 0$;

(iii) $f(n)=o(n)$.

Theorem 1 implies trivially

THEOREM 2. *A multiplicative arithmetic function* $f(n)$ *belongs to* $\mathcal{A}^*$ *if and only if there exists a positive integer* m *such that* $f(n)$ *is a character modulo* m.

In fact, $f(n)\in\mathcal{A}^*$ implies by Theorem 1 that $f(n)$ must satisfy (i) and (ii) in Theorem 1. The condition $f(n)=o(n)$ implies that $h=0$. If $f(n)$ is strictly multiplicative then (ii) in Theorem 1 yields that either $f(p_i)=0$ or $f(p_i)=\prod\limits_{\substack{1\leq l\leq r\\ l\neq i}}\chi_l(p_i)$. We may assume that

$$f(p_1)=\dots=f(p_s)=0 \quad\text{and}\quad f(p_{s+1})\neq 0,\dots,f(p_r)\neq 0.$$

Then it can be shown easily that $f(n)$ is a character modulo $p_1^{\alpha_1}p_2^{\alpha_2}\dots p_s^{\alpha_s}$.

2. Proof of Theorem 1. The essential part of the proof is to show that a multiplicative arithmetic function satisfies $f(n)\in\mathcal{A}$ if and only if $f(n)/n^h$ is *periodic* for a non-negative integer h; on the other hand, it can be shown easily that $f(n)/n^h$ is periodic if and only if $f(n)\equiv 0$ or both (i) and (ii) hold.

In this section, we prove that *if* $f(n)$ *is a multiplicative arithmetic function such that* $f(n)\equiv 0$ *or it satisfies both conditions* (i) *and* (ii) *in Theorem 1 then* $f(n)$ *belongs to* $\mathcal{A}$.

If $f(n)\equiv 0$ then, obviously, (1) holds for any complex numbers $a_0,a_1,\dots,a_k$ satisfying (2) thus we have $f(n)\in\mathcal{A}$.

Assume now that $f(n)$ is a multiplicative arithmetic function satisfying the conditions (i) and (ii) in Theorem 1. We are going to show that these conditions imply that the function $g(n)=f(n)/n^h$ is periodic with period $Q=m^2$, i.e., we have

$$g(n+Q)=g(n). \tag{4}$$

If $m=1$ then (i) implies that $g(n)\equiv 1$ thus (4) holds trivially. Assume now that $2\leq m=p_1^{\alpha_1}p_2^{\alpha_2}\dots p_r^{\alpha_r}$ and for $i=1,2,\dots,r$, define the multiplicative arithmetic function $g_i(n)$ in the following way: let

$$g_i(n)=\chi_i(n) \quad\text{if}\quad (n,p_i)=1 \tag{5}$$

where the character $\chi_i(n)$ is defined in (ii) and for $1\leq i\leq r$, $\gamma=0,1,2,\dots,$ let

$$g_i(p_i^\gamma)=\frac{f(p_i^\gamma)}{p_i^{h\gamma}\prod\limits_{\substack{1\leq l\leq r\\ l\neq i}}\chi_l(p_i^\gamma)}. \tag{6}$$

Let $n=p_1^{\gamma_1}p_2^{\gamma_2}\dots p_r^{\gamma_r}n_1$ be an arbitrary integer where $\gamma_1, \gamma_2, \dots, \gamma_r$ are non-negative integers and $(n_1, m)=1$. Then by (i) and (ii), we have

$$\begin{aligned}
\prod_{i=1}^{r} g_i(n) &= \prod_{i=1}^{r} g_i(p_1^{\gamma_1}p_2^{\gamma_2}\dots p_r^{\gamma_r}n_1) = \\
&= \prod_{i=1}^{r} g_i(p_1^{\gamma_1})g_i(p_2^{\gamma_2})\dots g_i(p_r^{\gamma_r})g_i(n_1) = \\
&= \prod_{i=1}^{r} \{(\prod_{\substack{1\leqq j\leqq r\\ j\neq i}} g_i(p_j^{\gamma_j}))g_i(p_i^{\gamma_i})g_i(n_1)\} = \\
&= \prod_{i=1}^{r} \Big\{(\prod_{\substack{1\leqq j\leqq r\\ j\neq i}} \chi_i(p_j^{\gamma_j}))\frac{f(p_i^{\gamma_i})}{p_i^{h\gamma_i}\prod\limits_{\substack{1\leqq l\leqq r\\ l\neq i}}\chi_l(p_i^{\gamma_i})}\chi_i(n_1)\Big\} = \\
&= \prod_{i=1}^{r}\frac{f(p_i^{\gamma_i})}{p_i^{h\gamma_i}}\prod_{i=1}^{r}\chi_i(n_1)\frac{\prod\limits_{\substack{1\leqq i,\, j\leqq r\\ i\neq j}}\chi_i(p_j^{\gamma_j})}{\prod\limits_{\substack{1\leqq i,\, j\leqq r\\ i\neq j}}\chi_i(p_j^{\gamma_j})} = \\
&= \frac{f(p_1^{\gamma_1}p_2^{\gamma_2}\dots p_r^{\gamma_r})}{(p_1^{\gamma_1}p_2^{\gamma_2}\dots p_r^{\gamma_r})^h}\chi(n_1) = \frac{f(p_1^{\gamma_1}p_2^{\gamma_2}\dots p_r^{\gamma_r})}{(p_1^{\gamma_1}p_2^{\gamma_2}\dots p_r^{\gamma_r})^h}\frac{f(n_1)}{n_1^h} = \\
&= \frac{f(p_1^{\gamma_1}p_2^{\gamma_2}\dots p_r^{\gamma_r}n_1)}{(p_1^{\gamma_1}p_2^{\gamma_2}\dots p_r^{\gamma_r}n_1)^h} = \frac{f(n)}{n^h}.
\end{aligned}\tag{7}$$

Now we are going to show that

$$g_i(n+Q) = g_i(n) \quad \text{for} \quad i = 1, 2, \dots, r \tag{8}$$

and for all n. Let us write n in the form $n=p_i^{\gamma}n_1$ where γ is a non-negative integer and $(p_i, n_1)=1$. By (5), we have

$$g_i(n) = g_i(p_i^{\gamma}n_1) = g_i(p_i^{\gamma})g_i(n_1) = g_i(p_i^{\gamma})\chi_i(n_1). \tag{9}$$

Assume first that $\gamma<\alpha_i$. Then there exists an integer t such that

$$\begin{aligned}
n+Q &= p_i^{\gamma}n_1+p_1^{2\alpha_1}\dots p_i^{2\alpha_i}\dots p_r^{2\alpha_r} = \\
&= p_i^{\gamma}(n_1+p_i^{\alpha_i}\cdot p_1^{2\alpha_1}\dots p_{i-1}^{2\alpha_{i-1}}p_i^{\alpha_i-\gamma}p_{i+1}^{2\alpha_{i+1}}\dots p_r^{2\alpha_r}) = p_i^{\gamma}(n_1+p_i^{\alpha_i}t)
\end{aligned}$$

thus

$$\begin{aligned}
g_i(n+Q) &= g_i\big(p_i^{\gamma}(n_1+p_i^{\alpha_i}t)\big) = g_i(p_i^{\gamma})g_i(n_1+p_i^{\alpha_i}t) = \\
&= g_i(p_i^{\gamma})\chi_i(n_1+p_i^{\alpha_i}t) = g_i(p_i^{\gamma})\chi_i(n_1).
\end{aligned}\tag{10}$$

(9) and (10) yield (8) in this case.

Assume now that $\gamma\geqq\alpha_i$. Then $p_i^{\alpha_i}|n$ and $p_i^{\alpha_i}|Q$ imply that $p_i^{\alpha_i}|n+Q$ thus $n+Q$ can be written in the form $n+Q=p_i^{\delta}n_2$ where $\delta\geqq\alpha_i$.

If $\chi_i(n)$ is not the principal character modulo $p_i^{\alpha_i}$ then by (ii), (6) and (9), we have

$$(11)\qquad g_i(n) = g_i(p_i^{\gamma})\chi_i(n_1) = \frac{f(p_i^{\gamma})}{p_i^{h\gamma} \prod\limits_{\substack{1\leq l\leq r\\ l\neq i}} \chi_l(p_i^{\gamma})}\chi_i(n_1) = 0$$

and

$$(12)\qquad g_i(n+Q) = g_i(p_i^{\delta})\chi_i(n_2) = \frac{f(p_i^{\delta})}{p_i^{h\delta} \prod\limits_{\substack{1\leq l\leq r\\ l\neq i}} \chi_l(p_i^{\delta})}\chi_i(n_2) = 0.$$

(11) and (12) yield (8).

If $\chi_i(n)$ is the principal character modulo $p_i^{\alpha_i}$ then again by (ii), (6) and (9), we have

$$(13)\qquad \begin{aligned} g_i(n) &= g_i(p_i^{\gamma})\chi_i(n_1) = \frac{f(p_i^{\gamma})}{p_i^{h\gamma} \prod\limits_{\substack{1\leq l\leq r\\ l\neq i}} \chi_l(p_i^{\gamma})}\cdot 1 = \\ &= \frac{p_i^{h(\gamma-\alpha_i)} f(p_i^{\alpha_i}) \prod\limits_{\substack{1\leq l\leq r\\ l\neq i}} \chi_l(p_i^{\gamma-\alpha_i})}{p_i^{h\gamma} \prod\limits_{\substack{1\leq l\leq r\\ l\neq i}} \chi_l(p_i^{\gamma-\alpha_i})\chi_l(p_i^{\alpha_i})} = \\ &= p_i^{-h\alpha_i} f(p_i^{\alpha_i}) \prod\limits_{\substack{1\leq l\leq r\\ l\neq i}} \left(\chi_l(p_i^{\alpha_i})\right)^{-1} \end{aligned}$$

and

$$(14)\qquad \begin{aligned} g_i(n+Q) &= g_i(p_i^{\delta})\chi_i(n_2) = \frac{f(p_i^{\delta})}{p_i^{h\delta} \prod\limits_{\substack{1\leq l\leq r\\ l\neq i}} \chi_l(p_i^{\delta})}\cdot 1 = \\ &= \frac{p_i^{h(\delta-\alpha_i)} f(p_i^{\alpha_i}) \prod\limits_{\substack{1\leq l\leq r\\ l\neq i}} \chi_l(p_i^{\delta-\alpha_i})}{p_i^{h\delta} \prod\limits_{\substack{1\leq l\leq r\\ l\neq i}} \chi_l(p_i^{\delta-\alpha_i})\chi_l(p_i^{\alpha_i})} = \\ &= p_i^{-h\alpha_i} f(p_i^{\alpha_i}) \prod\limits_{\substack{1\leq l\leq r\\ l\neq i}} \left(\chi_l(p_i^{\alpha_i})\right)^{-1}. \end{aligned}$$

(13) and (14) yield (8) also in this case and this completes the proof of (8).

(7) and (8) yield (4).

Now we are ready to prove that $f(n)\in\mathscr{A}$. Let us put $k=Q(h+1)$ and let

$$a_{iQ+j} = \begin{cases} (-1)^i\binom{h+1}{i} & \text{for}\quad i=0,1,\dots,h+1 \quad\text{and}\quad j=0,\\ 0 & \text{for}\quad i=0,1,\dots,h \quad\text{and}\quad j=1,2,\dots,Q-1. \end{cases}$$

Then by (4), we have

$$\sum_{l=0}^{k} a_l f(n+l) = \sum_{i=0}^{h+1} a_{iQ} f(n+iQ) =$$

$$= \sum_{i=0}^{h+1} (-1)^i \binom{h+1}{i} (n+iQ)^h g(n+iQ) = \tag{15}$$

$$= g(n) \sum_{i=0}^{h+1} (-1)^i \binom{h+1}{i} (n+iQ)^h.$$

Let us write

$$F(x) = Q^h x^h$$

and

$$\Delta_{h+1} F(x) = \sum_{i=0}^{h+1} (-1)^i \binom{h+1}{i} F(x+i).$$

Then we have

$$\Delta_{h+1} F(x) = 0$$

for all x since it is well-known that the k^{th} difference of a polynomial of degree $\leqq k-1$ is 0. (This can be shown, e.g., by straight induction on k). Thus (15) yields that

$$\sum_{l=0}^{k} a_l f(n+l) = g(n) \Delta_{h+1} F\left(\frac{n}{Q}\right) = 0$$

which proves that $f(n) \in \mathcal{A}$.

3. Now we have to prove that $f(n) \in \mathcal{A}$ *and*

$$f(n) \not\equiv 0 \tag{16}$$

imply that there exist h, m and χ satisfying (i) *and* (ii) *in Theorem 1.*

We need some lemmas; note that in Lemmas 1—6, we shall not use the multiplicativity of $f(n)$.

LEMMA 1. *If an arithmetic function $f(n)$ satisfies* (16) *and a linear recursion of form* (2) *then there exist positive integers $t, r_1, r_2, \dots, r_t$ and complex numbers $s_1, s_2, \dots, s_t, d_{11}, d_{12}, \dots, d_{1r_1}, d_{21}, d_{22}, \dots, d_{2r_2}, \dots, d_{t1}, d_{t2}, \dots, d_{tr_t}$ such that*

$$s_1 \neq 0,\ s_2 \neq 0, \dots, s_t \neq 0,\ s_i \neq s_j \text{ for } i \neq j,\ d_{1r_1} \neq 0,\ d_{2r_2} \neq 0, \dots, d_{tr_t} \neq 0 \tag{17}$$

and

$$f(n) = \sum_{i=1}^{t} \left(\sum_{j=1}^{r_i} d_{ij} \binom{n+j-1}{j-1} \right) s_i^n. \tag{18}$$

This lemma and its proof can be found in [3]. (See formula (12).)

By Lemma 1, $f(n)\in\mathscr{A}$ and (16) imply that $f(n)$ can be written in the form (18). Thus to complete the proof of Theorem 1, it is sufficient to show that if the function $f(n)$ is of form (18) (where (17) holds) then there exist h, m and χ satisfying (i) and (ii) in Theorem 1. We may assume that we have

$$|s_1| = |s_2| = \ldots = |s_u| > |s_{u+1}| \geqq |s_{u+2}| \geqq \ldots \geqq |s_t| \tag{19}$$

and

$$r_1 \geqq r_2 \geqq \ldots \geqq r_u \tag{20}$$

in (18) (also $u=1$ may occur). In order to simplify the notation, we put $r_1=r$ and $d_{1r_1}=d_{1r}$.

In the remaining part of the paper, $C_1, C_2, \ldots$ will denote positive constants which may depend on certain parameters $r_1, \ldots, r_t, s_1, \ldots, s_t, d_{11}, \ldots, d_{tr_t}$ but which are independent of the variables n, M, L. Similarly, we write, e.g., $g(n, r_1, \ldots, r_t, s_1, \ldots, s_t, d_{11}, \ldots, d_{tr_t}) = O(|h(n, r_1, \ldots, r_t, s_1, \ldots, s_t, d_{11}, \ldots, d_{tr_t})|)$ if $|g(n)|<C_1|h(n)|$ for $n=1, 2, \ldots$ where C_1 is a positive constant of type described above.

Lemma 2. *If $f(n)$ is an arithmetic function of form* (18) *where* (17), (19) *and* (20) *hold then there exists a constant $C_1=C_1(r_1, \ldots, r_t, d_{11}, \ldots, d_{tr_t}, s_1, \ldots, s_t)$ such that*

$$|f(n)| < C_1\binom{n+r-1}{r-1}|s_1|^n. \tag{21}$$

Proof. By (18), (19) and (20), we have

$$|f(n)| \leqq \sum_{i=1}^{u}\left(\sum_{j=1}^{r_i}|d_{ij}|\binom{n+r_1-1}{r_1-1}\right)|s_1|^n+$$

$$+\sum_{i=u+1}^{t}\left(\sum_{j=1}^{r_i}|d_{ij}|\binom{n+r_i-1}{r_i-1}\right)|s_{u+1}|^n =$$

$$= O\left(\binom{n+r-1}{r-1}|s_1|^n\right)+O\left(\max_{u\leqq i\leqq t}\binom{n+r_i-1}{r_i-1}|s_{u+1}|^n\right) =$$

$$= O\left(\binom{n+r-1}{r-1}|s_1|^n\right)+o(|s_1|^n) = O\left(\binom{n+r-1}{r-1}|s_1|^n\right)$$

which yields (21) and this proves Lemma 2.

Lemma 3. *Let U, V be positive integers such that $U\leqq V$, K a non-negative integer, z a complex number satisfying $|z|\leqq 1, z\neq 1$. Let*

$$S(U, V, K, z) = \binom{U+K}{K}z^U+\binom{U+K+1}{K}z^{U+1}+\ldots+\binom{V+K}{K}z^V.$$

Then we have

$$|S(U, V, K, z)| < 2^{K+2}\binom{V+K+1}{K}\frac{|z|^U}{|1-z|^{K+1}}.$$

This lemma is identical to Lemma 1 in [3].

LEMMA 4. *Let $f(n)$ be an arithmetic function of form* (18) *where* (17), (19) *and* (20) *hold and let D be a positive integer such that*

$$(s_i/s_1)^D \neq 1 \quad \textit{for} \quad i = 2, 3, \dots, t. \tag{22}$$

Then there exists a constant $C_2 = C_2(r_1, \dots, r_t, d_{11}, \dots, d_{tr_t}, s_1, \dots, s_t, D)$ such that for any positive integers M, L we have

$$\Bigg|\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \ (\mathrm{mod}\ D)}} f(n)s_1^{-n} - d_{1r}\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \ (\mathrm{mod}\ D)}}\binom{n+r-1}{r-1}\Bigg| <$$

$$< C_2\left(\binom{M+LD+r-1}{r-1} + (M+LD+R)^R|s_{u+1}/s_1|^M\right)$$

where $R = \max\limits_{u<i\leqq t} r_i$.

PROOF. (18) yields that

$$\Bigg|\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \ (\mathrm{mod}\ D)}} f(n)s_1^{-n} - d_{1r}\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \ (\mathrm{mod}\ D)}}\binom{n+r-1}{r-1}\Bigg| = \tag{23}$$

$$= \Bigg|\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \ (\mathrm{mod}\ D)}}\sum_{j=1}^{r-1} d_{ij}\binom{n+j-1}{j-1} + \sum_{\substack{M<n\leqq M+LD\\ n\equiv M \ (\mathrm{mod}\ D)}}\sum_{i=2}^{t}\left(\sum_{j=1}^{r_i} d_{ij}\binom{n+j-1}{j-1}\right)(s_i/s_1)^n\Bigg| \leqq$$

$$\leqq \sum_{j=1}^{r-1}|d_{1j}|\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \ (\mathrm{mod}\ D)}}\binom{n+r-2}{r-2} + \sum_{i=2}^{t}\sum_{j=1}^{r_i}|d_{ij}|\Bigg|\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \ (\mathrm{mod}\ D)}}\binom{n+j-1}{j-1}(s_i/s_1)^n\Bigg| =$$

$$= O\left\{L\binom{M+LD+r-2}{r-2} + \sum_{i=2}^{t}\sum_{j=1}^{r_i}\left|\frac{1}{D}\sum_{l=1}^{D} e^{-2\pi iMl/D}\cdot\sum_{n=M+D}^{M+LD}\binom{n+j-1}{j-1}\cdot(s_i e^{2\pi il/D}/s_1)^n\right|\right\} =$$

$$= O\left\{L\binom{M+LD+r-2}{r-2} + \sum_{i=2}^{t}\sum_{j=1}^{r_i}\sum_{l=1}^{D}\left|\sum_{n=M+D}^{M+LD}\binom{n+j-1}{j-1}(s_i e^{2\pi il/D}/s_1)^n\right|\right\}$$

(where the first term is 0 for $r=r_1=1$). In order to estimate the inner sum in the second term, we apply Lemma 3 with $z = s_i e^{2\pi il/D}/s_1$. Then $|z| \leqq 1$ and $z \neq 1$ hold by (17), (19), (20) and (22) thus we obtain from (23) that

$$\Bigg|\sum_{\substack{M<n\leqq M+LD\\ n\equiv M\ (\mathrm{mod}\ D)}} f(n)s_1^{-n}-d_{1r}\sum_{\substack{M<n\leqq M+LD\\ n\equiv M\ (\mathrm{mod}\ D)}}\binom{n+r-1}{r-1}\Bigg| =$$

$$=O\left\{L\binom{M+LD+r-2}{r-2}+\sum_{i=2}^{t}\sum_{j=1}^{r_i}\sum_{l=1}^{D}2^{j+1}\binom{M+LD+j}{j-1}\frac{|s_i/s_1|^{M+D}}{|1-s_ie^{2\pi il/D}/s_1|^j}\right\}=$$

$$=O\left\{L\binom{M+LD+r-2}{r-2}+\sum_{i=2}^{t}\binom{M+LD+r_i}{r_i-1}|s_i/s_1|^{M+D}\right\}=$$

$$=O\left\{L\frac{r-1}{M+LD+r-1}\binom{M+LD+r-1}{r-1}+\sum_{i=2}^{u}\binom{M+LD+r-1}{r-1}+\right.$$

$$\left.+\sum_{i=u+1}^{t}\binom{M+LD+r_i}{r_i-1}|s_{u+1}/s_1|^{M+D}\right\}=$$

$$=O\left\{L\frac{1}{LD}\binom{M+LD+r-1}{r-1}+\binom{M+LD+r-1}{r-1}+(M+LD+R)^{R-1}|s_{u+1}/s_1|^{M+D}\right\}=$$

$$=O\left(\binom{M+LD+r-1}{r-1}+(M+LD+R)^R|s_{u+1}/s_1|^M\right)$$

uniformly for $M, L=1, 2, 3, \ldots$ which proves the lemma.

LEMMA 5. *Let $f(n)$ be an arithmetic function of form* (18) *where* (17), (19) *and* (20) *hold and let D be a positive integer such that* (22) *holds. Then there exist positive constants C_3 and L_0 (which may depend on the parameters $r_1, \ldots, r_t, d_{11}, \ldots, d_{tr_t}, s_1, \ldots$ $\ldots s_t, D$) such that if L and M are positive integers satisfying $L>L_0$ and $M>M_0(L)$ then we have*

$$\max_{\substack{M<n\leqq M+LD\\ n\equiv M\ (\mathrm{mod}\ D)}}\left|\frac{f(n)}{\binom{n+r-1}{r-1}s_1^n}\right|>C_3. \tag{24}$$

PROOF. We are going to show that (24) holds with $C_3=\frac{|d_{1r}|}{2}$ (for sufficiently large L and $M>M_1(L)$). Assume indirectly that

$$\max_{\substack{M<n\leqq M+LD\\ n\equiv M\ (\mathrm{mod}\ D)}}\left|\frac{f(n)}{\binom{n+r-1}{r-1}s_1^n}\right|\leqq\frac{|d_{1r}|}{2}. \tag{25}$$

Then for all L and $M > M_2(L)$ we have

$$\left|\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \pmod D}} f(n)s_1^{-n} - d_{1r}\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \pmod D}}\binom{n+r-1}{r-1}\right| \geqq$$

$$\geqq |d_{1r}|\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \pmod D}}\binom{n+r-1}{r-1} - \sum_{\substack{M<n\leqq M+LD\\ n\equiv M \pmod D}}|f(n)s_1^{-n}| \geqq$$

$$\geqq |d_{1r}|\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \pmod D}}\binom{n+r-1}{r-1} - \sum_{\substack{M<n\leqq M+LD\\ n\equiv M \pmod D}}\frac{|d_{1r}|}{2}\binom{n+r-1}{r-1} = \tag{26}$$

$$= \frac{|d_{1r}|}{2}\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \pmod D}}\binom{n+r-1}{r-1} > \frac{|d_{1r}|}{2}\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \pmod D}}\frac{1}{2}\binom{M+LD+r-1}{r-1} =$$

$$= \frac{|d_{1r}|}{4}L\binom{M+LD+r-1}{r-1} \quad (\text{for } M > M_2(L))$$

since we have

$$\binom{n+r-1}{r-1} \sim \binom{M+LD+r-1}{r-1} \quad \text{for} \quad M < n \leqq M+LD,\ M \to +\infty.$$

On the other hand, applying Lemma 4, we obtain for $M > M_3(L)$ that

$$\left|\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \pmod D}} f(n)s_1^{-n} - d_{1r}\sum_{\substack{M<n\leqq M+LD\\ n\equiv M \pmod D}}\binom{n+r-1}{r-1}\right| < \tag{27}$$

$$< C_2\left(\binom{M+LD+r-1}{r-1} + (M+LD+R)^R |s_{u+1}/s_1|^M\right) < C_4\binom{M+LD+r-1}{r-1}$$

since, by (19), we have

$$(M+LD+R)^R |s_{u+1}/s_1|^M = o(1)$$

for fixed D, L and $M \to +\infty$.

(26) and (27) yield for all L and $M > M_4(L)$ that

$$\frac{|d_{1r}|}{4}L\binom{M+LD+r-1}{r-1} < C_4\binom{M+LD+r-1}{r-1}$$

or, in equivalent form,

$$L < \frac{4C_4}{|d_{1r}|}$$

but for sufficiently large L this inequality cannot hold. Thus in fact, the indirect assumption (25) leads to a contradiction for large L and $M > M_4(L)$ which proves Lemma 5.

Lemma 6. *Let $f(n)$ be an arithmetic function of form* (18) *where* (17), (19) *and* (20) *hold. If N is large enough then there exists an integer n for which*

$$[N/2] < n \leqq N, \tag{28}$$

$$|f(n)| > C_3 \binom{n+r-1}{r-1} |s_1|^n, \tag{29}$$

$$\text{if } p^\alpha | n \text{ then } p^\alpha \leqq \frac{N}{\log^2 N} \tag{30}$$

hold where C_3 denotes the (positive) constant defined in Lemma 5.

Proof. Let U denote the set of the integers n which satisfy (29). Applying Lemma 5 with $D=1$, we obtain that if L is large (but fixed) and $N > N_0(L)$ then

$$\begin{aligned} &\sum_{\substack{[N/2]<n\leqq N \\ n\in U}} 1 = \sum_{1\leqq j\leqq (N-[N/2])/L} \; \sum_{\substack{[N/2]+(j-1)L<n\leqq [N/2]+jL \\ n\in U}} 1 > \\ &> \sum_{1\leqq j\leqq (N-[N/2])/L} 1 = \left[\frac{N-[N/2]}{L}\right] \geqq \left[\frac{N/2}{L}\right] > \frac{N}{4L}. \end{aligned} \tag{31}$$

Furthermore, it is well-known that

$$\sum_{p^\alpha \leqq x} \frac{1}{p^\alpha} = \log\log x + C + o(1)$$

for some absolute constant C. Thus

$$\begin{aligned} &\sum_{N(\log N)^{-2}<p^\alpha\leqq N} \; \sum_{\substack{[N/2]<n\leqq N \\ p^\alpha | n}} 1 \leqq \\ &\leqq \sum_{N(\log N)^{-2}<p^\alpha\leqq N} \left[\frac{N}{p^\alpha}\right] \leqq N \sum_{N(\log N)^{-2}<p^\alpha\leqq N} \frac{1}{p^\alpha} = \\ &= N\big(\log\log N - \log\log (N(\log N)^{-2}) + o(1)\big) = N\cdot o(1) = o(N). \end{aligned} \tag{32}$$

(31) and (32) yield that the number of the integers n which satisfy (28), (29) and (30) is at least

$$\sum_{\substack{[N/2]<n\leqq N \\ n\in U}} 1 - \sum_{N(\log N)^{-2}<p^\alpha\leqq N} \; \sum_{\substack{[N/2]<n\leqq N \\ p^\alpha | n}} 1 > \frac{N}{4L} - \frac{N}{8L} = \frac{N}{8L} > 0$$

for large enough N and this completes the proof of Lemma 6.

(Note that so far we have not used the multiplicativity of $f(n)$.)

Lemma 7. *If a multiplicative arithmetic function $f(n)$ satisfies* (16) *and a linear recursion of form* (2) *then there exists a positive number p_0 such that if p is a prime number satisfying $p > p_0$ and α is a positive integer then we have*

$$f(p^\alpha) \neq 0.$$

PROOF. Assume indirectly that there exist infinitely many prime powers $p_1^{\alpha_1}, p_2^{\alpha_2}, \dots, p_k^{\alpha_k}, \dots$ such that $p_1<p_2<\dots<p_k$ and

$$f(p_1^{\alpha_1}) = f(p_2^{\alpha_2}) = \dots = f(p_k^{\alpha_k}) = \dots = 0. \tag{33}$$

Denote N the least positive integer satisfying the linear congruence system

$$\begin{aligned} N+1 &\equiv p_1^{\alpha_1} \pmod{p_1^{\alpha_1+1}} \\ N+2 &\equiv p_2^{\alpha_2} \pmod{p_2^{\alpha_2+1}} \\ &\dots\dots\dots\dots\dots\dots\dots \\ N+k &\equiv p_k^{\alpha_k} \pmod{p_k^{\alpha_k+1}}. \end{aligned}$$

Then for $i=1, 2, \dots, k$, we have $p_i^{\alpha_i} | N+i$ and $p_i^{\alpha_i+1} \nmid N+i$ thus $N+i$ can be written in the form $N+i=p_i^{\alpha_i} N_i$ where $(N_i, p_i)=1$. By (33), we have

$$f(N+i) = f(p_i^{\alpha_i} N_i) = f(p_i^{\alpha_i}) f(N_i) = 0 \cdot f(N_i) = 0 \quad \text{for} \quad i = 0, 1, \dots, k-1.$$

Thus with respect to (3), we obtain from (2) by straight induction that for $n=N+k+1, N+k+2, \dots,$

$$f(n) = -\sum_{i=1}^{k} \frac{a_{k-i}}{a_k} f(n-i) = -\sum_{i=1}^{k} \frac{a_{k-i}}{a_k} \cdot 0 = 0.$$

Define M by

$$f(M) \neq 0,\ f(M+1) = f(M+2) = \dots = f(M+i) = \dots = 0$$

(such an integer M exists by (16)). Then we have

$$\sum_{i=0}^{k} a_i f(M+i) = a_0 f(M) \neq 0$$

in contradiction with (2) and this contradiction completes the proof of Lemma 7.

LEMMA 8. *Let T be a positive integer and assume that the real numbers $y_1, y_2, \dots, y_T$ are linearly independent over the field of the rational numbers. Then the points*

$$(\{y_1\}, \{y_2\}, \dots, \{y_T\}),\ (\{2y_1\}, \{2y_2\}, \dots, \{2y_T\}), \dots, (\{ny_1\}, \{ny_2\}, \dots, \{ny_T\}), \dots$$

are uniformly distributed in the T-dimensional unit cube.

This lemma is well-known; see, e.g., [2].

LEMMA 9. *Let T be a positive integer, S a finite set of T-tuples of non-negative integers and assume that*

$$|S| \geqq 2. \tag{34}$$

For all $(j_1, j_2, \dots, j_T) \in S$, let $b_{j_1 j_2 \dots j_T}$ be a non-zero complex number. Put

$$F(x_1, x_2, \dots, x_T) = \sum_{(j_1, j_2, \dots, j_T) \in S} b_{j_1 j_2 \dots j_T} e^{2\pi i (j_1 x_1 + j_2 x_2 + \dots + j_T x_T)}$$

where $x_1, x_2, \dots, x_T$ are real variables. Then $|F(x_1, x_2, \dots, x_T)|$ assumes at least two distinct values, i.e., there exist real numbers $x_1', x_2', \dots, x_T', x_1'', x_2'', \dots, x_T''$ such that

$$|F(x_1', x_2', \dots, x_T')| \neq |F(x_1'', x_2'', \dots, x_T'')|.$$

PROOF. Assume indirectly that

$$|F(x_1, \dots, x_T)| = C$$

for all x. Then we have

$$\int_0^1 \dots \int_0^1 |F(x_1, \dots, x_T)|^2 \, dx_1 \dots dx_T = C^2 \tag{35}$$

and

$$\int_0^1 \dots \int_0^1 \big| |F(x_1, \dots, x_T)|^2 - C^2 \big|^2 \, dx_1 \dots dx_T = 0. \tag{36}$$

Let $(M_1, \dots, M_T)$ and $(L_1, \dots, L_T)$ denote the uniquely determined T-tuples for which $(M_1, \dots, M_T) \in S$, $(L_1, \dots, L_T) \in S$ hold, and for all $(j_1, j_2, \dots, j_T) \in S$, we have either $M_1 > j_1$ or $M_1 = j_1, \dots, M_{i-1} = j_{i-1}$, $M_i \neq j_i$ implies that $M_i > j_i$; similarly, either $L_1 < j_1$ or $L_1 = j_1, \dots, L_{i-1} = j_{i-1}, L_i \neq j_i$ implies that $L_i < j_i$. Using Parseval's formula, we obtain with respect to (34) and (35) that

$$\int_0^1 \dots \int_0^1 \big| |F(x_1, \dots, x_T)|^2 - C^2 \big|^2 \, dx_1 \dots dx_T =$$

$$= \int_0^1 \dots \int_0^1 |F(x_1, \dots, x_T) \overline{F(x_1, \dots, x_T)} - C^2|^2 \, dx_1 \dots dx_T =$$

$$= \int_0^1 \dots \int_0^1 \Big| \Big(\sum_{(j_1, \dots, j_T) \in S} b_{j_1 \dots j_T} e^{2\pi i (j_1 x_1 + \dots + j_T x_T)} \Big) \times$$

$$\times \Big(\sum_{(l_1, \dots, l_T) \in S} \overline{b_{l_1 \dots l_T}} e^{-2\pi i (l_1 x_1 + \dots + l_T x_T)} \Big) - C^2 \Big|^2 \, dx_1 \dots dx_T =$$

$$= \int_0^1 \dots \int_0^1 \Big| \Big(\sum_{(j_1, \dots, j_T) \in S} |b_{j_1 \dots j_T}|^2 - C^2 \Big) +$$

$$+ \sum_{(s_1, \dots, s_T) \neq (0, \dots, 0)} \Big(\sum_{j_1 - l_1 = s_1, \dots, j_T - l_T = s_T} b_{j_1 \dots j_T} \overline{b_{l_1 \dots l_T}} \Big) \times$$

$$\times e^{2\pi i (s_1 x_1 + \dots + s_T x_T)} \Big|^2 \, dx_1 \dots dx_T =$$

$$= \int_0^1 \dots \int_0^1 \Big| \Big(\int_0^1 \dots \int_0^1 |F(y_1, \dots, y_T)|^2 \, dy_1 \dots dy_T - C^2 \Big) +$$

$$+ \sum_{(s_1, \dots, s_T) \neq (0, \dots, 0)} \Big(\sum_{j_1 - l_1 = s_1, \dots, j_T - l_T = s_T} b_{j_1 \dots j_T} \overline{b_{l_1 \dots l_T}} \Big) \times$$

$$\times e^{2\pi i (s_1 x_1 + \dots + s_T x_T)} \Big|^2 \, dx_1 \dots dx_T =$$

$$= \int_0^1 \dots \int_0^1 \Big| \sum_{(s_1, \dots, s_T) \neq (0, \dots, 0)} \Big(\sum_{j_1 - l_1 = s_1, \dots, j_T - l_T = s_T} b_{j_1 \dots j_T} \overline{b_{l_1 \dots l_T}} \Big) \times$$

$$\times e^{2\pi i (s_1 x_1 + \dots + s_T x_T)} \Big|^2 \, dx_1 \dots dx_T =$$

$$= \sum_{(s_1, \dots, s_T) \neq (0, \dots, 0)} \Big| \sum_{j_1 - l_1 = s_1, \dots, j_T - l_T = s_T} b_{j_1 \dots j_T} \overline{b_{l_1 \dots l_T}} \Big|^2 \geqq$$

$$\geqq \Big| \sum_{j_1 - l_1 = M_1 - L_1, \dots, j_T - l_T = M_T - L_T} b_{j_1 \dots j_T} \overline{b_{l_1 \dots l_T}} \Big|^2 =$$

$$= |b_{M_1 \dots M_T} \overline{b_{L_1 \dots L_T}}|^2 > 0$$

(since $b_{M_1\dots M_T}\neq 0$ and $b_{L_1\dots L_T}\neq 0$) in contradiction with (36) and this contradiction proves Lemma 9.

4. Using Lemmas 2, 5, 6 and 7, we prove in this section

LEMMA 10. *Let* $f(n)$ *be a multiplicative arithmetic function of form* (18) *where* (17), (19) *and* (20) *hold. Then we have*

$$|s_1| = 1. \tag{37}$$

PROOF. First we prove that

$$|s_1| \leqq 1. \tag{38}$$

Assume indirectly that

$$|s_1| > 1. \tag{39}$$

For $N>N_0$, Lemma 6 yields that there exists a positive integer $n=p_1^{\alpha_1}p_2^{\alpha_2}\dots p_l^{\alpha_l}$ (where $p_1^{\alpha_1}<p_2^{\alpha_2}<\dots<p_l^{\alpha_l}$) for which (28), (29) and (30) hold. Then by (30), (39) and Lemma 2, we have

$$|f(n)| = \left|f\left(\prod_{i=1}^{l} p_i^{\alpha_i}\right)\right| = \prod_{i=1}^{l} |f(p_i^{\alpha_i})| <$$

$$< \prod_{i=1}^{l}\left(C_1\binom{p_i^{\alpha_i}+r-1}{r-1}|s_1|^{p_i^{\alpha_i}}\right) \leqq \prod_{i=1}^{l} C_1(rp_i^{\alpha_i})^r|s_1|^{N(\log N)^{-2}} \leqq \tag{40}$$

$$\leqq (C_1r^r)^l n^r|s_1|^{lN(\log N)^{-2}} \leqq (C_5)^{\frac{\log N}{\log 2}} N^r|s_1|^{\frac{\log N}{\log 2}N(\log N)^{-2}} = N^{C_6}|s_1|^{2N(\log N)^{-1}}$$

since obviously,

$$l = \sum_{p|n} 1 \leqq \frac{\log n}{\log 2} \leqq \frac{\log N}{\log 2}.$$

On the other hand, (28) and (29) yield with respect to (39) for large N that

$$|f(n)| \geqq C_3\binom{n+r-1}{r-1}|s_1|^n > C_3|s_1|^{N/2}. \tag{41}$$

(40) and (41) yield that

$$N^{C_6}|s_1|^{2N(\log N)^{-1}} \geqq C_3|s_1|^{N/2}$$

hence for large N,

$$\frac{1}{C_3}N^{C_6} \geqq |s_1|^{N/2-2N(\log N)^{-1}} > |s_1|^{N/3}.$$

But by (39), this inequality cannot hold for large N. Thus (39) leads to a contradiction which proves (38).

Now we are going to prove that

$$|s_1| \geqq 1. \tag{42}$$

Assume indirectly that

$$|s_1| < 1. \tag{43}$$

By (17), we have $s_2/s_1 \neq 1, s_3/s_1 \neq 1, \dots, s_t/s_1 \neq 1$. Thus there exists a positive number p_1 such that if p is a prime number satisfying $p > p_1$ then

$$(44) \qquad (s_i/s_1)^p \neq 1 \quad \text{for} \quad i = 2, 3, \dots, t.$$

Furthermore, by Lemma 7, we have

$$(45) \qquad f(p) \neq 0$$

for $p > p_0$. Let us fix a prime number satisfying $p > \max\{p_0, p_1\}$; then both (44) and (45) hold.

By (44), we may apply Lemma 5 with $D = p$ and $M \equiv 1 \pmod{p}$. Then Lemma 5 yields that there exist infinitely many integers m such that

$$|f(mp+1)| \geqq C_3 \binom{mp+1+r-1}{r-1} |s_1|^{mp+1} \geqq C_3 |s_1|^{mp+1}$$

hence

$$(46) \qquad |f(mp^2+p)| = \left|f\big(p(mp+1)\big)\right| = |f(p)|\,|f(mp+1)| \geqq C_3 |f(p)|\,|s_1|^{mp+1}$$

since $(p, mp+1) = 1$. On the other hand, by using Lemma 2, we obtain for all m that

$$(47) \qquad \begin{aligned} |f(mp^2+p)| &< C_1 \binom{mp^2+p+r-1}{r-1} |s_1|^{mp^2+p} \leqq \\ &\leqq C_1 (mp^2+p+r-1)^{r-1} |s_1|^{mp^2+p}. \end{aligned}$$

(46) and (47) yield that for infinitely many m we have

$$C_3 |f(p)|\,|s_1|^{mp+1} \leqq C_1 (mp^2+p+r-1)^{r-1} |s_1|^{mp^2+p}$$

hence with respect to (43),

$$C_7 |f(p)| \leqq m^{C_8} |s_1|^{(p-1)(mp+1)} \leqq m^{C_8} |s_1|^{mp+1} \leqq m^{C_8} |s_1|^m$$

where $C_7 > 0$, and C_7, C_8 may depend on $p, r_1, \dots, r_t, d_{11}, \dots, d_{tr_t}, s_1, \dots, s_t$ but these numbers are independent of m. But by (43) and (45), this inequality cannot hold for large m, and this contradiction proves (42).

(38) and (42) yield (37) and this completes the proof of Lemma 10.

5. We have to show that if $f(n)$ is a multiplicative function which satisfies (16) and a linear recursion of form (2) then there exist h, m and χ satisfying the conditions in Theorem 1. We are going to prove that $h = r-1$ can be chosen where $r = r_1$ is defined in Lemma 1 (while m and χ will be defined later). In order to show this, we prove two lemmas concerning the function

$$(48) \qquad g(n) = f(n)/n^h = f(n)/n^{r-1}$$

in this section. This function is multiplicative since it is the product of the multiplicative functions $f(n)$ and n^{-h}.

LEMMA 11. *Assume that $f(n)$ is a multiplicative arithmetic function which satisfies* (16) *and a linear recursion of form* (2). *Define the multiplicative function $g(n)$ by* (48) *where $r = r_1$ is defined in Lemma* 1. *For $j = 1, 2, \dots$, let R_j denote the set* *of the positive integers n for which* $\left(n, \prod_{p \leqq j} p\right)=1$ *holds. Then for all* $\varepsilon>0$, *there exists a number* $j_0=j_0(\varepsilon)$ *such that* $j>j_0$, $n \in R_j$ *imply that*

$$1-\varepsilon<|g(n)|<1+\varepsilon \quad (\textit{for } j>j_0,\ n \in R_j). \tag{49}$$

PROOF. By Lemmas 2 and 10, we have

$$\begin{aligned}|g(n)| &= |f(n)|n^{-r+1} < C_1\binom{n+r-1}{r-1}|s_1|^n n^{-r+1} = \\ &= C_1\binom{n+r-1}{r-1} n^{-r+1} < C_1(n+r-1)^{r-1} n^{-r+1} = \\ &= C_1\left(1+\frac{r-1}{n}\right)^{r-1} \leqq C_1\big(1+(r-1)\big)^{r-1} = C_9.\end{aligned} \tag{50}$$

On the other hand, let us apply Lemma 5 with $D=1$ (then (22) holds by (17)). We obtain that for $L>L_0$ and $M>M_0(L)$ we have

$$\max_{M<n \leqq M+L}\left|\frac{f(n)}{\binom{n+r-1}{r-1}s_1^n}\right| > C_3. \tag{51}$$

By Lemma 10, we have

$$\left|\frac{f(n)}{\binom{n+r-1}{r-1}s_1^n}\right| = \frac{|f(n)|}{\binom{n+r-1}{r-1}} < \frac{|f(n)|(r-1)!}{n^{r-1}} = (r-1)!\,|g(n)|. \tag{52}$$

(51) and (52) yield that for $L=[L_0]+1$ and $M>M_0(L)$ we have

$$\max_{M<n \leqq M+L}|g(n)| > \frac{C_3}{(r-1)!} = C_{10} \quad (\text{for } M>M_0(L)) \tag{53}$$

where $C_{10}>0$.

In order to prove (49), let us assume indirectly that either for a fixed $\varepsilon>0$ and for $j=1, 2, \ldots$ there exists a positive integer n_j such that

$$|g(n_j)| \geqq 1+\varepsilon \quad \text{and} \quad n_j \in R_j \quad (\text{for } j=1, 2, \ldots) \tag{54}$$

or for a fixed $(1>)\varepsilon>0$ and for $j=1, 2, \ldots$ there exists a positive integer u_j such that

$$|g(u_j)| \leqq 1-\varepsilon \quad \text{and} \quad u_j \in R_j \quad (\text{for } j=1, 2, \ldots). \tag{55}$$

Note that (54) and (55) imply that $n_j>1$ and $u_j>1$ since we have $g(1)=1$ by the multiplicativity of $g(n)$ ($g(n) \not\equiv 0$ by (53)).

Assume first that (54) holds and for each j, fix a positive integer n_j satisfying (54). Define the positive integers $N_1<N_2<\ldots$ by the following recursion: let $N_1=n_1$ and for $i \geqq 2$, let $N_i=n_{N_{i-1}}$. Then we have $N_i \in R_{N_{i-1}}$ thus $\left(N_i, \prod_{p \leqq N_{i-1}} p\right)=1$, hence $(N_i, N_j)=1$ for all $j<i$. Thus (54) yields for all x that

$$|g(N_1 N_2 \ldots N_x)| = |g(N_1)g(N_2)\ldots g(N_x)| \geqq (1+\varepsilon)^x. \tag{56}$$

On the other hand, by (49) we have

$$|g(N_1 N_2 \dots N_x)| < C_9. \tag{57}$$

(56) and (57) yield that

$$(1+\varepsilon)^x < C_9$$

but, for large x, this inequality cannot hold and this contradiction proves that (54) cannot hold.

Assume now that (55) holds. Again, for each j fix a positive integer u_j satisfying (55) and define the positive integers $U_1 < U_2 < \dots$ by the following recursion: let $U_1 = u_1$ and for $i \geqq 2$, let $U_i = u_{U_{i-1}}$. Then we have $U_i \in R_{U_{i-1}}$ thus again, $(U_i, U_j) = 1$ for all $j < i$. For a positive integer x, let us write

$$V_i = \prod_{j=(i-1)x+1}^{ix} U_j \quad \text{for} \quad i = 1, 2, \dots, L$$

where $L(=[L_0]+1)$ has been defined above (see (53)) and define the positive integer M by

$$M + i \equiv V_i \ (\operatorname{mod} V_i^2) \quad \text{for} \quad i = 1, 2, \dots, L \tag{58}$$

and

$$M_0(L) < M \leqq M_0(L) + \prod_{i=1}^{L} V_i^2.$$

Then by (53), there exists a positive integer i such that $1 \leqq i \leqq L$ and

$$|g(M+i)| \geqq C_{10} \quad (>0). \tag{59}$$

Then by (57), there exists a non-negative integer t such that $M+i = tV_i^2 + V_i$ thus with respect to (50) and (55), we have

$$|g(M+i)| = |g(tV_i^2+V_i)| = \left|g\bigl(V_i(tV_i+1)\bigr)\right| =$$

$$= |g(V_i)g(tV_i+1)| = |g(V_i)|\,|g(tV_i+1)| =$$

$$= \left|g\Bigl(\prod_{j=(i-1)x+1}^{ix} U_j\Bigr)\right| |g(tV_i+1)| = \Bigl(\prod_{j=(i-1)x+1}^{ix} |g(U_j)|\Bigr) |g(tV_i+1)| < \tag{60}$$

$$< \Bigl(\prod_{j=(i-1)x+1}^{ix} (1-\varepsilon)\Bigr) C_9 = C_9(1-\varepsilon)^x$$

since $(V_i, tV_i+1) = 1$ and $(U_i, U_j) = 1$ for $i \neq j$. (59) and (60) yield that

$$(0<)\ C_{10} < C_9(1-\varepsilon)^x$$

but for large x, this inequality cannot hold and this contradiction proves that (55) cannot hold which completes the proof of (49).

LEMMA 12. *Assume that $f(n)$ is a multiplicative arithmetic function which satisfies* (16) *and a linear recursion of form* (2). *Then the multiplicative arithmetic function*

$g(n)$ defined by (48) *can be written in the form*

$$g(n) = f(n)/n^h = f(n)/n^{r-1} = \varphi(n) + \psi(n) + \varrho(n) \tag{61}$$

where

(i) *the function $\varphi(n)$ is of the form*

$$\varphi(n) = \sum_{j=1}^{v} d_j s_j^n$$

where $d_1, d_2, \dots, d_v$ are non-zero complex numbers and $s_1, s_2, \dots, s_v$ are roots of unity, i.e., there exist positive integers $M_1, M_2, \dots, M_v$ such that

$$s_j^{M_j} = 1 \quad \textit{for} \quad j = 1, 2, \dots, v; \tag{62}$$

furthermore, we have

$$s_i \neq s_j \quad \textit{for} \quad 1 \leqq i < j \leqq v \tag{63}$$

and

$$\varphi(n) \not\equiv 0; \tag{64}$$

(ii) *The function $\psi(n)$ is of the form*

$$\psi(n) = \sum_{j=1}^{w} b_j t_j^n \tag{65}$$

where $b_1, \dots, b_w$ are non-zero complex numbers and

$$t_i \neq t_j \quad \textit{for} \quad 1 \leqq i < j \leqq w,$$

$$|t_1| = |t_2| = \dots = |t_w| = 1, \tag{66}$$

but $t_1, t_2, \dots, t_w$ are not roots of unity, i.e.,

$$t_j^n \neq 1 \quad \textit{for} \quad j = 1, 2, \dots, w \quad \textit{and} \quad n = 1, 2, \dots \tag{67}$$

(also $\psi(n) \equiv 0$ may occur);

(iii) *We have*

$$\varrho(n) = O\left(\frac{1}{n}\right)$$

(also $\varrho(n) \equiv 0$ may occur).

PROOF. If $f(n)$ is multiplicative, and it satisfies (16) and a linear recursion of form (2) then by Lemmas 1 and 10, $f(n)$ can be written in the form

$$f(n) = \sum_{i=1}^{t} \left(\sum_{j=1}^{r_i} d_{ij} \binom{n+j-1}{j-1} \right) s_i^n \tag{68}$$

where (17) holds and we have

$$1 = |s_1| = |s_2| = \dots = |s_u| > |s_{u+1}| \geqq |s_{u+2}| \geqq \dots \geqq |s_t|$$

and

$$r_1 \geqq r_2 \geqq \dots \geqq r_u.$$

We may assume that

$$r_1 = r_2 = \ldots = r_x > r_{x+1} \geqq \ldots \geqq r_u,$$

furthermore, the numbers $s_1, s_2, \ldots, s_v$ are roots of unity but the numbers $s_{v+1}, \ldots, s_x$ are not (it may occur that $v=0$, i.e., none of the s_i''s is root of unity). Then putting $h=r-1$ and dividing (68) by n^h, we obtain for $n \to +\infty$ that

$$g(n) = \frac{f(n)}{n^h} = \frac{f(n)}{n^{r-1}} =$$

$$= \sum_{i=1}^{x} d_{ir_i} n^{-(r-1)} \binom{n+r_i-1}{r_i-1} s_i^n + \sum_{i=1}^{x} \left(\sum_{j=1}^{r_i-1} d_{ij} n^{-(r-1)} \binom{n+j-1}{j-1} \right) s_i^n +$$

$$+ \sum_{i=x+1}^{u} \left(\sum_{j=1}^{r_i} d_{ij} n^{-(r-1)} \binom{n+j-1}{j-1} \right) s_i^n + \sum_{i=u+1}^{t} \left(\sum_{j=1}^{r_i} d_{ij} n^{-(r-1)} \binom{n+j-1}{j-1} \right) s_i^n =$$

$$= \sum_{i=1}^{x} d_{ir_i} n^{-(r-1)} \binom{n+r-1}{r-1} s_i^n + \sum_{i=1}^{x} \sum_{j=1}^{r-1} O(n^{-(r-1)} n^{j-1}) +$$

(69) $$+ \sum_{i=x+1}^{u} \left(\sum_{j=1}^{r_i} O(n^{-(r-1)} n^{j-1}) \right) + \sum_{i=u+1}^{t} \sum_{j=1}^{r_i} O(n^{j-1} |s_i|^n) =$$

$$= \sum_{i=1}^{x} d_{ir_i} n^{-(r-1)} \left(\frac{n^{r-1}}{(r-1)!} + O(n^{r-2}) \right) s_i^n + O(n^{-(r-1)} n^{(r-1)-1}) +$$

$$+ \sum_{i=x+1}^{u} O(n^{-(r-1)} n^{r_i-1}) + \sum_{i=u+1}^{t} O(n^{r_i-1} |s_i|^n) =$$

$$= \sum_{i=1}^{x} \frac{d_{ir_i}}{(r-1)!} s_i^n + O\left(\frac{1}{n}\right) + O\left(\frac{1}{n}\right) + \sum_{i=x+1}^{u} O(n^{-(r-1)} n^{(r-1)-1}) +$$

$$+ \frac{1}{n} \sum_{i=u+1}^{t} O(n^{r_i} |s_i|^n) =$$

$$= \sum_{i=1}^{x} \frac{d_{ir_i}}{(r-1)!} s_i^n + O\left(\frac{1}{n}\right) + \frac{1}{n} \cdot o(1) =$$

$$= \sum_{i=1}^{v} \frac{d_{ir_i}}{(r-1)!} s_i^n + \sum_{i=v+1}^{x} \frac{d_{ir_i}}{(r-1)!} s_i^n + O\left(\frac{1}{n}\right) = \varphi(n) + \psi(n) + O\left(\frac{1}{n}\right)$$

where

$$\varphi(n) = \sum_{i=1}^{v} \frac{d_{ir_i}}{(r-1)!} s_i^n \quad \text{and} \quad \psi(n) = \sum_{i=v+1}^{x} \frac{d_{ir_i}}{(r-1)!} s_i^n.$$

Define $\varrho(n)$ by

$$\varrho(n) = g(n) - \varphi(n) - \psi(n).$$

Then with respect to (69), the arithmetic functions $\varphi(n)$, $\psi(n)$ and $\varrho(n)$ satisfy all the conditions in Lemma 12, except at most (64). Thus in order to complete the proof of Lemma 12, we have to show that also (64) holds.

Assume indirectly that

$$g(n) = \frac{f(n)}{n^h} = \psi(n)+O\left(\frac{1}{n}\right) = \sum_{j=1}^{w} b_j t_j^n + O\left(\frac{1}{n}\right) \tag{70}$$

where $b_1, \dots, b_w$ are non-zero complex numbers and $t_1, t_2, \dots, t_w$ satisfy all the conditions in (ii). For $j=1, 2, \dots, w$, write t_j in the form

$$t_j = e^{2\pi i\theta_j}$$

where θ_j is a real number.

Assume first that $w=1$, i.e., $g(n)$ can be written in the form

$$g(n) = b_1 t_1^n + O\left(\frac{1}{n}\right) = b_1 e^{2\pi i n\theta_1} + O\left(\frac{1}{n}\right)$$

where $b_1 \neq 0$ and by (67), θ_1 is an irrational number. Put $b_1 = Ae^{2\pi i\alpha}$ where $A>0$ and α are real numbers. Then we have

$$g(n) = Ae^{2\pi i(n\theta_1+\alpha)} + O\left(\frac{1}{n}\right)$$

thus if ε is a small positive number (which will be fixed later) then for $n>n_0(\varepsilon)$ we have

$$|g(n) - Ae^{2\pi i(n\theta_1+\alpha)}| < \varepsilon \quad (\text{for } n > n_0(\varepsilon)). \tag{71}$$

Obviously, there exists a positive number $\delta=\delta(\varepsilon)$ such that if $|x|<\delta$ then

$$|1-e^{2\pi i x}| < \varepsilon \quad (\text{for } |x| < \delta). \tag{72}$$

Let $a=[n_0(\varepsilon)]+2$ (where $n_0(\varepsilon)$ is defined by (70)). (70) yields that

$$|g(a) - Ae^{2\pi i(a\theta_1+\alpha)}| < \varepsilon. \tag{73}$$

Write $a(a-1)\theta_1=\beta$. Then β is irrational number thus by Lemma 8 (applying it with $T=1$), the numbers $\{\beta\}, \{2\beta\}, \dots, \{j\beta\}, \dots$ are uniformly distributed in the interval $[0, 1)$. Thus there exists a positive integer j for which

$$|j\beta - \theta_1 - \alpha - 1/2| < \delta. \tag{74}$$

Let us write $b=aj+1$. Then, obviously, $(a, b)=1$ thus

$$g(ab) = g(a)g(b) \tag{75}$$

and $b>a>n_0(\varepsilon)$ thus by (71),

$$|g(b) - Ae^{2\pi i(b\theta_1+\alpha)}| < \varepsilon \tag{76}$$

and

$$|g(ab) - Ae^{2\pi i(ab\theta_1+\alpha)}| < \varepsilon. \tag{77}$$

(73), (76) and (77) yield with respect to (72) and (74) that

$$\begin{aligned}
|g(ab)-g(a)g(b)| &= \Big|\big\{Ae^{2\pi i(ab\theta_1+\alpha)}+\big(g(ab)-Ae^{2\pi i(ab\theta_1+\alpha)}\big)\big\}-\\
&\quad -\big\{A\varepsilon^{2\pi i(a\theta_1+\alpha)}\cdot Ae^{2\pi i(b\theta_1+\alpha)}+g(a)\big(g(b)-Ae^{2\pi i(b\theta_1+\alpha)}\big)+\\
&\quad +Ae^{2\pi i(b\theta_1+\alpha)}\big(g(a)-Ae^{2\pi i(a\theta_1+\alpha)}\big)\big\}\Big| \geqq\\
&\geqq |Ae^{2\pi i(ab\theta_1+\alpha)}-A^2e^{2\pi i((a+b)\theta_1+2\alpha)}|-|g(ab)-Ae^{2\pi i(ab\theta_1+\alpha)}|-\\
&\quad -|g(a)|\,|g(b)-Ae^{2\pi i(b\theta_1+\alpha)}|-A\,|g(a)-Ae^{2\pi i(a\theta_1+\alpha)}| >\\
&> A\,|e^{2\pi i((a+b)\theta_1+2\alpha)}(e^{2\pi i((ab-a-b)\theta_1-\alpha)}-A)|-\varepsilon-|g(a)|\varepsilon-A\varepsilon =\\
&= A\,\Big|-e^{2\pi i\left\{(a(aj+1)-a-(aj+1))\theta_1-\alpha-\frac{1}{2}\right\}}-A\Big|-\\
&\quad -\varepsilon-\varepsilon\Big|Ae^{2\pi i(a\theta_1+\alpha)}+\big(g(a)-Ae^{2\pi i(a\theta_1+\alpha)}\big)\Big|-\varepsilon A \geqq\\
&\geqq A\,\Big|e^{2\pi i\left(j(a^2-a)\theta_1-\theta_1-\alpha-\frac{1}{2}\right)}+A\Big|-\varepsilon-\varepsilon A-\varepsilon\Big|g(a)-Ae^{2\pi i(a\theta_1+\alpha)}\Big|-\varepsilon A \geqq\\
&\geqq A\,\Big|e^{2\pi i\left(j\beta-\theta_1-\alpha-\frac{1}{2}\right)}+A\Big|-\varepsilon-\varepsilon A-\varepsilon^2-\varepsilon A \geqq\\
&\geqq A(1+A)-A\,\Big|e^{2\pi i\left(j\beta-\theta_1-\alpha-\frac{1}{2}\right)}-1\Big|-\varepsilon-2\varepsilon A-\varepsilon^2 \geqq\\
&\geqq A(1+A)-\varepsilon A-\varepsilon-2\varepsilon A-\varepsilon^2 = A(1+A-3\varepsilon)-\varepsilon-\varepsilon^2
\end{aligned}$$

but if ε is sufficiently small (in terms of A_1) then this lower bound is positive in contradiction with (75), and this contradiction proves that we cannot have $w=1$ in (70).

Assume now that $w\geqq 2$ and denote T the number of the elements of a maximal linearly independent subset of $\theta_1, \theta_2, \dots, \theta_w$ over the field of the rational numbers. We may assume that $\theta_1, \theta_2, \dots, \theta_T$ form such a maximal linearly independent subset of $\theta_1, \theta_2, \dots, \theta_w$. Then there exist integers p_{ij} and positive integers q_{ij} (where $i=1, 2, \dots, w-T, j=1, 2, \dots, T$) such that

$$\theta_{T+i}=\sum_{j=1}^{T}\frac{p_{ij}}{q_{ij}}\theta_j \quad \text{for} \quad i=1, 2, \dots, w-T.$$

Let us write $Q=[q_{11}, q_{12}, \dots, q_{1T}, q_{21}, \dots, q_{w-T,T}]$. Then writing $\Phi_j=\theta_j/Q$ for $j=1, 2, \dots, T$, we obtain that there exist integers n_{ij} (where $i=1, 2, \dots, w-T$, $j=1, 2, \dots, T$) such that

$$\theta_{T+i}=\sum_{j=1}^{T}n_{ij}\Phi_j \quad \text{for} \quad i=1, 2, \dots, w-T.$$

Putting this into (65), we obtain that

$$\begin{aligned}
\psi(n) &= \sum_{j=1}^{T}b_je^{2\pi iQ\Phi_jn}+\sum_{j=1}^{w-T}b_{T+j}e^{2\pi i(n_{j1}\Phi_1+n_{j2}\Phi_2+\dots+n_{jT}\Phi_T)n} =\\
&= F(\Phi_1n, \Phi_2n, \dots, \Phi_Tn)
\end{aligned} \tag{78}$$

where the function $F(x_1, x_2, \ldots, x_T)$ is defined by

$$F(x_1, x_2, \ldots, x_T) = \sum_{j=1}^{T} b_j e^{2\pi i Q x_j} + \sum_{j=1}^{w-T} b_{T+j} e^{2\pi i (n_{j1} x_1 + n_{j2} x_2 + \cdots + n_{jT} x_T)}.$$

There are $w \geqq 2$ non-zero coefficients in this trigonometric polynomial thus we may apply Lemma 9. (Note that by $t_i \neq t_j$ the T-tuples $(n_{j1}, n_{j2}, \ldots, n_{jT})$ in the exponents are different.) We obtain that there exist T-tuples $(x_1', x_2', \ldots, x_T')$, $(x_1'', x_2'', \ldots, x_T'')$ such that

$$|F(x_1', \ldots, x_T')| - |F(x_1'', \ldots, x_T'')| = \eta > 0. \tag{79}$$

The function $F(x_1, x_2, \ldots, x_T)$ is continuous and periodic in each variable with period 1 thus there exists a positive number δ such that we have

$$|F(x_1, \ldots, x_T) - F(x_1', \ldots, x_T')| < \frac{\eta}{4} \quad \text{for} \quad |x_1 - x_1'| < \delta, \ldots, |x_T - x_T'| < \delta \tag{80}$$

and

$$|F(x_1, \ldots, x_T) - F(x_1'', \ldots, x_T'')| < \frac{\eta}{4} \quad \text{for} \quad |x_1 - x_1''| < \delta, \ldots, |x_T - x_T''| < \delta. \tag{81}$$

Applying Lemma 11 with $\varepsilon = \eta/9$, we obtain that for $j > j_1 = j_1(\eta)$, $n \in R_j$ implies that

$$1 - \frac{\eta}{9} < |g(n)| < 1 + \frac{\eta}{9} \quad (\text{for } j > j_1,\ n \in R_j). \tag{82}$$

By (70), we have

$$\psi(n) = g(n) + O\left(\frac{1}{n}\right). \tag{83}$$

$n \in R_j, n \neq 1$ imply that $n > j$. Thus for $j > j_1$, (82) and (83) yield that

$$1 - \frac{\eta}{8} < |\psi(n)| < 1 + \frac{\eta}{8} \tag{84}$$

for

$$j > j_1, \quad n \in R_j, \quad n \neq 1. \tag{85}$$

Let us write $J = [j_1] + 1$ and $S = \prod_{p \leqq J} p$. The numbers $\theta_1, \theta_2, \ldots, \theta_T$ are linearly independent over the field of the rational numbers and thus also the numbers $S\Phi_1 = S\frac{\theta_1}{Q}, \ldots, S\Phi_T = S\frac{\theta_T}{Q}$ are linearly independent. Then by Lemma 8, the points

$$(\{S\Phi_1\}, \{S\Phi_2\}, \ldots, \{S\Phi_T\}), (\{2S\Phi_1\}, \{2S\Phi_2\}, \ldots, \{2S\Phi_T\}), \ldots,$$
$$(\{lS\Phi_1\}, \{lS\Phi_2\}, \ldots, \{lS\Phi_T\}), \ldots$$

are uniformly distributed in the T-dimensional unit cube. Thus there exist positive integers l_1, l_2 such that

$$|\{l_1 S\Phi_j\}-\{x_j'-\Phi_j\}|<\delta \quad \text{for} \quad j=1,2,\dots,T$$

and

$$|\{l_2 S\Phi_j\}-\{x_j''-\Phi_j\}|<\delta \quad \text{for} \quad j=1,2,\dots,T$$

or in equivalent form,

(86) $|(l_1S+1)\Phi_j-x_j'-u_j|<\delta$ for an integer u_j and for $j=1,2,\dots,T$

and

(87) $|(l_2S+1)\Phi_j-x_j''-v_j|<\delta$ for an integer v_j and for $j=1,2,\dots,T$.

The function $F(x_1, x_2, \dots, x_T)$ is periodic in each variable with period 1 thus (78), (80), (81), (86) and (87) yield that

$$\begin{aligned}
&|\psi(l_1S+1)|=\big|F(\Phi_1(l_1S+1),\dots,\Phi_T(l_1S+1))\big|=\\
&=\big|F(x_1'+((l_1S+1)\Phi_1-x_1'),\dots,x_T'+((l_1S+1)\Phi_T-x_T'))\big|=\\
(88)\quad &=\big|F(x_1'+((l_1S+1)\Phi_1-x_1'-u_1),\dots,x_T'+((l_1S+1)\Phi_T-x_T'-u_T))\big|>\\
&>|F(x_1',\dots,x_T')|-\frac{\eta}{4}
\end{aligned}$$

and similarly,

$$\begin{aligned}
&|\psi(l_2S+1)|=\big|F(\Phi_1(l_2S+1),\dots,\Phi_T(l_2S+1))\big|=\\
(89)\quad &=\big|F(x_1''+((l_2S+1)\Phi_1-x_1''-v_1),\dots,x_T''+((l_2S+1)\Phi_T-x_T''-v_T))\big|<\\
&<|F(x_1'',\dots,x_T'')|+\frac{\eta}{4}.
\end{aligned}$$

(79), (88) and (89) yield that

$$\begin{aligned}
&|\psi(l_1S+1)|-|\psi(l_2S+1)|>\left(|F(x_1',\dots,x_T')|-\frac{\eta}{4}\right)-\left(|F(x_1'',\dots,x_T'')|+\frac{\eta}{4}\right)=\\
(90)\quad &=|F(x_1',\dots,x_T')|-|F(x_1'',\dots,x_T'')|-\frac{\eta}{2}=\frac{\eta}{2}.
\end{aligned}$$

On the other hand, we have $(S, l_1S+1)=1$ and $(S, l_2S+1)=1$ where $S=\prod_{p\leqq J}p$, thus $l_1S+1\in R_J$ and $l_2S+1\in R_J$ where $J>j_1$; furthermore, $l_1S+1>1$ and $l_2S+1>1$. Thus all the conditions in (85) hold with $j=J, n=l_1S+1$, and $j=J, n=l_2S+1$, respectively. Then (84) yields that

$$|\psi(l_1S+1)|-|\psi(l_2S+1)|<\left(1+\frac{\eta}{8}\right)-\left(1-\frac{\eta}{8}\right)=\frac{\eta}{4}$$

in contradiction with (90). Thus the indirect assumption (70) leads to a contradiction also in the case $w\geqq 2$ which completes the proof of (64) and thus also of Lemma 12.

6. In this section, we prove (using Lemma 12)

LEMMA 13. *Assume that $f(n)$ is a multiplicative arithmetic function which satisfies* (6) *and a linear recursion of form* (2). *Then the multiplicative arithmetic function $g(n)$ defined by* (48) *is periodic, i.e., there exists a positive integer M such that we have*

$$g(n+M) = g(n) \quad \textit{for} \quad n = 1, 2, \dots . \tag{91}$$

PROOF. By Lemma 12, $g(n)$ can be written in the form (61) where $\varphi(n)$, $\psi(n)$ and $\varrho(n)$ satisfy all the conditions (i), (ii) and (iii) in Lemma 12. Let us write $M=[M_1, M_2, \dots, M_v]$ where the positive integers $M_1, M_2, \dots, M_v$ satisfy (62). Then, obviously, we have

$$\varphi(n+M) = \varphi(n) \quad \text{for} \quad n = 1, 2, \dots . \tag{92}$$

Let x, y be arbitrary fixed positive integers and N a positive integer. Then with respect to (65), (66), (67), (92) and (iii), we obtain from (61) that for fixed x, y and $N \to +\infty$ we have

$$\begin{aligned}
\sum_{l=1}^{N} g(x+x^2yMl) &= \sum_{l=1}^{N} \varphi(x+x^2yMl) + \sum_{l=1}^{N} \psi(x+x^2yMl) + \sum_{l=1}^{N} \varrho(x+x^2yMl) = \\
&= \sum_{l=1}^{N} \varphi(x) + \sum_{l=1}^{N} \sum_{j=1}^{w} b_j t_j^{x+x^2yMl} + O\left(\sum_{l=1}^{N} \frac{1}{x+x^2yMl}\right) = \\
&= N\varphi(x) + \sum_{j=1}^{w} b_j t_j^{x+x^2yM} \frac{1-t_j^{x^2yMN}}{1-t_j^{x^2yM}} + O(\log N) = \\
&= N\varphi(x) + O\left(\sum_{j=1}^{w} \frac{1}{|1-t_j^{x^2yM}|}\right) + O(\log N) - \\
&= N\varphi(x) + O(1) + O(\log N) = N\varphi(x) + O(\log N).
\end{aligned} \tag{93}$$

On the other hand, $(x, 1+xyMl)=1$ thus by the multiplicativity of the function $g(n)$ we have

$$\sum_{l=1}^{N} g(x+x^2yMl) = \sum_{l=1}^{N} g(x)g(1+xyMl) = g(x) \sum_{l=1}^{N} g(1+xyMl). \tag{94}$$

(93) and (94) yield that

$$g(x) \sum_{l=1}^{N} g(1+xyMl) = N\varphi(x) + O(\log N). \tag{95}$$

Let us put $x=1$. By (16) and the multiplicativity of $g(n)$, we have $g(1)=1$, thus we obtain that

$$\sum_{l=1}^{N} g(1+yMl) = N\varphi(1) + O(\log N) \tag{96}$$

for all y. Let us put $y=1$ in (95) and apply (96) with x in place of y. Then the left-hand side of (95) is

$$(97)\quad g(x)\sum_{l=1}^{N} g(1+xMl) = g(x)\big(N\varphi(1)+O(\log N)\big) = Ng(x)\varphi(1)+O(\log N).$$

(95) and (97) yield that

$$Ng(x)\varphi(1) = N\varphi(x)+O(\log N)$$

hence

$$g(x)\varphi(1) = \varphi(x)+O\left(\frac{\log N}{N}\right).$$

For $N\to+\infty$, we obtain that

$$(98)\qquad g(x)\varphi(1) = \varphi(x) \quad \text{for} \quad x = 1, 2, \dots .$$

By (64), there exists a positive integer x_0 for which $\varphi(x_0)\neq 0$. Then (97) yields that

$$g(x_0)\varphi(1) = \varphi(x_0) \neq 0$$

which implies that

$$\varphi(1) \neq 0.$$

But then we obtain from (97) that

$$(99)\qquad g(x) = \frac{1}{\varphi(1)}\,\varphi(x).$$

(92) and (99) yield (91) and this completes the proof of Lemma 13.

7. In this section we complete the proof of Theorem 1 by showing that $f(n)\in\mathcal{A}$ *and* (16) *imply that there exist h, m and* χ *satisfying* (i) *and* (ii) *in Theorem 1.*

By Lemma 13, $f(n)\in\mathcal{A}$ and (16) imply that the multiplicative arithmetic function $g(n)$ defined by (48) is *periodic*, i.e., there exists a positive integer M for which (91) holds. By Lemma 7, for $p>p_0$, $\alpha=1, 2, \dots$ we have $f(p^\alpha)\neq 0$. Let $q_1, q_2, \dots, q_z$ denote the finitely many prime numbers for which

$$g(q_i^{\alpha_i}) = \frac{f(q_i^{\alpha_i})}{q_i^{h\alpha_i}} = 0$$

holds for a positive integer α_i and let us write $m=[M, q_1, \dots, q_z]$. Then (91) implies that we have

$$(100)\qquad g(n+m) = g(n) \quad \text{for} \quad n = 1, 2, \dots .$$

Furthermore, we have

$$(101)\qquad g(n)\neq 0 \quad \text{for} \quad (n, m) = 1.$$

Let us define the multiplicative arithmetic function $G(n)$ in the following way: let

$$G(n) = g(n) \quad \text{if} \quad (n, m) = 1$$

and

$$G(p^\alpha) = 0 \quad \text{if} \quad p|m \quad \text{and} \quad \alpha = 1, 2, \dots .$$

Then by (100) and (101), we have

$$G(n) = \begin{cases} = 0 & \text{if} \quad (n, m) > 1, \\ \neq 0 & \text{if} \quad (n, m) = 1 \end{cases} \tag{102}$$

and

$$G(n+m) = G(n) \quad \text{for} \quad n = 1, 2, \ldots . \tag{103}$$

We are going to show that the function $G(n)$ is *strictly* multiplicative, i.e.,

$$G(ab) = G(a)G(b) \quad \text{for} \quad a = 1, 2, \ldots, \quad b = 1, 2, \ldots . \tag{104}$$

In fact, if $(ab, m) > 1$ then we have 0 on both sides of (104), while if $(ab, m) = 1$ then by Dirichlet's theorem, there exist prime numbers p_1, p_2 such that

$$a \equiv p_1 \pmod{m},$$
$$b \equiv p_2 \pmod{m}$$

and $p_1 \neq p_2$, i.e., $(p_1, p_2) = 1$. Then by the multiplicativity of $G(n)$, we have

$$G(p_1 p_2) = G(p_1)G(p_2) \tag{105}$$

and by (103),

$$G(p_1 p_2) = G(ab), \quad G(p_1) = G(a) \quad \text{and} \quad G(p_2) = G(b). \tag{106}$$

(105) and (106) yield (104).

It is well-known (see e.g. [4], p. 102) that (102), (103) and (104) imply that the function $G(n)$ must be a character modulo m:

$$G(n) = \chi(n) \quad \text{for} \quad n = 1, 2, \ldots$$

where $\chi(n)$ is a character modulo m. But then for $(n, m) = 1$, we have

$$f(n) = n^h g(n) = n^h G(n) = n^h \chi(n)$$

which proves that (i) in Theorem 1 holds.

In order to show that also (ii) holds, assume that $m \geqq 2$, write m in the form $m = p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_r^{\alpha_r}$ (where $p_1, p_2, \ldots, p_r$ are distinct prime numbers, $\alpha_1, \alpha_2, \ldots, \alpha_r$ are positive integers) and let

$$\chi(n) = \chi_1(n)\chi_2(n)\ldots\chi_r(n) \tag{107}$$

where $\chi_1(n), \chi_2(n), \ldots, \chi_r(n)$ are the uniquely determined characters modulo $p_1^{\alpha_1}, p_2^{\alpha_2}, \ldots, p_r^{\alpha_r}$, respectively, for which (107) holds for all n.

Let $1 \leqq i \leqq r$. Assume first that $\chi_i(n)$ is the principal character modulo p_ι and let j be a positive integer (the case $j = 0$ is trivial). Then by (100) and (107), we have

$$f(p_i^{\alpha_i + j}) = p_i^{h(\alpha_i + j)} g(p_i^{\alpha_i + j}) = p_i^{h(\alpha_i + j)} g(m + p_i^{\alpha_i + j}) =$$
$$= p_i^{h(\alpha_i + j)} g\big(p_i^{\alpha_i}(m p_i^{-\alpha_i} + p_i^j)\big) = p_i^{h(\alpha_i + j)} g(p_i^{\alpha_i}) g(m p_i^{-\alpha_i} + p_i^j) =$$
$$= p_i^{h(\alpha_i + j)} p_i^{-h\alpha_i} f(p_i^{\alpha_i}) \prod_{l=1}^{r} \chi_l(m p_i^{-\alpha_i} + p_i^j) =$$
$$= p_i^{hj} f(p_i^{\alpha_i}) \chi_i(m p_i^{-\alpha_i} + p_i^j) \prod_{\substack{1 \leqq l \leqq r \\ l \neq i}} \chi_l(m p_i^{-\alpha_i} + p_i^j) = p_i^{hj} f(p_i^{\alpha_i}) \prod_{\substack{1 \leqq l \leqq r \\ l \neq i}} \chi_l(p_i^j)$$

since $(mp_i^{-\alpha_i}+p_i^j, p_l)=1$ for $l=1, 2, \dots, r$ and $mp_i^{-\alpha_i}+p_i^j\equiv p_i^j \pmod{p_l^{\alpha_l}}$ for $1\leqq l\leqq r$, $l\neq i$. Thus (ii) holds if $\chi_i(n)$ is the principal character modulo $p_i^{\alpha_i}$.

Assume now that $\chi_i(n)$ is not the principal character modulo $p_i^{\alpha_i}$. Then there exists a positive integer n_1 for which

$$\chi_i(n_1)\neq 0 \tag{108}$$

and

$$\chi_i(n_1)\neq 1 \tag{109}$$

hold; (108) implies that

$$(n_1, p_i)=1. \tag{110}$$

We have $(mp_i^{-\alpha_i}, p_i)=1$ thus there exists a positive integer x for which the linear congruence

$$mp_i^{-\alpha_i}x+1\equiv n_1 \pmod{p_i^{\alpha_i}} \tag{111}$$

holds. Then for $j=1, 2, \dots,$ we obtain with respect to (100), (107), (108), (110) and (111) that

$$\begin{aligned} g(p_i^{\alpha_i+j}) &= g(mp_i^j x+p_i^{\alpha_i+j}) = \\ &= g\big(p_i^{\alpha_i+j}(mp_i^{-\alpha_i}x+1)\big)=g(p_i^{\alpha_i+j})\,g(mp_i^{-\alpha_i}x+1)= \\ &= g(p_i^{\alpha_i+j})\prod_{l=1}^{r}\chi_l(mp_i^{-\alpha_i}x+1)= \\ &= g(p_i^{\alpha_i+j})\chi_i(mp_i^{-\alpha_i}x+1)\prod_{\substack{1\leqq l\leqq r\\ l\neq r}}\chi_l(mp_i^{-\alpha_i}x+1)= \\ &= g(p_i^{\alpha_i+j})\chi_i(n_1)\prod_{\substack{1\leqq l\leqq r\\ l\neq i}}\chi_l(1)=g(p_i^{\alpha_i+j})\chi_i(n_1) \end{aligned} \tag{112}$$

since

$$(mp_i^{-\alpha_i}x+1, p_l)=1$$

holds trivially for $l\neq i$ while for $l=i$, this follows from (110) and (111). (109) and (112) imply that

$$g(p_i^{\alpha_i+j})=0$$

hence

$$f(p_i^{\alpha_i+j})=p_i^{h(\alpha_i+j)}g(p_i^{\alpha_i+j})=0$$

thus (ii) holds also in this case which completes the proof of Theorem 1.

REFERENCES

[1] HASSE, H., *Zahlentheorie,* Akademie-Verlag, Berlin, 1949.

[2] KUIPERS, L.—NIEDERREITER, H., *Uniform distribution of sequences,* Wiley Interscience Publications, New York, 1974.

[3] LOVÁSZ, L.—SÁRKÖZY, A.—SIMONOVITS, M., On additive arithmetic functions satisfying a linear recursion, *Ann. Univ. Sci. Budapest. Eötvös Sect. Math.* (to appear).

[4] PRACHAR, K., *Primzahlverteilung,* Springer-Verlag, Berlin—Göttingen—Heidelberg, 1957.

Mathematical Institute of the Hungarian Academy of Sciences
H—1053 Budapest, Reáltanoda u. 13—15.

(Received April 20, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 105—124.

ON THE QUESTION OF DESCRIPTION OF THE BEHAVIOUR OF FINITE AUTOMATA

by

A. ÁDÁM

To the memory of Professor László Kalmár

Summary

In § 1 the concepts of automaton mapping, finite (Moore) automaton and some fundamental notions concerning these are introduced. Some known facts (among them, the connection of finite realizability and right-congruences) are also stated.

§ 2 starts with two formulations of the description problem of finitely realizable automaton mappings. In the continuation some criticisms on the method that strives to deal with these problems by using regular expressions are contained.

From among each family of isomorphic automata, a single automaton with particular notation of states and output signs is selected in § 3. These automata are called standard. A technique — using certain tables, named codes — for describing all standard automata is contained in § 4.

The considerations of the paper culminate with §§ 5—6. A complexity notion for automata is defined so that precisely the reduced automata are of finite complexity. After exposing a new version of the description problem (mentioned already in § 2) four more particular problems of this are raised. These particular problems surround the question how all the codes characterizing reduced automata may be obtained constructively.

§ 1. Fundamental notions

1.1. How does a finite abstract automaton[1] behave towards the outer world? More precisely: how can be described all the possible reactions of finite automata to the various incoming stimuli, so that the inner processes of the automata are left out of consideration and the description does not comprise the types of behaviour which remain outside of the abilities of *finite* automata?

The notion of finitely realizable automaton mapping[2] is a straightforward tool in investigating this question. The question of studying the manners of behaviour can be raised, in a more exact form, such that these automaton mappings are to be studied.

Now we give two (essentially coinciding) definitions of automaton mapping. (In §§ 1—3 a broader family will be considered; the property of finite realizability is not required unless explicitly stated. Beginning with § 4, all automaton mappings are meant finitely realizable.)

Let X, Y be (disjoint) finite sets. $X=\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ is viewed as the same ordered set throughout the paper. The (non-commutative) free monoid over X is denoted by $F(X)$. The unit element of $F(X)$ (the empty word) is denoted by e.

[1] In what follows, always flinite deterministic automata with a distinguished (initial) state in the sense of Moore will be considered. For the sake of definiteness, the reader may think of an automaton of this type in these introductory sentences, too.

[2] Instead of "automaton mapping", the terminology "(retrospective) sequential function" is also used in the literature.

The elements of X are identified with the elements (words) p in $F(X)$ whose length $L(p)$ equals 1.

A pair (P, σ) is called an *automaton mapping in the first sense* if P is a partition of $F(X)$ and σ is a bijection of the factor set $F(X)/P$ onto Y. (This definition implies obviously $\operatorname{ind} P = |Y| < \infty$ where $|Y|$ denotes the number of elements of Y and $\operatorname{ind} P$ (the index of P) denotes the number of classes modulo P, i.e., $|F(X)/P|$.)

A mapping α of $F(X)$ into $F(Y)$ is called an *automaton mapping in the second sense* if

(i) the image $\alpha(e)$ of e is an element of Y;

(ii) to each choice of $p(\in F(X))$ and $x(\in X)$ there is a $y(\in Y)$ such that $\alpha(px) = \alpha(p)y$ and

(iii) to each $y(\in Y)$ there is a $p(\in F(X))$ such that the last letter of $\alpha(p)$ is y. (It follows that $L(\alpha(p)) = 1 + L(p)$ for every $p(\in F(X))$.)

These two versions of the notion of automaton mapping are essentially equivalent. Indeed, to any (P, σ) we can define an α by putting

$$\alpha(p) = \sigma(\bar{p}_0)\sigma(\bar{p}_1)\sigma(\bar{p}_2)\dots\sigma(\bar{p}_k)$$

where

$p = x_1 x_2 \dots x_k$ ($k = L(p)$ and $x_1, \dots, x_k$ are arbitrary — not necessarily distinct — elements of X),

$p_i = x_1 x_2 \dots x_i$ for each i $(0 \leqq i \leqq k)$,

$\bar{p}$ is the class mod P which contains p.

It is easy to see that each α is yielded exactly once in this manner.

In what follows, we shall always consider the first definition of automaton mapping. The second one was exposed only because this is familiar in the literature (possibly with the modification in (i) that $\alpha(e)$ is the unit element of $F(Y)$).

Two automaton mappings (P_1, σ_1) and (P_2, σ_2) are called *isomorphic* if $P_1 = P_2$ holds. This definition reflects that the second component σ does not play an essential role in the (first) notion of automaton mapping.

1.2. The concept of initial finite Moore automaton is thought to be known. We denote by $A, X, Y, \delta, \lambda, a^*$ the state set, input set, output set, transition function, output function, initial state (respectively) of the automaton **A**. We consider always initially connected automata, i.e., ones in which

$$\forall a(\in A) \exists p(\in F(X)) (\delta(a^*, p) = a)$$

is fulfilled. We suppose that λ is surjective.

Let

$$\mathbf{A}_1 = (A_1, X, Y_1, \delta_1, \lambda_1, a_1^*), \quad \mathbf{A}_2 = (A_2, X, Y_2, \delta_2, \lambda_2, a_2^*)$$

be two automata. If $\alpha_A\colon A_1 \to A_2$ and $\alpha_Y\colon Y_1 \to Y_2$ are bijections and the three equalities

$$(1.1) \qquad (a_1^*)^{\alpha_A} = a_2^*,$$

$$(1.2) \qquad (\delta_1(a_1, x))^{\alpha_A} = \delta_2(a_1^{\alpha_A}, x),$$

$$(1.3) \qquad (\lambda_1(a_1))^{\alpha_Y} = \lambda_2(a_1^{\alpha_A})$$

hold for each choice of $a_1(\in A_1)$, $x(\in X)$, then we say that $\mathbf{A}_1, \mathbf{A}_2$ are *isomorphic* and the pair (α_A, α_Y) is an *isomorphism.* [3]

If the correspondences α_A, α_Y between two automata $\mathbf{A}_1, \mathbf{A}_2$ satisfy (1.1), (1.2), (1.3) and α_A is a surjection only [4], then we say that the pair (α_A, α_Y) is a *homomorphism* of $\mathbf{A}_1$ onto $\mathbf{A}_2$. If such a homomorphism may be established, then $\mathbf{A}_2$ is called a *homomorphic image* of $\mathbf{A}_1$.

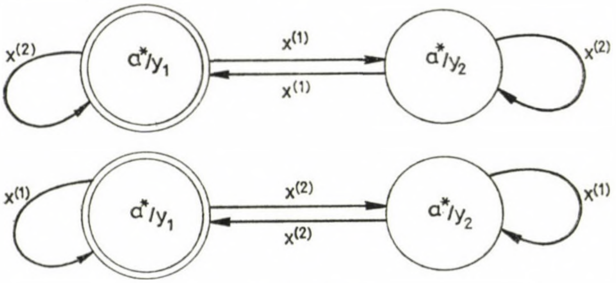


Fig. 1

If each possible homomorphism of an automaton $\mathbf{A}$ is an isomorphism, then we say that $\mathbf{A}$ is *simple* (or *reduced*).

It is clear that the notion of isomorphism is a particular case of the notion of homomorphism.

Let (α_A, α_Y) be a homomorphism (of $\mathbf{A}_1$ onto $\mathbf{A}_2$) such that $\mathbf{A}_1, \mathbf{A}_2$ have the same output set Y. If α_Y is the identical mapping of Y, then we say that α_A is a *state-homomorphism* and $\mathbf{A}_2$ is a *state-homomorphic image of* $\mathbf{A}_1$.

1.3. Let $A=(A, X, Y, \delta, \lambda, a^*)$ be an automaton. Define an automaton mapping in the following manner:

let $p\equiv q \pmod P$ hold if and only if

$$\lambda(\delta(a^*, p)) = \lambda(\delta(a^*, q))$$

where p, q are arbitrary elements of $F(X)$,

let $\sigma(\bar{p})$ equal $\lambda(\delta(a^*, p))$ where $\bar{p}$ is the class mod P containing p.
We say that $\mathbf{A}$ *realizes* (or *induces*) the pair (P, σ).

1.4.

PROPOSITION 1. *If the automaton* $\mathbf{A}_2$ *is a homomorphic image of the automaton* $\mathbf{A}_1$, *then they induce isomorphic automaton mappings.*

PROOF. By use of (1.1)—(1.3) we can deduce

$$(\delta_1(a_1^*, p))^{\alpha_A} = \delta_2(a_2^*, p)$$

and

$$\lambda_2(\delta_2(a_2^*, p)) = \lambda_2((\delta_1(a_1^*, p))^{\alpha_A}) = (\lambda_1(\delta_1(a_1^*, p)))^{\alpha_Y}. \tag{1.4}$$

[3] This kind of isomorphism is called (A, Y)-isomorphism in [7]. The automata in Fig. 1 are not isomorphic in this sense.

[4] α_Y is a bijection in this case as well.

The conclusion follows from the equality of the first and third expressions in (1.4) by the bijectivity of α_Y and the corresponding definitions. □

SUPPLEMENT. *If* $\mathbf{A}_2$ *is a state-homomorphic image of* $\mathbf{A}_1$, *then* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *realize the same automaton mapping.* □

Let P be a partition of $F(X)$. Define the partition $\mathfrak{M}(P)$ of $F(X)$ in the following manner:

$$p \equiv q\left(\mathrm{mod}\, \mathfrak{M}(P)\right) \Leftrightarrow \forall\, r\left(pr \equiv qr (\mathrm{mod}\, P)\right)$$

where p, q, r are elements of $F(X)$. One can easily show that[5]

$\mathfrak{M}(P) \subseteqq P$,

$\mathfrak{M}(P)$ is a right-congruence of $F(X)$ and

$P^* \subseteqq P$ implies $P^* \subseteqq \mathfrak{M}(P)$ for each right-congruence P^* of $F(X)$.

PROPOSITION 2. *Let* P *be a partition of* $F(X)$. *Denote by* $\mathfrak{A}_p$ *the set of (all) finite automata* $\mathbf{A}$ *such that the first component of the automaton mapping induced by* $\mathbf{A}$ *equals to* P. *If* $\mathfrak{A}_\pi \neq \emptyset$, *then* $\mathfrak{A}_p$ *contains a reduced automaton* $\mathbf{A}_0$, *this* $\mathbf{A}_0$ *is unique apart from isomorphy, and* $\mathbf{A}_0$ *is a homomorphic image of each element of* $\mathfrak{A}_p$.

REMARKS. The condition $\mathfrak{A}_p \neq \emptyset$ is needed because we consider finite automata only; in this case, $\mathfrak{A}_p \neq \emptyset$ implies $\mathrm{ind}\, P < \infty$ evidently. If automata with an infinity of states (and with a finite number of output signs) are also taken into account, then the assertion remains valid and $\mathfrak{A}_p \neq \emptyset$ can be replaced by $\mathrm{ind}\, P < \infty$ equivalently. (We shall not deal, however, with infinite automata.)

PROOF of Proposition 2.

Case 1. $\mathrm{ind}\, \mathfrak{M}(P)$ is finite. *We construct* $\mathbf{A}_0$ by what follows. The states of $\mathbf{A}_0$ correspond to the classes $\mathrm{mod}\, \mathfrak{M}(P)$ in a one-to-one manner. The initial state a^* corresponds to the class containing the empty word. Let x be an input sign, a be a state and p be an element of $F(X)$ such that $\bar{p}$ (the class $\mathrm{mod}\, \mathfrak{M}(P)$ containing p) corresponds to a; we define $\delta(a, x)$ as the state that corresponds to the class $\overline{px}$. Let the output function λ be determined in such a manner that $\lambda(a) = \lambda(b)$ holds exactly if $p \equiv q$ (mod P) where p, q are chosen so that a corresponds to $\bar{p}$ and b corresponds to $\bar{q}$. It may easily be checked that $\mathbf{A}_0$ has been defined meaningfully and (up to isomorphy) uniquely.

We verify that $\mathbf{A}_0$ *belongs to* $\mathfrak{A}_p$. Indeed, an easy induction shows that $\bar{p}$ corresponds to $\delta(a^*, p)$ for each $p(\in F(X))$, thus $p \equiv q$ (mod P) and $\lambda(\delta(a^*, p)) = \lambda(\delta(a^*, q))$ are equivalent.

Let an arbitrary element $\mathbf{A}$ of $\mathfrak{A}_p$ be considered. Our next aim is *to establish a homomorphism of* $\mathbf{A}$ *onto* $\mathbf{A}_0$. First we note that the transition function δ of $\mathbf{A}$ fulfils

$$(1.5) \qquad \delta(a^*, p) = \delta(a^*, q) \Rightarrow p \equiv q\left(\mathrm{mod}\, \mathfrak{M}(P)\right).$$

(In fact, if the left-hand side of (1.5) holds, then $p \equiv q$ (mod P) because $\mathbf{A} \in \mathfrak{A}_p$. Furthermore, the left-hand side of (1.5) defines clearly a right-congruence of $F(X)$.

[5] Cf. § 10.1 in [6].

This right-congruence is a refinement of P by the first sentences of the proof.) The assignment

$$\delta(a^*, p) \Rightarrow \delta_0(a_0^*, p)$$

(where δ_0 is the transition function and a_0^* is the initial state of $\mathbf{A}_0$) is a surjection α_A of A onto A_0 (where the state sets of $\mathbf{A}$, $\mathbf{A}_0$ are denoted by A, A_0, resp.), the homomorphism properties (1.1)—(1.3) with some α_Y may be checked without difficulty since both $\mathbf{A}$, $\mathbf{A}_0$ belong to $\mathfrak{A}_p$.

Let $\mathbf{A}_0'$ be an arbitrary homomorphic image of $\mathbf{A}_0$, denote its state set by A_0'. Then $|A_0'| \leqq |A_0|$. On the other hand, $\mathbf{A}_0' \in \mathfrak{A}_p$, this implies $|A_0'| \geqq |A_0|$ by the preceding considerations of this proof. The resulting equality $|A_0'| = |A_0|$ is impossible unless the *finite* automata $\mathbf{A}_0'$, $\mathbf{A}_0$ are isomorphic.

We have got that $\mathbf{A}_0$ *is reduced.*

The *unicity of the reduced automaton* $\mathbf{A}_0$ is obvious.

Case 2. ind $\mathfrak{M}(P)$ is infinite. Let $\mathbf{A}$ be an arbitrary element of $\mathfrak{A}_p$. The implication (1.5) shows that $\mathbf{A}$ cannot be a finite automaton. Consequently, $\mathfrak{A}_p = \emptyset$. □

The bifurcation in the previous proof leads to the

COROLLARY 1. *Let (P, σ) be an arbitrary automaton mapping. The following three assertions are equivalent:*

(i) *The index of $\mathfrak{M}(P)$ is finite.*

(ii) *There exists a (finite) automaton realizing (P, σ).*

(iii) *There exists a reduced (and, apart from isomorphy, uniquely determined; finite) automaton realizing (P, σ).* □

REMARK. It is known that every right-congruence P^* of $F(X)$ with ind $P^* < \infty$ has a refinement R such that[6] R is a congruence and ind $R < \infty$. This fact implies that (i) is equivalent to the following assertion:

(iv) *The index of the largest (two-sided) congruence that is a refinement of P is finite.*

1.5. The equivalence of the assertions (i)—(iv) in Corollary 1 (and the Remark after it) shows the importance of a class of partitions of $F(X)$. We introduce therefore the following terminology:

A partition P of the free monoid $F(X)$ (generated by the finite set X) is called *hyper-finite* if $\mathfrak{M}(P)$ is of finite index. It is well-known that the family of hyper-finite partitions of $F(X)$ is *properly* included in the family of all partitions of $F(X)$ with finite index.

[6] For a simple proof, see Statement 2 in Section 3 of [4].

§ 2. The basic problem and some comments on the theory of regular expressions

2.1.

BASIC PROBLEM *(first version).* For an arbitrary finite X, let a constructive description of all hyper-finite partitions of $F(X)$ be given.

BASIC PROBLEM *(second version).* For an arbitrary finite X, let a constructive description of all reduced finite Moore automata, whose input set equals to X, be given (apart from isomorphy).

By virtue of Proposition 2 and Corollary 1, these versions of the basic problem are equivalent to each other. Both of them is a formulation of the question how we can get an overview of all the (non-isomorphic) finitely realizable automaton mappings.

2.2. The way we choose for making some approach towards solving the basic problem *will be different* from the study of regular expressions. Sections 2.3—2.5 contain some considerations concerning why we prefer another way. The reader may omit these sections and jump to § 3.

2.3. The notion of regular expressions in sense of KLEENE [10] is supposed to be known in the remainder of this paragraph.

Let P be a hyper-finite partition of $F(X)$. We mention a procedure, starting with P, which consists of two steps.

Step I: We take some sets $M_1, M_2, \dots, M_t (\subseteqq F(X))$ such that the equivalence

$$p \equiv q \pmod P) \Leftrightarrow \forall\, i (p \in M_i \Leftrightarrow q \in M_i)$$

holds $(1 \leqq i \leqq t)$.

(It is possible that we put $t = \operatorname{ind} P$ and we choose the M_i's as the classes modulo P. This choice is not economical, the smallest possible value for t is $\lceil \log_2 \operatorname{ind} P \rceil$ where $\lceil a \rceil$ denotes the integer fulfilling $a \leqq \lceil a \rceil < a+1$.)

Step II: We represent each set M_i by a regular expression.

(We have to explain why M_i is regular. Indeed, P is hyper-finite, it occurs in a finitely realizable automaton mapping by our Corollary 1, and an arbitrary set of states of a finite automaton represents a regular event by a well-known theorem of KLEENE ([10], Theorem 5; [7], Theorem 2.5.1), hence each class modulo P is regular. M_i is equal to the union of certain classes, thus M_i is again a regular event.)

These steps yield a collection of t regular expressions (over X). We can view this collection as a description of an arbitrary automaton mapping (P, σ) (where P is the given hyper-finite partition, σ is arbitrary).

Let us consider the situation when we start with an automaton $\mathbf{A}$, we introduce (P, σ) as the automaton mapping represented by $\mathbf{A}$, and finally we apply Steps I and II. This procedure is essentially equivalent to the following one: we replace $\mathbf{A}$ by certain automata $\mathbf{A}_1, \dots, \mathbf{A}_t$ such that $\mathbf{A}_i$ (where $1 \leqq i \leqq t$) is obtained from $\mathbf{A}$ by some modification of the output function, in fact, $\mathbf{A}_i$ has two output signs

y_0, y_1 only, and its output function λ_i is defined by

$$\lambda_i(a) = \begin{cases} y_1 & \text{if} \quad p \in M_i \\ y_0 & \text{if} \quad p \notin M_i \end{cases}$$

where $p(\in F(X))$ satisfies $\delta(a^*, p) = a$. ($\lambda_i(a)$ does not depend on the particular choice of p.) We characterize — for each i — the set of words p, fulfilling $\lambda_i(\delta(a^*, p)) = y_1$, by a regular expression.

We can say that we pay a special attention to the automata with *only two* output signs if we examine the automaton mappings by use of Steps I and II.

2.4. The idea, seen in Section 2.3, has great and (it appears) insurmountable disadvantages from the view point of the unicity. The system consisting of $M_1, \dots, M_t$ is not uniquely determined by the partition P (unless we choose [7] $t = \text{ind}\, P$, this choice is not in the least economical). To any set M_i, an immense diversity of regular expression representing M_i exists.

The problems of this type remain essentially unchanged even if the automaton **A** (in Section 2.3) is supposed to be simple.

Although the theory of regular expressions is widely elaborated (see the survey [8]—[9]), the lack of unicity (as seen above) seems to be a very serious difficulty, and this feature justifies the endeavour to seek other methods for attacking the basic problem [8].

2.5. The line of considerations we follow in the sequel can be characterized by saying that we give key-role to automata having the *possibly largest number* of output signs (i.e., to automata whose output function is bijective).

§ 3. Standardization

3.1. The notion of isomorphy of two automata was introduced in Section 1.2. We regard two automata to be essentially different only if they are not isomorphic in sense of this definition.

Consider an *isomorphy family,* i.e., a maximal family of pairwise isomorphic automata. The main goal of § 3 is to select precisely one automaton (provided with a special notation of the states and output signs) in an (arbitrary) isomorphy family; these will be called standard automata. Analogously, a standard automaton mapping will be picked out in each isomorphy family of automaton mappings.

Since the equality of automata or automaton mappings will occur in Propositions 7 and 9, now we give formal definitions of how these equalities are understood.

Let **A**, **A**$'$ be two automata whose states are denoted in form $a_1, a_2, \dots,$ and whose output signs are denoted in form $y_1, y_2, \dots$. The automata **A**, **A**$'$ are considered *equal* if there exists an isomorphism (α_A, α_Y) between them such that α_A preserves the labelling (subscripts) of the states and α_Y preserves the labelling of the output signs.

[7] Just in case of this choice the following difficult problem arises: if a family of regular expressions is given, how can we decide whether or not the regular events represented by them are pairwise disjoint?

[8] The dissertation [11] illustrates the difficulties that may arise when questions on regular expressions are dealt with.

Let $(P, \sigma), (P', \sigma')$ be two automaton mappings such that the classes mod P (or mod P') are written in the form $K_1, K_2, \ldots$. The mappings $(P, \sigma), (P', \sigma')$ are called *equal* if

(1) each class K_i modulo P is exactly a class modulo P', moreover, it is written as K_i again mod P', and

(2) for each class K_i, the equality $\sigma(K_i)=\sigma'(K_i)$ holds.

3.2. We introduce a full ordering $\lessdot$ in the monoid $F(X)$ by the following rule:

$$p \lessdot q \Leftrightarrow \begin{cases} \text{either } L(p) < L(q), \\ \text{or } L(p) = L(q) \text{ and } p \text{ precedes } q \text{ lexicographically.} \end{cases}$$

Next we state some immediate consequences of this definition.

PROPOSITION 3. *Let p, q, r, s be arbitrary elements of $F(X)$. If one of the three formulae $p \lessdot q, pr \lessdot qr, rp \lessdot rq$ holds, then the other two formulae are also valid. If $p \lessdot q$ and $L(r)=L(s)$, then $pr \lessdot qs$.* □

Let us demonstrate $F(X)$ in the manner seen in Fig. 2 (with $n=3$). The $\lessdot$-enumeration is obviously the following: we begin with the empty word e, ..., we pass through the k^{th} level from left to right, (we jump from the rightmost element of the k^{th} level to the leftmost element of the $(k+1)^{\text{th}}$ level,) we pass through the $(k+1)^{\text{th}}$ level from left to right, ... (ad infinitum).

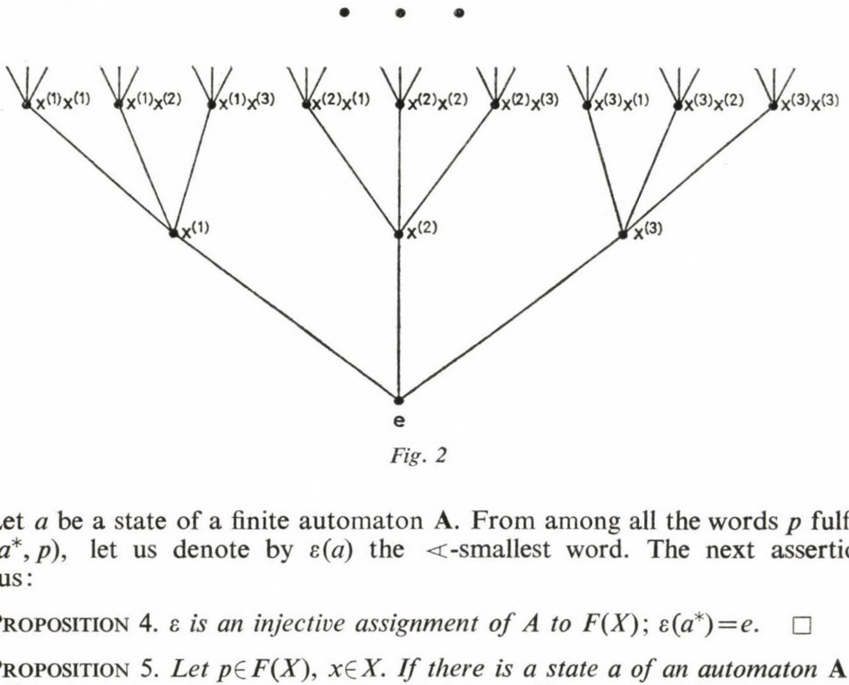


Fig. 2

Let a be a state of a finite automaton $\mathbf{A}$. From among all the words p fulfilling $a=\delta(a^*, p)$, let us denote by $\varepsilon(a)$ the $\lessdot$-smallest word. The next assertion is obvious:

PROPOSITION 4. *ε is an injective assignment of A to $F(X)$; $\varepsilon(a^*)=e$.* □

PROPOSITION 5. *Let $p \in F(X)$, $x \in X$. If there is a state a of an automaton $\mathbf{A}$ such that $\varepsilon(a)=px$, then $\mathbf{A}$ has a state b satisfying $\varepsilon(b)=p$ and $\delta(b, x)=a$.*

REMARK. b is uniquely determined by the equality $\varepsilon(b)=p$.

PROOF. Denote $\delta(a^*, p)$ by b and $\varepsilon(b)$ by q. If $q \prec p$ were true, then $qx \prec px$ would follow from Proposition 3; this relation and

$$\delta(a^*, qx) = \delta(\delta(a^*, q), x) = \delta(b, x) =$$
$$= \delta(\delta(a^*, p), x) = \delta(a^*, px) = a$$

would contradict the equality $\varepsilon(a)=px$. The Remark holds by Proposition 4. □

3.3. Let us visualize **A** by a graph in the customary manner. Then to each transition $\delta(b, x)=a$ an edge (marked by x) from the vertex b to the vertex a is assigned. Call an edge $\overrightarrow{ba}$ a *distinguished edge* if (a is not the initial state of **A** and) the word p and the input sign x, determined by $\varepsilon(a)=px$, satisfy the following conditions (i) and (ii):

(i) $b=\delta(a^*, p)$,

(ii) $\overrightarrow{ba}$ is marked by x.

By Proposition 5, the distinguished edges form a spanning tree in the graph of **A**. Each edge of this tree is directed towards the vertex which is more distant from the initial state.

3.4. A finite automaton $\mathbf{A}=(A, X, Y, \delta, \lambda, a_1)$ is called a *standard automaton* if the following two requirements are fulfilled:

(α) $A=\{a_1, a_2, \ldots, a_v\}$ where $v=|A|$ and the equivalence

$$l < m \Leftrightarrow \varepsilon(a_l) \prec \varepsilon(a_m)$$

holds for every choice of l and m $(1 \leqq l \leqq v, 1 \leqq m \leqq v)$,

(β) $Y=\{y_1, y_2, \ldots, y_t\}$ where $t=|Y|, \lambda(a_1)=y_1$ and whenever $\lambda(a_m)=y_k$, then there is an $l(<m)$ satisfying $\lambda(a_l)=y_{k-1}$ $(1 \leqq m \leqq v, 1 \leqq l \leqq v, 2 \leqq k \leqq t)$.

PROPOSITION 6. *To each finite automaton* **A** *there is a standard automaton* **A**′ *such that they are isomorphic.*

PROOF. By an appropriate notation, we can denote the words being in the range of ε by $p_1, p_2, \ldots, p_v$ such that

$$p_1 \prec p_2 \prec \ldots \prec p_v.$$

Having done this, we introduce a new notation of the states such that $\varepsilon(a_i)=p_i$ $(1 \leqq i \leqq v)$. (α) is satisfied and obviously a_1 is the initial state (by $p_1=e$).

Let $A^+(\subseteqq A)$ be the set of states a_i fulfilling

$$\lambda(a_i) \notin \{\lambda(a_1), \lambda(a_2), \ldots, \lambda(a_{i-1})\} \tag{3.1}$$

$(1 \leqq i \leqq v)$. We introduce the new notation y_k for $\lambda(a_i)$ (where $a_i \in A^+$) such that k is defined by the formula

$$k = 1+|A^+ \cap \{a_1, a_2, \ldots, a_{i-1}\}|. \tag{3.2}$$

(Especially, $a_1 \in A^+$ and $\lambda(a_1)=y_1$.) We can easily see that the restriction of λ to A^+ is bijective and $a_i \in A^+$, $\lambda(a_i)=y_k$, $k>1$ imply (by (3.2)) $\lambda(a_l)=y_{k-1}$ where

l is the largest number such that $l<i$ and $a_l \in A^+$. For the complete verification of (β), let a_m be an arbitrary element of A (possibly $a_m \notin A^+$). Independently of the fulfilment of (3.1), there is an $a_i (\in A^+)$ such that $i \leqq m$ and $\lambda(a_i)=\lambda(a_m)$. We obtain the validity of (β) by summarizing these considerations (for each pair m, k such that $\lambda(a_m)=y_k$, $k>1$). □

PROPOSITION 7. *If two standard automata* $\mathbf{A}_1, \mathbf{A}'$ *are isomorphic, then they are equal.*

PROOF. Consider the assignments $\varepsilon, \varepsilon'$ for $\mathbf{A}, \mathbf{A}'$, resp. It is evident that the ranges of $\varepsilon, \varepsilon'$ are equal, moreover, (α) implies that $\varepsilon(a_i)=p$ and $\varepsilon'(a_i)=p$ are equivalent. The fact $\lambda(a_i)=\lambda'(a_i)$ follows easily from (β). □

3.5. Let (P, σ) be an automaton mapping. For each class K (mod P), we denote by $\varkappa(K)$ the $\lessdot$-smallest word from among the elements of K. Introduce the notation

$$(3.3) \qquad K_1, K_2, \ldots, K_t$$

(where $t=\operatorname{ind} P$) for the classes mod P such that for arbitrary classes K_i, K_j, the inequality $i<j$ and the formula $\varkappa(K_i) \lessdot \varkappa(K_j)$ are equivalent. (P, σ) is called a *standard automaton mapping* if $\sigma(K_i)=y_i$ is universally satisfied $(1 \leqq i \leqq t)$.

PROPOSITION 8. *To each automaton mapping* (P, σ) *there is a standard automaton mapping* (P', σ') *such that they are isomorphic.*

PROOF. Take $P'=P$; the standardness can be reached by a suitable notation of the output signs. □

PROPOSITION 9. *If two standard automaton mappings* (P, σ) *and* (P', σ') *are isomorphic, then they are equal.*

PROOF. $P=P'$ by the isomorphy. It is clear that the notation in (3.3) is determined by P uniquely, hence $\sigma=\sigma'$ by the definition of standard mappings. □

3.6. In this last section of § 3 we make some observations on the standardness of an automaton and the standardness of the mapping induced by it.

The notions of isomorphism and state-homomorphism were introduced in Section 1.2. If a pair (α_A, α_Y) is both an isomorphism and a state-homomorphism (i.e., if it is a homomorphism, α_A is a bijection and α_Y is the identical mapping of Y), then we call (α_A, α_Y) a *state-isomorphism.*

LEMMA 1. *If the automata* $\mathbf{A}_1, \mathbf{A}_2$ *are isomorphic and the automaton mappings induced by them are equal, then* $\mathbf{A}_1, \mathbf{A}_2$ *are state-isomorphic, too.*

PROOF. $\mathbf{A}_1$ and $\mathbf{A}_2$ have the same set Y of output signs. Consider an isomorphism (α_A, α_Y) of $\mathbf{A}_1=(A_1, X, Y, \delta_1, \lambda_1, a^*_{(1)})$ onto $\mathbf{A}_2=(A_2, X, Y, \delta_2, \lambda_2, a^*_{(2)})$. For an arbitrary $p(\in F(X))$, denote by a the state $\delta_1(a^*_{(1)}, p)$. The deduction

$$\lambda_1(a) = \lambda_1(\delta_1(a^*_{(1)}, p)) = \sigma(\bar{p}) =$$
$$= \lambda_2(\delta_2(a^*_{(2)}, p)) = \lambda_2(a^{\alpha_A}) = (\lambda_1(a))^{\alpha_Y}$$

shows that α_Y is the identical mapping of Y. □

PROPOSITION 10. *If* **A** *is a standard automaton, then the automaton mapping* (P, σ) *induced by* **A** *is standard.*

PROOF. Let K be a class mod P. Clearly, $\varkappa(K)$ is the $\lessdot$-smallest of the words $\varepsilon(a_{i_1}), \varepsilon(a_{i_2}), \dots, \varepsilon(a_{i_w})$ where $a_{i_1}, a_{i_2}, \dots, a_{i_w}$ run through the states a_i of **A** for which $\lambda(a_i)=\sigma(K)$. Let A^+ be as in the proof of Proposition 6. The standardness of **A** implies that the following three statements are equivalent for each $a_m(\in A^+)$ and $k(\in\{2, 3, \dots, |Y|\})$:

(i) $\lambda(a_m)=y_k$,

(ii) $\{\lambda(a_1), \lambda(a_2), \dots, \lambda(a_{m-1})\}$ does not contain y_k but it contains y_{k-1},

(iii) $a_m \in K_k$ where the classes K are numbered as in (3.3).

The equivalence of (i), (iii) and the obvious facts $a_1 \in K_1 \cap A^+$, $\lambda(a_1)=y_1$ mean that $\sigma(K_k)=y_k$ for each k, hence (P, σ) is standard. □

PROPOSITION 11. *Let* (P, σ) *be a finitely realizable standard automaton mapping. If* **A** *is an automaton inducing* (P, σ), *then* **A** *is state-isomorphic to a standard automaton.*

PROOF. Consider a standard automaton **A′** isomorphic to **A** (existing by Proposition 6). Let (P', σ') be the automaton mapping realized by **A′**. (P, σ) and (P', σ') are isomorphic by Proposition 1. (P', σ') is a standard mapping by Proposition 10, hence $(P, \sigma)=(P', \sigma')$ by Proposition 9. The state-isomorphy of **A** and **A′** follows by using Lemma 1. □

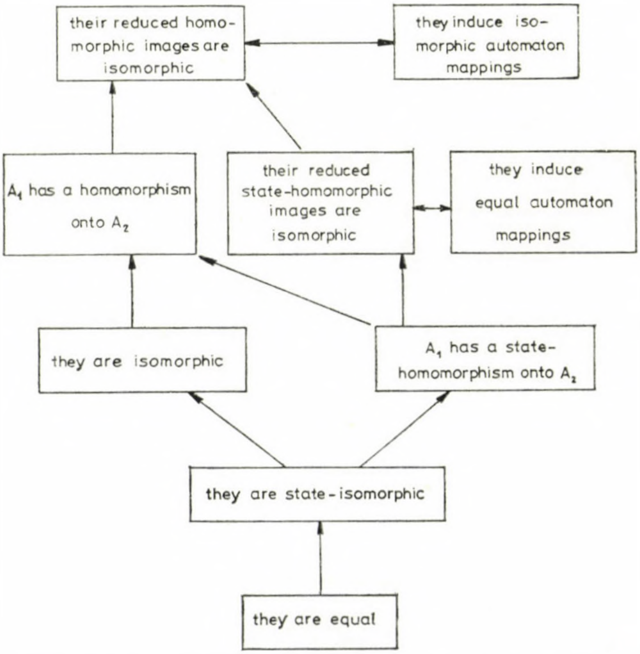

their reduced homomorphic images are isomorphic
they induce isomorphic automaton mappings
A_1 has a homomorphism onto A_2
their reduced state-homomorphic images are isomorphic
they induce equal automaton mappings
they are isomorphic
A_1 has a state-homomorphism onto A_2
they are state-isomorphic
they are equal

Table 1

The hierarchy of some assertions on two automata $\mathbf{A}_1, \mathbf{A}_2$

The last proposition of this paragraph summarizes the facts stated in Section 3.6.

PROPOSITION 12. *Let (P, σ) be a finitely realizable automaton mapping. Consider the isomorphy families of all automata realizing (P, σ). Every isomorphy family consists of pairwise state-isomorphic automata. If (P, σ) is standard, then each isomorphy family contains exactly one standard automaton. If (P, σ) is not standard, then all the automata inducing (P, σ) are non-standard.* □

The considerations of § 1 and § 3 (together with a few easy consequences of them) show a hierarchy among some assertions concerning two automata $\mathbf{A}_1, \mathbf{A}_2$. This hierarchy is illustrated in Table 1.

§ 4. Codes and pre-codes

4.1. The considerations of Section 3.4 show that it suffices to pay our attention to standard automata when our aim is to view the automata up to isomorphy. In Sections 4.2—4.4 a formal construction yielding all standard automata will be established. First we define a particular kind of tables — named pre-codes and codes — axiomatically, then we assign an automaton to each code (i.e., maximal pre-code).

The idea of our construction is:

first we take the initial state a_1 and its output sign y_1,

afterwards, we build up the other states $a_2, a_3, \ldots$ consecutively, together with their output signs and the transitions corresponding to the distinguished edges (cf. Section 3.3),

finally, we determine the remaining transitions $\delta(a_i, x^{(k)})$ consecutively, according to the lexicographic order of the pairs (i, k).

4.2. Denote by $\mathbf{N}_i^j$ the set of the integers $i, i+1, i+2, \ldots, j-1, j$ (where $i \leqq j$). Evidently, $|\mathbf{N}_i^j| = j-i+1$.

The sextuple $\mathbf{D}=(r, s, \beta, \gamma, \mu, \varphi)$ is called a *pre-code* if it satisfies the following postulates (I)—(VIII):

(I) r, s are non-negative integers; $\beta, \gamma, \mu, \varphi$ are functions.

(II) The domains of $\beta, \gamma, \mu, \varphi$ are $\mathbf{N}_2^{r+s+1}, \mathbf{N}_2^{r+s+1}, \mathbf{N}_1^{r+1}, \mathbf{N}_{r+2}^{r+s+1}$, respectively.

(III) The target[9] of each of β, μ, φ is $\mathbf{N}_1^{r+1}$.

(IV) The target of γ is $\mathbf{N}_1^n$.

(V) We have $\beta(2)=1$. If $i \in \mathbf{N}_3^{r+1}$, then (a) & ((b)$\vee$(c)) holds where the meaning of the statements (a), (b), (c) is:

(a) $\beta(i-1) \leqq \beta(i) < i$,

(b) $\beta(i-1) < \beta(i)$,

(c) $\gamma(i-1) < \gamma(i)$.

[9] By the target T_f of a function f, the set of its possible values is meant. The range R_f is a subset of T_f (R_f consists of the values that are actually taken by f).

(VI) If $i\in\mathbf{N}_1^{r+1}$, then

$$\mu(i)-1\in\{0,\mu(1),\mu(2),\dots,\mu(i-1)\}.$$

(VII) If $i\in\mathbf{N}_{r+2}^{r+s+1}$, then $(\beta(i),\gamma(i))$ is the lexicographically smallest pair fulfilling

$$j\in\mathbf{N}_2^{i-1}\Rightarrow(\beta(i)\neq\beta(j))\vee(\gamma(i)\neq\gamma(j))$$

for every j.

(VIII) If $i\in\mathbf{N}_{r+2}^{r+s+1}$, then either $\varphi(i)=1$ or (d) & ((e)$\vee$(f)) holds where the meaning of the statements (d), (e), (f) is:

(d) $\beta(\varphi(i))\leqq\beta(i)$,

(e) $\beta(\varphi(i))<\beta(i)$,

(f) $\gamma(\varphi(i))<\gamma(i)$.

The definition of the pre-code is finished.

Any pre-code can be illustrated by a table each row of which contains the values i, $\beta(i)$, $\gamma(i)$, $\mu(i)$, $\varphi(i)$. (See Table 2 as an example.)

i	$\beta(i)$	$\gamma(i)$	$\mu(i)$	$\varphi(i)$
1	—	—	1	—
2	1	1	2	—
3	1	2	1	—
4	2	1	3	—
5	2	2	2	—
6	3	2	4	—
7	3	1	—	4
8	4	1	—	2
9	4	2	—	1
10	5	1	—	3
11	5	2	—	5
12	6	1	—	6
13	6	2	—	1

Table 2

An example for code (with $n=2$; $r=5$, $s=7$)

PROPOSITION 13. *If the formulae* $i\in\mathbf{N}_2^{r+s+1}$, $j\in\mathbf{N}_2^{r+s+1}$, $\beta(i)=\beta(j)$, $\gamma(i)=\gamma(j)$ *are valid (for a pre-code), then* $i=j$.

PROOF. We want to verify that $2 \leqq j < i \leqq r+s+1$ implies

$$\beta(i) \neq \beta(j) \vee \gamma(i) \neq \gamma(j). \tag{4.1}$$

Case 1. $i \leqq r+1$. If $\beta(j)=\beta(i)$, then we get

$$\beta(i) = \beta(i-1) = \beta(i-2) = \ldots = \beta(j)$$

and

$$\gamma(i) > \gamma(i-1) > \gamma(i-2) > \ldots > \gamma(j)$$

by an iterated application of (V), hence $\gamma(i) \neq \gamma(j)$.

Case 2. $i \geqq r+2$. Then (4.1) follows from (VII) immediately. □

4.3. Let $\mathbf{D}_1=(r_1, s_1, \beta_1, \gamma_1, \mu_1, \varphi_1)$ and $\mathbf{D}_2=(r_2, s_2, \beta_2, \gamma_2, \mu_2, \varphi_2)$ be two pre-codes. Let the relation $\mathbf{D}_1 < \mathbf{D}_2$ be true if

$$(r_1 < r_2 \ \& \ s_1 = 0) \vee (r_1 = r_2 \ \& \ s_1 < s_2)$$

and $\beta_2, \gamma_2, \mu_2, \varphi_2$ is an extension of $\beta_1, \gamma_1, \mu_1, \varphi_1$, respectively. This relation $<$ is a partial ordering of all pre-codes (n is fixed). If $\mathbf{D}_1 < \mathbf{D}_2$ and $r_1+s_1+1=r_2+s_2$ hold, then we write $\mathbf{D}_1 \prec \mathbf{D}_2$ (this means that we get $\mathbf{D}_1$ if we delete the last row of $\mathbf{D}_2$).

If $\mathbf{D}_1$ is a pre-code and no pre-code $\mathbf{D}_2$ with $\mathbf{D}_1 < \mathbf{D}_2$ exists, then $\mathbf{D}_1$ is called a *code*. (The pre-code given by Table 2 is a code.)

REMARK. It can be shown that $s \leqq rn+n-r$ is valid for each pre-code; equality holds for the codes and only for them.

4.4. Let $\mathbf{C}=(r, s, \beta, \gamma, \mu, \varphi)$ be an arbitrary code. Determine an automaton

$$\psi(\mathbf{C}) = \mathbf{A} = (A, X, Y, \delta, \lambda, a^*)$$

by the following rules (a)—(e):

(a) $A=\{a_1, a_2, \ldots, a_{r+1}\}$.

(b) (The initial state of $\mathbf{A}$ is a_1, i.e.,) $a^*=a_1$.

(c) $Y=\{y_1, y_2, \ldots, y_t\}$ where

$$t = \max\big(\mu(1), \mu(2), \ldots, \mu(r+1)\big).$$

(d) For each $i (\in \mathbf{N}_2^{r+s+1})$ let

$$\delta(a_{\beta(i)}, x^{(\gamma(i))}) = \begin{cases} a_i & \text{if } \ i \leqq r+1, \\ a_{\varphi(i)} & \text{if } \ i \geqq r+2 \end{cases}$$

be true.

(e) For each $i (\in \mathbf{N}_2^{r+1})$ let

$$\lambda(a_i) = y_{\mu(i)}$$

hold.

REMARK. The transition function δ is consistently and totally defined because (VII) was stipulated, Proposition 13 holds and $\mathbf{C}$ is a code.

REMARK. If $\mathbf{D} < \mathbf{C}$ is valid for a pre-code $\mathbf{D}$, then we can assign to $\mathbf{D}$ a partial sub-automaton of $\psi(\mathbf{C})$ in a natural manner.

PROPOSITION 14. *$\psi(\mathbf{C})$ is a standard automaton for each code* $\mathbf{C}$.

Sketch of the PROOF. Let a_i be a state of $\psi(\mathbf{C})$, consider the word

$$p = x^{(\gamma(\beta^{k-1}(i)))}x^{(\gamma(\beta^{k-2}(i)))}\dots x^{(\gamma(\beta(i)))}x^{(\gamma(i))}$$

where $\beta^h(i)$ stands for

$$\overset{1}{\breve{\beta}}(\overset{2}{\breve{\beta}}\dots\overset{h}{\breve{\beta}}(i)\dots)$$

and k is the (unique) number fulfilling $\beta^k(i)=1$. We can show that p equals to $\varepsilon(a_i)$ and $\varepsilon(a_i) \lessdot \varepsilon(a_j)$ is equivalent to $i<j$. This means that the first property defining the standard automata (in Section 3.4) is valid. The second property follows from the postulate (VI). □

PROPOSITION 15. *To each standard automaton* $\mathbf{A}$, *there exists precisely one code* $\mathbf{C}$ *with* $\psi(\mathbf{C})=\mathbf{A}$.

Sketch of the PROOF. Starting with $\mathbf{A}=(A, X, Y, \delta, \lambda, a_1)$, we can establish an assignment ψ^{-1} such that ψ^{-1} yields a code and ψ, ψ^{-1} are inverse of each other.

$\psi^{-1}(\mathbf{A})$ is introduced in the following way:

(A) Denote $|A|-1$ by r, let s be $rn+n-r$.

(B) For an arbitrary i $(2 \leqq i \leqq r+1)$, let $\varepsilon(a_i)$ be written in the form $p'x^{(m)}$ $(p' \in F(X),\ x^{(m)} \in X)$. Let

$\beta(i)$ be the number satisfying $a_{\beta(i)}=\delta(a_1, p')$,

$\gamma(i)$ be m,

$\mu(i)$ be the number fulfilling $y_{\mu(i)}=\lambda(a_i)$.

(C) For an arbitrary i $(r+2 \leqq i \leqq r+s+1)$, $\beta(i)$ and $\gamma(i)$ are uniquely determined by Postulate (VII). Let $\varphi(i)$ be the number for which

$$a_{\varphi(i)} = \delta(a_{\beta(i)}, x^{(\gamma(i))}).$$

It can be shown that the rules (A), (B), (C) determine a code $\psi^{-1}(\mathbf{A})$, and $\psi(\psi^{-1}(\mathbf{A}))=\mathbf{A}$, $\psi^{-1}(\psi(\mathbf{C}))=\mathbf{C}$ are always valid. □

Fig. 3 shows the automaton $\psi(\mathbf{C})$ when $\mathbf{C}$ is the code given by Table 2.

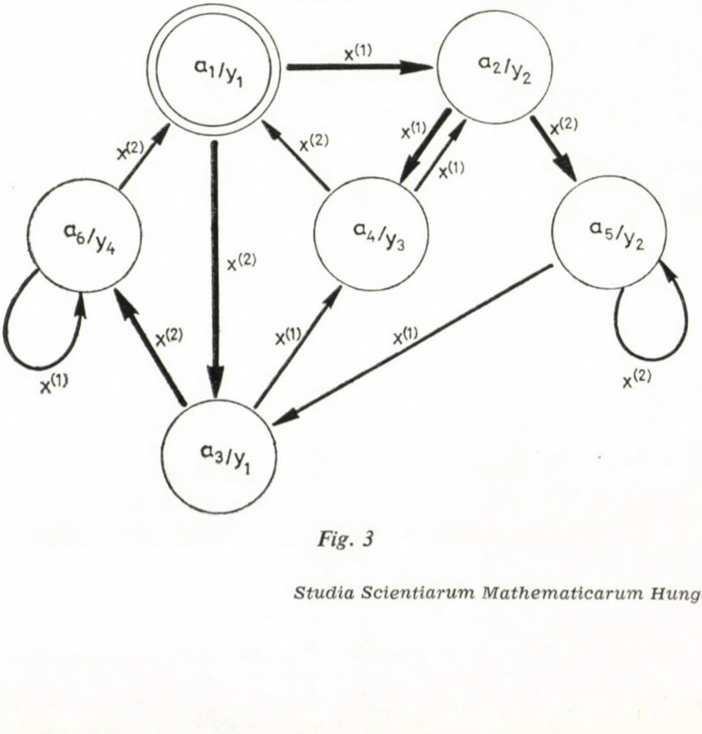


Fig. 3

§ 5. Notions of complexity

5.1. As we have seen in § 4, the codes give a constructive description of all finite Moore automata. Recall the second version of the basic problem (see Section 2.1). From the view point of this problem, the notion of code is too extended. Therefore, we want to restrict our attention to the codes **C** for which the automaton $\psi(\mathbf{C})$ is simple.

For this aim, we are going to introduce some complexity notions.

5.2. Let a, b be two states of a finite Moore automaton **A**. By the *distinguishability number* $\omega(a, b)$ of these states, we understand the length $L(p)$ of a shortest word p such that

$$\lambda(\delta(a, p)) \neq \lambda(\delta(b, p)). \tag{5.1}$$

(Of course, $\omega(a, b)=\infty$ if $\lambda(\delta(a, p))=\lambda(\delta(b, p))$ for every $p(\in F(X))$.) Evidently, $\omega(a, a)=\infty$ and $\omega(a, b)=\omega(b, a)$ are universally true.

The maximum

$$\max \omega(a, b)$$

is called the *complexity* $\Omega_A(\mathbf{A})$ of the automaton **A** where (a, b) runs through all pairs of states of **A** such that $a \neq b$. ($\Omega_A(\mathbf{A})=0$ if **A** has only one state.)

The *complexity* $\Omega_C(\mathbf{C})$ of a code is defined by

$$\Omega_C(\mathbf{C}) = \Omega_A(\psi(\mathbf{C})).$$

5.3.

LEMMA 2. *If a, b, c are arbitrary states of a finite automaton, then*

$$\omega(b, c) \geqq \min(\omega(a, b),\ \omega(a, c)).$$

PROOF. Let p be a shortest word such that

$$\lambda(\delta(b, p)) \neq \lambda(\delta(c, p)).$$

Denote $\lambda(\delta(a, p))$ by y_a, the notations y_b and y_c are meant analogously. If $y_a \neq y_b$, then $\omega(a, b) \leqq \omega(b, c)$. If $y_a \neq y_c$, then $\omega(a, c) \leqq \omega(b, c)$. □

PROPOSITION 16. *Let a_1, a_2, a_3 be arbitrary states of a finite Moore automaton. There exists a permutation π of the set $\{1, 2, 3\}$ such that*

$$\omega(a_{\pi(1)}, a_{\pi(2)}) = \omega(a_{\pi(2)}, a_{\pi(3)}) \leqq \omega(a_{\pi(1)}, a_{\pi(3)}).$$

PROOF. Obviously, there is a π fulfilling

$$\omega(a_{\pi(1)}, a_{\pi(2)}) \leqq \omega(a_{\pi(2)}, a_{\pi(3)}) \leqq \omega(a_{\pi(1)}, a_{\pi(3)}).$$

If $\omega(a_{\pi(1)}, a_{\pi(2)}) < \omega(a_{\pi(2)}, a_{\pi(3)})$ were true, then we should get a contradiction to Lemma 2 (with $a=a_{\pi(3)}, b=a_{\pi(2)}, c=a_{\pi(1)}$). □

REMARK. Let k be a non-negative integer. Define the relation $R_k(a, b)$ so that it holds if $\omega(a, b) \geqq k$. Then R_k is an equivalence for each k; if we form the sequence $R_0, R_1, R_2, \ldots$, then we do substantially the same as if we should minimize the automaton in the customary sense.

5.4.

LEMMA 3. *Let b, c be two states of a finite automaton* **A**. $\omega(b, c)=\infty$ *if and only if there is a homomorphism* (α_A, α_Y) *of* **A** *such that* $b^{\alpha_A}=c^{\alpha_A}$.

Sketch of the PROOF.

Necessity. Suppose $\omega(b, c)=\infty$. Introduce a binary relation ϱ (depending on b, c) in the set of states of **A** by what follows: $\varrho(a, a^+)$ is true precisely when there exist a sequence of states

$$(a =) a_0, a_1, a_2, \dots, a_k (= a^+)$$

and a sequence of words

$$p_1, p_2, \dots, p_k$$

(k is common) such that

$$\{\delta(b, p_i),\ \delta(c, p_i)\} = \{a_{i-1}, a_i\}$$

holds for every choice of i $(1 \leqq i \leqq k)$. We can show that ϱ is an equivalence relation, thus it determines a partition R of the set of states. The factor automaton with respect to R can be introduced in a consequent manner[10]; hence the pair (α_A, α_Y) where α_A is the natural assignment associated to the partition and α_Y is the identical permutation of the output alphabet is a homomorphism.

Sufficiency. Assume that $b^{\alpha_A}=c^{\alpha_A}$ for a homomorphism (α_A, α_Y). Then

$$\lambda'\big((\delta(b, p))^{\alpha_A}\big) = \lambda'\big(\delta'(b^{\alpha_A}, p)\big) = \lambda'\big(\delta'(c^{\alpha_A}, p)\big) = \lambda'\big((\delta(c, p))^{\alpha_A}\big)$$

(where the transition function δ' and the output function λ' are meant in the homomorphic image) for every choice of p, consequently

$$\lambda\big(\delta(b, p)\big) = \lambda\big(\delta(c, p)\big)$$

(since α_Y is a bijection). □

PROPOSITION 17. *Let* **A** *be a (finite) Moore automaton.* **A** *is simple if and only if* $\Omega_A(\mathbf{A})<\infty$.

PROOF. Consider the following five assertions:

(i) **A** is not reduced;

(ii) **A** has a homomorphism which is not an isomorphism;

(iii) **A** has two different states b, c and a homomorphism (α_A, α_Y) such that $b^{\alpha_A}=c^{\alpha_A}$;

(iv) **A** has two different states b, c such that $\omega(b, c)=\infty$;

(v) $\Omega_A(\mathbf{A})=\infty$.

The equivalences (i)⇔(ii), (ii)⇔(iii), (iv)⇔(v) are evident. The equivalence (iii)⇔(iv) was stated in Lemma 3. Hence (i) and (v) are equivalent. □

[10] The supposition $\omega(b, c)=\infty$ is utilized when we define the output function of the factor automaton.

PROPOSITION 18. *The following three statements are equivalent for a (finite) Moore automaton* $\mathbf{A}$:

(a) *The output function* λ *of* $\mathbf{A}$ *is bijective.*

(b) $\Omega_A(\mathbf{A})=0$.

(c) $\Omega_C(\psi^{-1}(\mathbf{A}))=0$.

PROOF. The equivalence of (b) and (c) is trivial. The equivalence of (a) and (b) follows from the evident facts that the statements $\lambda(a)\neq\lambda(b)$ and $\omega(a, b)=0$ are equivalent (where a, b are different states of $\mathbf{A}$). □

5.5. By comparing the second version of the basic problem, Proposition 17 and the definition of $\Omega_C(\mathbf{C})$, we get a new equivalent formulation:

BASIC PROBLEM *(third version)*. Let a constructive description of all codes $\mathbf{C}$ satisfying $\Omega_C(\mathbf{C})<\infty$ be given.

§ 6. Some open problems

6.1. The third version of the basic problem seems to be preferable (in comparison to the first and second versions) for being attacked in the future. This opinion can be supported by the fact that any code $\mathbf{C}$ can be viewed as the last member of the sequence

$$\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3, \dots$$

where $\mathbf{D}_i$ is the pre-code consisting of the first, second, ..., i-th rows of $\mathbf{C}$, and this sequence is uniquely determined by $\mathbf{C}$. If we extend the complexity notion Ω_C to all pre-codes (in some natural manner), then we may hope that the complexity of an arbitrary code $\mathbf{C}$ can be established as the result of an algorithmic procedure the steps of which are associated to $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3, \dots$; one may also hope that this aspect can yield a sharp separation of the codes with finite complexity from the codes whose complexity is infinite.

6.2. We define the *complexity* $\Omega_C(\mathbf{D})$ of a pre-code $\mathbf{D}$ by

$$\Omega_C(\mathbf{D}) = \min \Omega_C(\mathbf{C})$$

where $\mathbf{C}$ runs through all codes $\mathbf{C}$ fulfilling $\mathbf{D}\leqq\mathbf{C}$.

Next some immediate consequences of this definition are formulated.

PROPOSITION 19. *If the pre-codes* $\mathbf{D}_1$ *and* $\mathbf{D}_2$ *satisfy* $\mathbf{D}_1<\mathbf{D}_2$, *then* $\Omega_C(\mathbf{D}_1)\leqq\leqq\Omega_C(\mathbf{D}_2)$. □

PROPOSITION 20. *To each pre-code* $\mathbf{D}$ *there exists a code* $\mathbf{C}$ *such that*

$$\mathbf{D} \leqq \mathbf{C} \quad \& \quad \Omega_C(\mathbf{D}) = \Omega_C(\mathbf{C}). \quad \square$$

PROPOSITION 21. *If a pre-code* $\mathbf{D}_1$ *is not a code, then there exists a pre-code* $\mathbf{D}_2$ *such that*

$$\mathbf{D}_1 \prec \mathbf{D}_2 \quad \& \quad \Omega_C(\mathbf{D}_1) = \Omega_C(\mathbf{D}_2). \quad \square$$

6.3. The solution of each of the following four problems (more precisely: the algorithmic solution of Problems 1, 2, and a sufficiently constructive determination of the function mentioned in Problem 3) would be a useful contribution towards elaborating a theory for the elucidation of the third version of our basic problem.

PROBLEM 1. Let $\mathbf{D}$ be a pre-code with finite complexity. Let a code $\mathbf{C}$ be constructed which satisfies

$$\mathbf{D} \leqq \mathbf{C} \quad \& \quad \Omega_C(\mathbf{D}) = \Omega_C(\mathbf{C}).$$

PROBLEM 2. Let $\mathbf{D}_1$ be a pre-code with finite complexity. Let $\mathbf{D}_2$ run through all pre-codes such that $\mathbf{D}_1 \prec \mathbf{D}_2$. Let us partition the set of the $\mathbf{D}_2$'s into three classes concerning that where is sharp inequality in the formula

$$\Omega_C(\mathbf{D}_1) \leqq \Omega_C(\mathbf{D}_2) \leqq \infty.$$

PROBLEM 3. Let all the pairs $(\mathbf{D}_1, \mathbf{D}_2)$ be considered where $\mathbf{D}_1$ and $\mathbf{D}_2$ are precodes with finite complexity and they fulfil $\mathbf{D}_1 \prec \mathbf{D}_2$. Either determine the maximum of the differences

$$\Omega_C(\mathbf{D}_2) - \Omega_C(\mathbf{D}_1)$$

(as a function of $n = |X|$) or prove that the set of these differences is unbounded.

PROBLEM 4. Let the set of the complexities $\Omega_C(\mathbf{D})$ be considered where $\mathbf{D}$ runs through all pre-codes with $s=0$. Is this set equal to the set of all non-negative integers (without ∞)?[11]

The exposed four problems are (more or less) obviously related to each other. I am not able to predict which of them would occur earlier, which would occur later in an expected theory answering to these questions.

Appendix

(Some bibliographical notes [12])

Corollary 1 and the Remark to it are due to NERODE and MYHILL essentially, see § 2.4 in [7].

More detailed presentation of some fundamental notions and assertions touched here can be found in some parts of [7] and in [1].

The basic idea of §§ 3—4 of the present paper is similar to the basic idea of [2] (but the purpose of [2] differs from our present purposes). In [3] I have tried to begin to explain the same topics as now; my attitude, however, has considerably changed in the meantime.

[11] (Added in proof.) After completing the present paper, the author succeeded in solving Problem 4 affirmatively. This result will appear in Vol. 5 of the *Acta Cybernetica.*

[12] A few bibliographical remarks were already done in course of §§ 1—6. These are not recapitulated here.

REFERENCES

[1] Ádám, A., Automata-leképezések, félcsoportok, automaták (Automaton mappings, semigroups, automata), *Mat. Lapok* **19** (1968), 327—343.

[2] Ádám, A., A description of the finite right-congruences of finitely generated free semigroups, *Period. Math. Hungar.* **1** (1971), 135—144.

[3] Ádám, A., On some aspects of the algebraic description of automaton mappings, *Acta Cybernet.* **2** (1973), 1—21.

[4] Andréka, H.—Horváth, S.—Németi, I., Notes on maximal congruence relations, automata and related topics, *Acta Cybernet.* **2** (1973), 71—88.

[5] Clifford, A. H.—Preston, G. B., *The algebraic theory of semigroups, I*, Amer. Math. Soc., Providence, 1961.

[5a] Клиффорд, А. Х.—Престон, Г. Б. *Алгебраическая теория полугрупп, I,* Мир, Москва, 1972.

[6] Clifford, A. H.—Preston, G. B., *The algebraic theory of semigroups, II*, Amer. Math. Soc., Providence, 1967.

[6a] Клиффорд, А. Х.—Престон, Г. Б. *Алгебраическая теория полугрупп, II*, Мир, Москва, 1972.

[7] Gécseg, F.—Peák, I., *Algebraic theory of automata,* Akadémiai Kiadó, Budapest, 1972.

[8] Havel, I. M., The theory of regular expressions, I, *Kybernetika* (Praha) **5** (1969), 400—419.

[9] Havel, I. M., The theory of regular expressions, II, *Kybernetika* (Praha) **5** (1969), 520—544.

[10] Kleene, S. C., Representation of events in nerve nets and finite automata, *Automata Studies,* Princeton, 1956, 3—41.

[10a] Клини, С. К., Представление событий в нервных сетях и конечных автоматах, *Автоматы,* Москва, 1956, 15—67.

[11] Simon, I., *Hierarchies of events with dot-depth one* (Dissertation), Univ. of Waterloo, Canada, 1972.

[12] Simon, I., Piecewise testable events, *Automata Theory and Formal Languages (Proc. Conf. Kaiserslautern, 1975),* Springer-Verlag, Berlin, 1975, 214—222.

Mathematical Institute of the Hungarian Academy of Sciences
H—1053 Budapest, Reáltanoda u. 13—15

(Received April 25, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 125—132.

ON THE PERMEABILITY OF LAYERS OF DISCS

by

A. FLORIAN

To my friend Prof. L. Fejes Tóth on his 65th birthday

By a *layer of discs* we mean a set of convex compact discs which do not overlap and which lie between two parallel lines. We shall always consider the smallest parallel strip containing the layer. Its width w is called the *width of the layer.*

We shall consider continuous rectifiable curves lying completely within the parallel strip. Such a curve *crosses* the layer if it connects the two boundary lines without containing any interior point of a disc. According to L. FEJES TÓTH [2], the *permeability of the layer* is defined by

$$p=\frac{w}{\inf \Lambda},$$

where w denotes the width of the layer and Λ the length of a curve crossing the layer. Obviously, we have $0<p\leqq 1$.

The permeability of a layer consisting of circles or other kinds of convex discs was investigated in the papers [1, 2, 3, 5]. Our object is to improve two results contained in [2] and the theorem proved in [5].

§ 1

In this paragraph, we solve the problem of determining the least permeable layer of prescribed width consisting of congruent circles of given diameter.

THEOREM 1. *Let w be the width of a layer consisting of congruent circles of diameter d. Then the permeability p of the layer satisfies*

$$(1)\qquad p\geqq\frac{w}{d}\left(n\frac{\pi}{3}+1+\arcsin\frac{w-d-n\frac{d}{2}\sqrt{3}}{d}\right)^{-1},$$

where $n=\left[\frac{2}{\sqrt{3}}\frac{w-d}{d}\right]$.

Let w be an arbitrary number greater than d. Consider an equliateral triangular section T of a densest lattice packing of circles of diameter d, consisting of $(n+1)+n+\ldots+1$ circles. Then the only case in which equality is attained in (1) may be described as follows: triangular blocks of circles congruent to T are based alternately on the upper and the lower bordering line of the strip, such that each

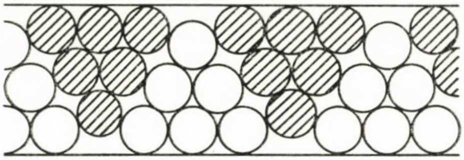


Fig. 1a

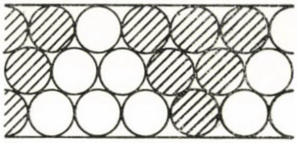


Fig. 1b

block touches its two neighbours (Fig. 1a, b with $n=2$). J. MOLNÁR conjectured this arrangement to be a densest packing of circles of diameter d contained in a parallel strip of width w. It is easy to see that this is true for $d\leqq w\leqq\left(1+\frac{\sqrt{3}}{2}\right)d$. Recently, G. KERTÉSZ succeeded in proving the conjecture in the case when $\left(1+\frac{\sqrt{3}}{2}\right)d<w\leqq(1+\sqrt{2})d$. Molnár's conjecture proves to be true also in the cases when $w=\left(1+n\frac{\sqrt{3}}{2}\right)d$, $n=2, 3, \dots$. This is a consequence of GROEMER's inequality [4]

$$\sqrt{12}n\leqq F-\frac{2-\sqrt{3}}{2}U+\sqrt{12}-\pi(\sqrt{3}-1),$$

where F and U denote the area and the perimeter of a convex disc containing a packing of n unit circles.

Since

$$d\left(n\frac{\pi}{3}+1+\arcsin\frac{w-d-n\frac{d}{2}\sqrt{3}}{d}\right)=\frac{2\pi}{\sqrt{27}}(w-d)+d+$$

$$+d\left\{\arcsin\left(\frac{\sqrt{3}}{2}\left(\frac{2}{\sqrt{3}}\frac{w-d}{d}-n\right)\right)-\frac{\pi}{3}\left(\frac{2}{\sqrt{3}}\frac{w-d}{d}-n\right)\right\}$$

and

$$\arcsin(xy)<x\arcsin y\qquad(0<x<1,\ 0<y<1),$$

(1) implies the weaker inequality

$$p\geqq\frac{w}{\frac{2\pi}{\sqrt{27}}(w-d)+d}\tag{2}$$

proved by Fejes Tóth [2]. In (2) equality is claimed if and only if, $\frac{2}{\sqrt{3}}\frac{w-d}{d}$ is an integer n and the circles form $n+1$ consecutive complete rows of a densest lattice packing of circles (Fig. 1b with $n=2$).

For the proof of Theorem 1 we modify Fejes Tóth's proof of (2). Let the layer of circles lie in a vertical plane so that the two boundary lines are horizontal. We now describe a curve crossing the layer. From a point U of the upper boundary line, a point P moves vertically downwards into the layer. Whenever striking a circle, P moves downwards along the circle to the nearest end-point of the horizontal diameter. (If P hits the 'top' of a circle, it is supposed to deviate either to the right or to the left.) P leaves the circle at this point and again moves vertically downwards. P covers a curve K composed of straight segments and circular arcs of length $\leqq \pi d/4$ and terminates at a point L of the lower boundary line. We choose the point of departure U so that P, before striking against a circle, touches another circle.

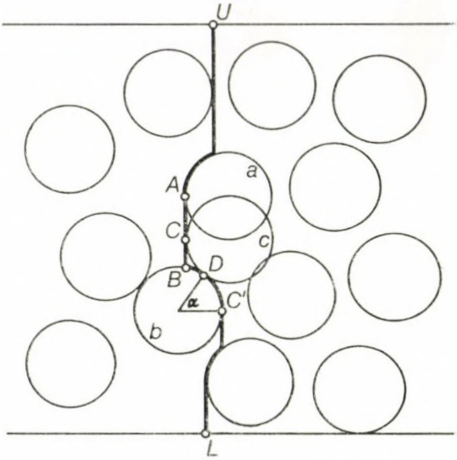


Fig. 2

Now let us shorten the curve $K = UL$ by rounding off its corners in the following manner: let AB be a straight segment of K connecting the circle a with the circle b so that the straight line AB touches a and cuts b. Let c be a circle of diameter d touching the segment AB at a point C and the circle b at a point D belonging to K. We replace the segment CBD of K by the circular arc CD. Rounding off all corners of K in this way, we obtain a new curve K' (Fig. 2) crossing the layer of circles.

We proceed to show that the length Λ' of K' satisfies the inequality

$$\Lambda' \leqq d\left(n\frac{\pi}{3}+1+\arcsin\frac{w-d-n\frac{d}{2}\sqrt{3}}{d}\right). \tag{3}$$

Reflecting the circular arc CD in D we obtain a new circular arc DC' of central angle α also belonging to K'. Denoting the number of arcs of the curve K' of type DC' by m and the total length of the straight segments contained in K' by s, we have

$$\Lambda' = d(\alpha_1+\ldots+\alpha_m)+s \tag{4}$$

and

$$w = d(\sin\alpha_1+\ldots+\sin\alpha_m)+s. \tag{5}$$

According to the construction we have

$$0 \leqq \alpha_i \leqq \frac{\pi}{3} \qquad (i = 1, \ldots, m), \tag{6}$$

$$s \geqq d.$$

We shall now determine the maximum of Λ' for a given value $w > d$ and some positive integer m under the conditions (5) and (6). Putting

$$\sin\alpha_i = x_i \qquad (i = 1, \ldots, m)$$

and using the strict convexity of $\arcsin x$ for $0 \leqq x \leqq 1$, we see that, in the case when Λ' attains its maximum, at most one of the x_i's, say x_m, is >0 and $<\frac{1}{2}\sqrt{3}$. Since $\arcsin x - x$ is strictly increasing for $0 \leqq x \leqq 1$, we have now $s=d$. Therefore, Λ' attains its maximum under the conditions (5) and (6) if and only if, for some $m \geqq 1$,

$$\alpha_1 = \ldots = \alpha_{m-1} = \frac{\pi}{3}, \quad 0 \leqq \alpha_m < \frac{\pi}{3}, \tag{7}$$

$$s = d.$$

(7) implies that $m-1=n$, and (3) is proved.

From (3) it follows the inequality

$$\inf \Lambda \leqq d\left(n\frac{\pi}{3}+1+\arcsin\frac{w-d-n\frac{d}{2}\sqrt{3}}{d}\right),$$

which completes the proof of (1).

Now we suppose that $w>d$ and that the permeability of the layer is equal to the right side of (1). If Λ' denotes the length of any curve K' constructed as above, then (3) holds with equality. Thus (7) is fulfilled for all admissible positions of the point U. This fact implies that the layer consists of triangular blocks of circles as described above. Conversely, it is easy to show that the permeability of a layer of this kind is given by the right side of (1).

§ 2

We shall now show that Theorem 1 continues to hold if we replace the circles by translates of a disc of constant width d.

THEOREM 2. *Let w be the width of a layer of translates of a disc of constant width d. Then the permeability p satisfies the inequality* (1).

The only cases in which equality is attained in (1) may be described similarly as for circles. Note that the minimum value of the permeability depends only on

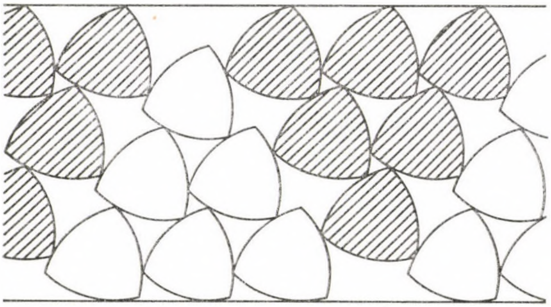


Fig. 3

w and d, but neither on the orientation nor on the shape of the disc. Fig. 3 exhibits a layer of translated Reuleaux triangles which has minimal permeability.

As in the case of circles, Theorem 2 implies the inequality (2), already proved by HORTOBÁGYI [5].

A disc of constant width is automatically strictly convex, what means, that its boundary does not contain straight segments. This and further properties of discs of constant width may be found in [6].

For the proof of Theorem 2 we need a lemma which is due to Hortobágyi [5].

LEMMA. *Let A_1A_2 and B_1B_2 be two diameters of a disc of constant width d. Then*

$$\widehat{A_1B_1}+\widehat{A_2B_2}=d\alpha,$$

where α is the angle between A_1A_2 and B_1B_2 (Fig. 4).

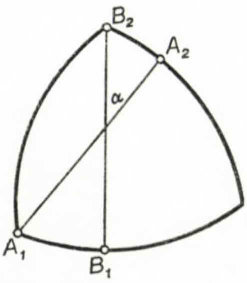


Fig. 4

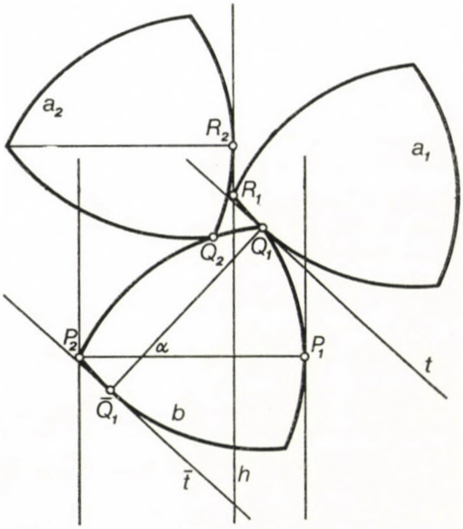


Fig. 5

The proof rests on the possibility of approximating an arbitrary disc of constant width d by discs of the same constant width whose boundaries are composed of circular arcs of diameter d (see [6]). It is easy to see that for such discs the lemma is true.

PROOF of Theorem 2. Let P_1P_2 be a horizontal diameter of a disc b of constant width d (Fig. 5). Let the line h intersect the segment P_1P_2 perpendicularly. Let a_i $(i=1, 2)$ be translates of b touching b and h at the points Q_i and R_i, respectively. Without loss of generality we suppose that $\overline{R_1P_1}\leqq\overline{R_1P_2}$; otherwise replace a_1 by a_2. Let t be a line through Q_1 separating b and a_1. The parallel supporting line $\bar{t}$ of b touches b at a point $\bar{Q}_1$, where $\overline{Q_1\bar{Q}_1}=d$. The translation $\bar{Q}_1\to Q_1$ carries b into a_1 and P_2 into R_1 so that $P_2R_1\parallel\bar{Q}_1Q_1$ and $\overline{P_2R_1}=d$. Let α

be the acute angle between P_1P_2 and $\bar{Q}_1Q_1$. The supposition $\overline{R_1P_1} \leqq \overline{R_1P_2}$ implies that R_1P_1 is a shortest side of the triangle $P_1P_2R_1$, so that $\alpha \leqq \frac{\pi}{3}$. The difference of levels between R_1 and P_2, i.e., the distance of the lines through these points parallel to the strip, is $d \sin \alpha$. Since $\overset{\frown}{R_1Q_1} = \overset{\frown}{P_2\bar{Q}_1}$, we have, by the lemma,

$$\overset{\frown}{R_1Q_1} + \overset{\frown}{Q_1P_1} = \overset{\frown}{P_2\bar{Q}_1} + \overset{\frown}{\bar{Q}_1P_1} = d\alpha.$$

Similarly as in § 1, the layer can be crossed by a curve K' of length Λ', given by (4), where the variables α_i $(i=1, \dots, m)$ and s are subject to the conditions (5) and (6). The further course of the proof coincides with that of Theorem 1.

§ 3

We now consider translates of a strictly convex disc having an axis of symmetry perpendicular to the layer (shortly: an axis). We recall that a convex disc whose boundary does not contain straight segments is said to be strictly convex. Let a and b be two such discs touching each other at a point S so that none of them is cut by the axis of the other. Let A be the boundary point of a nearest to the axis of b and B the boundary point of b nearest to the axis of a. Applying a construction analogous to that of K' to a layer of translates of a, we obtain a path consisting of vertical segments and pairs of arcs like AS and SB, which crosses the layer. We introduce some notations.

$(\mathcal{N})$ Let λ be the length of the arc ASB and δ the difference of levels between the points A and B. We fix the disc b and consider the function

$$\lambda = f(\delta),$$

where $0 \leqq \delta \leqq \bar{\delta}$ and $\bar{\delta}$ is determined by that position of the disc a in which it touches the axis of b.

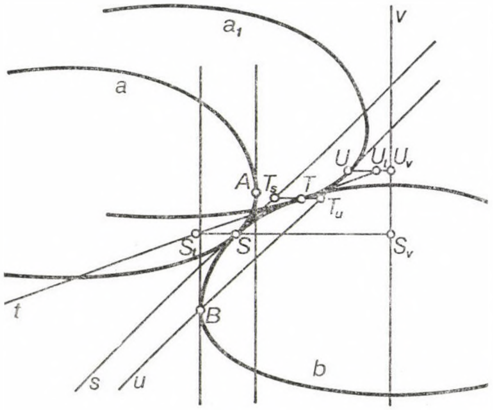


Fig. 6

Let us translate a into a position a_1 touching b at a point T, so that $\delta < \delta_1 \leqq \bar{\delta}$ (Fig. 6). Since $\lambda_1 = f(\delta_1) > \lambda$, $f(\delta)$ is strictly increasing for $0 \leqq \delta \leqq \bar{\delta}$. We shall prove that for $0 \leqq \delta \leqq \bar{\delta}$

(i) $f(\delta)$ is strictly convex,
(ii) $f(\delta) - \delta$ is strictly increasing.

Let s be a line separating a and b. Let the translation $a \to a_1$ carry S and s into U and u. We denote the intersection of the horizontal line through a point X with a line y by X_y and consider the points S_v, T_s, T_u and U_v, where v is a vertical line. Ob-serving that

$$\lambda_1-\lambda=\widehat{UT}+\widehat{TS}>\overline{UT}+\overline{TS}>\overline{UT_u}+\overline{T_sS}$$

and

$$\delta_1-\delta=\overline{U_vS_v}$$

and denoting the angle between a line y and v by φ_y, we have

$$\frac{\lambda_1-\lambda}{\delta_1-\delta}>\frac{\overline{UT_u}+\overline{T_sS}}{\overline{U_vS_v}}=\sec\varphi_s. \tag{8}$$

Similarly, let t be a line separating a_1 and b. Let u and s intersect t at U' and S', respectively. Then we obtain

$$\lambda_1-\lambda=\widehat{UT}+\widehat{TS}<\overline{UU'}+\overline{U'T}+\overline{TS'}+\overline{S'S}<\overline{U_tU'}+\overline{U'T}+$$
$$+\overline{TS'}+\overline{S'S_t}=\overline{U_tS_t}=\overline{U_vS_v}\sec\varphi_t.$$

Thus

$$\frac{\lambda_1-\lambda}{\delta_1-\delta}<\sec\varphi_t. \tag{9}$$

Let a_1 and a_2 be two translates of a satisfying $0\leqq\delta<\delta_1<\delta_2\leqq\bar{\delta}$. By (8), we have

$$\frac{\lambda_2-\lambda_1}{\delta_2-\delta_1}>\sec\varphi_t.$$

Combining this with (9) we obtain

$$\frac{\lambda_2-\lambda_1}{\delta_2-\delta_1}>\frac{\lambda_1-\lambda}{\delta_1-\delta},$$

which proves (i). In view of $\sec\varphi_s>1$, (8) implies (ii).

A precisely similar argument as in the proof of Theorem 1 yields

THEOREM 3. *Let w be the width of a layer consisting of translates of a strictly convex disc b having an axis of symmetry perpendicular to the layer. Let p be the permeability of the layer and d the width of b perpendicular to the layer. Then, using the notations ($\mathcal{N}$) and writing $\bar{\lambda}=f(\bar{\delta})$, we have*

$$p\geqq w\big(n\bar{\lambda}+d+f(w-d-n\bar{\delta})\big)^{-1}, \tag{10}$$

where $n=\left[\dfrac{w-d}{\bar{\delta}}\right]$.

The case in which (10) holds with equality may be described in the same way as for circles (Fig. 7).

It follows from (i), that $\lambda/\delta\leqq\bar{\lambda}/\bar{\delta}$. Using this, we obtain as a simple consequence of (10) the inequality

$$p\geqq\frac{w}{q(w-d)+d},\quad q=\frac{\bar{\lambda}}{\bar{\delta}},$$

which was previously proved by Fejes Tóth [2].

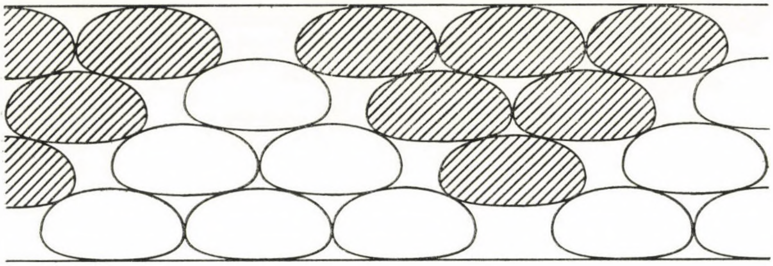


Fig. 7

REFERENCES

[1] Bollobás, B., Remarks to a paper of L. Fejes Tóth, *Studia Sci. Math. Hungar.* **3** (1968), 373—379.

[2] Fejes Tóth, L., On the permeability of a circle-layer, *Studia Sci. Math. Hungar.* **1** (1966), 5—10.

[3] Fejes Tóth, L., On the permeability of a layer of parallelograms, *Studia Sci. Math. Hungar.* **3** (1968), 195—200.

[4] Groemer, H., Über die Einlagerung von Kreisen in einen konvexen Bereich, *Math. Z.* **73** (1960), 285—294.

[5] Hortobágyi, I., Über die Durchlässigkeit einer aus Scheiben konstanter Breite bestehenden Schicht, *Studia Sci. Math. Hungar.* **11** (1976), 383—387.

[6] Jaglom, I. M.—Boltjanski, W. G., *Konvexe Figuren,* Berlin, 1956.

Mathematisches Institut der Universität Salzburg
A—5020 Salzburg, Petersbrunnstraße 19

(Received May 3, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 133—137.

STACK OF PANCAKES

by

ERVIN GYŐRI and GYÖRGY TURÁN

The following open problem was proposed in [1], [2]:

The chef in our place is sloppy, and when he prepares a stack of pancakes they come out all different sizes. Therefore, when I deliver them to a customer, on the way to the table I rearrange them (so that the smallest winds up on the top, and so on, down the largest on the bottom) by grabling several from the top and flipping them over, repeating this (varying the number I flip) as many times as necessary. If there are n pancakes, what is the maximum number of flips (as a function on n) that I shall ever have to use to rearrange them?

A mathematical formulation of this problem is the following:

Let $\pi = i_1 i_2 \dots i_n$ be a permutation of the number-set $\{1, 2, \dots, n\}$. Let admissible step be the reversing of the "end" of the sequence (permutation). (The ends of this sequence are the subsequences $i_{n-k} i_{n-k+1} \dots i_{n-1} i_n$ for $0 \leqq k \leqq n-1$.) This step is denoted by

$$i_1 i_2 \dots i_{n-k-1} \underline{i_{n-k} i_{n-k+1} \dots i_{n-1} i_n} \to i_1 i_2 \dots i_{n-k-1} i_n i_{n-1} \dots i_{n-k+1} i_{n-k}.$$

Let $f(\pi)$ be the minimal number of the admissible steps by means of which we get the permutation $1, 2, \dots, n$. Let

$$f(n) = \max_{\pi \in S_n} f(\pi).$$

What can one say on $f(n)$?

M. R. Garey, D. S. Johnson and S. Lin proved

$$n+1 \leqq f(n) \leqq 2n-6 \quad \text{for} \quad n \geqq 7.$$

We prove that

$$f(n) \leqq \frac{5}{3}(n+1) \quad \text{for arbitrary } n.$$

Definition. Suppose that the sequence is $i_1 i_2 \dots i_n$. We say that the subsequence $i_j i_{j+1} \dots i_{j+k-1} i_{j+k}$ $(1 \leqq j \leqq j+k \leqq n)$ is a chain if

$$i_{j+h} \equiv i_j + h \pmod n \quad 1 \leqq h \leqq n$$

or if

$$i_{j+h} \equiv i_j - h \pmod n \quad 1 \leqq h \leqq n.$$

(That is 1 and n may be neighbours in a chain.) The last element of the chain is the element i_{j+k}. The length of the chain is $k+1$. The chains of length one are called trivial chains, the chains of length at least two are called proper chains. The maximal chains are unique in a fixed sequence and the maximal chains constitute a partition of the sequence, obviously.

In this algorithm we gradually decrease the number of the maximal chains. We do four kinds of operations:

1st kind. At this kind of operation we perform one step, the number of the proper maximal chains does not change, the number of the trivial maximal chains decreases by one.

2nd kind. At this kind of operation we perform one step, the number of the proper maximal chains increases by one, the number of the trivial maximal chains decreases by two.

3rd kind. At this kind of operation we perform two steps, the number of the proper maximal chains decreases by one, the number of the trivial maximal chains does not change.

4th kind. At this kind of operation we perform four steps, the number of the proper maximal chains decreases by one, the number of the trivial maximal chains decreases by one.

(The concrete description of operations will be given at the case analysis of the algorithm.)

If the number of the maximal chains is one then only at most four steps have to be done to finish the algorithm.

Suppose that $i_1 i_2 \dots i_n$ be the sequence for the moment. Consider the maximal chain containing i_n. Let $i_{n-j} i_{n-j+1} \dots i_n$ be this maximal chain.

Case 1. $j=0$, that is the maximal chain containing i_n is trivial.

Let i_k be the unique element satisfying

$$i_k \equiv i_n + 1 \pmod n$$

and let i_h be the unique element satisfying

$$i_h \equiv i_n - 1 \pmod n.$$

Case 1.1. One of i_k and i_h is the last element of a maximal chain. E.g., i_k is the last element of a maximal chain.

Then the following operation has to be done:

$$i_1 i_2 \dots i_k \underline{i_{k+1} \dots i_{n-1} i_n} \to i_1 i_2 \dots i_k i_n i_{n-1} \dots i_{k+1}.$$

(If i_k belongs to a proper maximal chain then this is an operation of the 1st kind otherwise this is an operation of the 2nd kind.)

Case 1.2. Neither i_k nor i_h is the last element of a maximal chain. (Then they belong to proper maximal chains.)

In this case an operation of the 4th kind has to be done: (We may assume that $k<h$ because of the symmetry.)

$$i_1 i_2 \dots i_{k-1} i_k i_{k+1} \dots i_{h-1} \underline{i_h \dots i_{n-1} i_n} \rightarrow i_1 i_2 \dots i_{k-1} i_k i_{k+1} \dots i_{h-1} i_n \underline{i_{n-1} \dots i_h} \rightarrow$$

$$i_1 i_2 \dots i_{k-1} \underline{i_k i_{k+1} \dots i_{h-1} i_n i_h \dots i_{n-1}} \rightarrow i_1 i_2 \dots i_{k-1} i_{n-1} \dots i_h i_n \underline{i_{h-1} \dots i_{k+1} i_k} \rightarrow$$

$$i_1 i_2 \dots i_{k-1} i_{n-1} \dots i_h i_n i_k i_{k+1} \dots i_{h-1}.$$

Case 2. $j>0$, that is i_n belongs to a proper maximal chain. If the unique element i_k satisfying

$$i_k - i_n \equiv i_n - i_{n-1} \pmod n$$

is the last element of a maximal chain then the step

$$i_1 i_2 \dots i_k \underline{i_{k+1} \dots i_{n-1} i_n} \rightarrow i_1 i_2 \dots i_k i_n i_{n-1} \dots i_{k+1}$$

is either an operation of the 1st kind or a better operation (in the sense of the final analysis of the algorithm, see below) than an operation of the 3rd kind. Therefore we may assume that i_k is not the last element of any maximal chain. Let i_h be the unique element satisfying

$$i_{n-j+1} - i_{n-j} \equiv i_{n-j} - i_h \pmod n.$$

Case 2.1. i_h constitutes a trivial maximal chain.
Then an operation of the 4th kind has to be done:

Case 2.1.1. $k<h$.

$$i_1 i_2 \dots i_{k-1} i_k i_{k+1} \dots i_{h-1} \underline{i_h i_{h+1} \dots i_{n-j-1} i_{n-j} i_{n-j+1} \dots i_{n-1} i_n} \rightarrow$$

$$i_1 i_2 \dots i_{k-1} i_k i_{k+1} \dots i_{h-1} i_n i_{n-1} \dots i_{n-j+1} i_{n-j} \underline{i_{n-j-1} \dots i_{h+1} i_h} \rightarrow$$

$$i_1 i_2 \dots i_{k-1} \underline{i_k i_{k+1} \dots i_{h-1} i_n i_{n-1} \dots i_{n-j+1} i_{n-j} i_h i_{h+1} \dots i_{n-j-1}} \rightarrow$$

$$i_1 i_2 \dots i_{k-1} i_{n-j-1} \dots i_{h+1} i_h i_{n-j} i_{n-j+1} \dots i_{n-1} i_n \underline{i_{h-1} \dots i_{k+1} i_k} \rightarrow$$

$$i_1 i_2 \dots i_{k-1} i_{n-j-1} \dots i_{h+1} i_h i_{n-j} i_{n-j+1} \dots i_{n-1} i_n i_k i_{k+1} \dots i_{h-1}$$

Case 2.1.2. $k>h$.

$$i_1 i_2 \dots i_{h-1} \underline{i_h i_{h+1} \dots i_{k-1} i_k i_{k+1} \dots i_{n-j-1} i_{n-j} i_{n-j+1} \dots i_{n-1} i_n} \rightarrow$$

$$i_1 i_2 \dots i_{h-1} i_n i_{n-1} \dots i_{n-j+1} i_{n-j} \underline{i_{n-j-1} \dots i_{k+1} i_k i_{k-1} \dots i_{h+1} i_h} \rightarrow$$

$$i_1 i_2 \dots i_{h-1} \underline{i_n i_{n-1} \dots i_{n-j+1} i_{n-j} i_h i_{h+1} \dots i_{k-1} i_k i_{k+1} \dots i_{n-j-1}} \rightarrow$$

$$i_1 i_2 \dots i_{h-1} i_{n-j-1} \dots i_{k+1} i_k \underline{i_{k-1} \dots i_{h+1} i_h i_{n-j} i_{n-j+1} \dots i_{n-1} i_n} \rightarrow$$

$$i_1 i_2 \dots i_{h-1} i_{n-j-1} \dots i_{k+1} i_k i_n i_{n-1} \dots i_{n-j+1} i_{n-j} i_h i_{h+1} \dots i_{k-1}.$$

Case 2.2. i_h belongs to a proper maximal chain.
Then an operation of the 3rd kind has to be done.
Case 2.2.1. i_h belongs to the maximal chain $i_h i_{h+1} \dots i_{h+m}$ $(m \geqq 1, h+m \leqq n-j)$. (That is i_h is not a last element.)

Then the following operation has to be done:

$$i_1 i_2 \dots i_{h-1} \underline{i_h i_{h+1} \dots i_{h+m-1} i_{h+m} i_{h+m+1} \dots i_{n-j-1} i_{n-j} i_{n-j+1} \dots i_{n-1} i_n} \to$$

$$i_1 i_2 \dots i_{h-1} i_n i_{n-1} \dots i_{n-j+1} i_{n-j} \underline{i_{n-j-1} \dots i_{h+m+1} i_{h+m} i_{h+m-1} \dots i_{h+1} i_h} \to$$

$$i_1 i_2 \dots i_{h-1} i_n i_{n-1} \dots i_{n-j+1} i_{n-j} i_h i_{h+1} \dots i_{h+m-1} i_{h+m} i_{h+m+1} \dots i_{n-j-1}.$$

Case 2.2.2. i_h belongs to the maximal chain $i_{h-m} i_{h-m+1} \dots i_h$ $(m \geqq 1, h-m \geqq 1)$. (That is i_h is a last element.)

Then the following operation has to be done:

$$i_1 i_2 \dots i_{h-m-1} i_{h-m} i_{h-m+1} \dots i_{h-1} i_h i_{h+1} \dots i_{n-j-1} \underline{i_{n-j} i_{n-j+1} \dots i_{n-1} i_n} \to$$

$$i_1 i_2 \dots i_{h-m-1} i_{h-m} i_{h-m+1} \dots i_{h-1} i_h \underline{i_{h+1} \dots i_{n-j-1} i_n i_{n-1} \dots i_{n-j+1} i_{n-j}} \to$$

$$i_1 i_2 \dots i_{h-m-1} i_{h-m} i_{h-m+1} \dots i_{h-1} i_h i_{n-j} i_{n-j+1} \dots i_{n-1} i_n i_{n-j-1} \dots i_{h+1}.$$

If the number of the maximal chains is one then we can reach the natural order $1, 2, \dots, n$ by means of at most four steps. (It is almost trivial.)

PROPOSITION. *The number of the steps to reach the natural order* $1, 2, \dots, n$ *in the above-mentioned way is at most* $\frac{5}{3}(n+1)$.

PROOF. Let n_i denote the number of the operations of i-th kind. Then the number of the steps done during the above-mentioned algorithm is at most

$$s = n_1 + n_2 + 2n_3 + 4n_4 + 4.$$

Let p be the number of the proper maximal chains at the beginning of the algorithm and let q be the number of the trivial maximal chains at the beginning of the algorithm. Then obviously

$$p-1 = n_3 + n_4 - n_2, \quad q = 2n_2 + n_1 + n_4$$

because finally the number of the proper maximal chains will be one and the number of the trivial maximal chains will be zero. Hence

$$s = n_1 + n_2 + 2n_3 + 4n_4 + 4 \leqq \frac{5}{3} n_1 + n_2 + \frac{7}{3} n_3 + 4n_4 + 4 =$$

$$= \frac{5}{3}(n_1 + 2n_2 + n_4) + \frac{7}{3}(n_3 + n_4 - n_2) + 4 = \frac{5}{3} q + \frac{7}{3} p + \frac{5}{3} \leqq$$

$$\leqq \frac{5}{3}(2p + q + 1) \leqq \frac{5}{3}(n+1).$$

$(2p + q \leqq n$ because a proper maximal chain contains at least two elements.)

NOTE. We learned that independently of us W. H. GATES and C. H. PAPADIMITROU proved

$$\frac{17}{16} n \leqq f(n) \leqq \frac{5}{3}(n+1).$$

REFERENCES

[1] *Amer. Math. Monthly* **82** (1975), 1010.
[2] *Amer. Math. Monthly* **84 (1977),** 296.
[3] Gates, W. H.—Papadimitrou, C. H., Bounds for sorting by prefix reversal, *Discrete Mathematics* (to appear).

Mathematical Institute of the Hungarian Academy of Sciences
H—1053 Budapest, Reáltanoda u. 13—15
and
Bolyai Institute of the Attila József University
H—6720 Szeged, Aradi vértanúk tere 1

(Received May 4, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 139—142.

SOME PROBLEMS ON LATTICE AUTOMORPHISMS

by

L. BABAI

§ **1.** G. BIRKHOFF proved in 1949 [6] that *every group is isomorphic to the automorphism group of a distributive lattice.* For finite groups, the lattice obtained was finite but of large size, height and order-dimension, and this is necessarily so for *distributive* lattices with prescribed automorphism groups. The situation may be different if selecting our lattice with prescribed automorphism group from a larger lattice variety. The aim of this note is to raise interest in questions, relating automorphism groups of finite lattices in a variety to bounds on various parameters of the lattice.

c_1, c_2, etc. denote absolute constants.

§ **2. Size of the lattice.** For $\mathscr{V}$ a lattice variety and G a finite group, let $\alpha(G; \mathscr{V})$ denote the minimum number of elements of a lattice from $\mathscr{V}$ whose automorphism group is isomorphic to G. Let $\mathscr{L}$, $\mathscr{M}$ and $\mathscr{D}$ denote the variety of all lattices, of the modular and of the distributive lattices, resp.

R. FRUCHT proved in 1950 [10] that

$$\alpha(G; \mathscr{L}) \leqq c_1 n \log n$$

(n denotes the order of G throughout).

PROBLEM 2.1. *Does there exist a constant c_2 such that*

$$\alpha(G; \mathscr{L}) \leqq c_2 n$$

for every finite group $G(|G|=n)$?

REMARK 2.2. For every finite group G there exists a commutative semigroup S of order $\leqq 2n+2$ such that $\operatorname{Aut} S \cong G$ [3, Theorem 4.9].

PROBLEM 2.3. *Does there exist a constant c_3 such that*

$$\alpha(G; \mathscr{M}) \leqq n^{c_3}?$$

REMARK 2.4. If a distributive lattice L admits an automorphism of prime order p, then $|L| \geqq 2^p$ [3, Prop. 4.8]. This implies that $\alpha(G; \mathscr{D}) \geqq 2^n$ for infinitely many finite groups G. On the other hand,

(i) $\alpha(G; \mathscr{D}) \leqq 2^{3n}$ for every finite group G ([3, Corollary 4.5]).

This follows from [3, Corollary 4.3]:

(ii) Given a finite group G there exists a poset P of $3n$ elements such that $\operatorname{Aut} P \cong G$.

Indeed, as proved by Birkhoff [6], if D is the (distributive) lattice of order-ideals of P then $|D|\leqq 2^{|P|}$, and $\operatorname{Aut} P\cong\operatorname{Aut} D$. So (ii) implies (i).

§ 3. Height of the lattice. Let $\eta(G;\mathscr{V})$ denote the minimum height of a finite lattice from the variety $\mathscr{V}$ whose automorphism group is isomorphic to G.

One can easily show that

$$\eta(G;\mathscr{L})\leqq 3$$

for every finite group G (Frucht [10]).

As a matter of fact, take a finite graph $X=(V,E)$ such that $\operatorname{Aut} X\cong G$. Let $L=\{0,1\}\cup V\cup E$ be a poset under the relation

$$\left.\begin{aligned} x&\leqq x\\ 0\leqq x&\leqq 1\end{aligned}\right\}\quad\text{for every } x \text{ in } L;$$

$$v\leqq e \text{ if } v\in V,\ e\in E \text{ and } v \text{ is an endpoint of } e.$$

Clearly, L is a lattice of height 3 and $\operatorname{Aut} L\cong\operatorname{Aut} X\cong G$.

One can also require L to be geometric (matroid lattice).

Theorem 3.1 ([4]). *Given a finite group G there exists a finite geometric lattice L of height* 3 *such that* $\operatorname{Aut} L\cong G$.

If infinite lattices are admitted, L can be chosen to be modular (using free completion of a partial projective plane, E. Mendelsohn [13]).

For the finite case, however, we conjecture:

Conjecture 3.2. *For every integer N there exists a finite group G such that $\eta(G;\mathscr{M})>N$. In other words, finite modular lattices of bounded height do not represent all finite groups as their groups of automorphisms.*

To my knowledge, even the case $N=3$ is still undecided. It appears to be an open problem to find a finite group which is not isomorphic to the automorphism group of any finite projective plane. C. Hering has obtained interesting results in this direction [11], [12] (cf. [8, chap. 4]).

Problem 3.3. *Are there infinitely many finite projective planes with no non-identity automorphism?*

We remark that 3.2 clearly holds for distributive lattices. In fact,

Proposition 3.4. *If the prime p divides the order of* $\operatorname{Aut} L$, *L a finite distributive lattice, then the Boolean algebra with p atoms has a cover-preserving embedding in L* [3, Prop. 4.8.].

From this it follows that

$$\eta(G;\mathscr{D})\geqq n$$

for infinitely many finite groups G.

On the other hand, by 2.4 (ii) we have

$$\eta(G;\mathscr{D})\leqq 3n$$

for every finite group G (since the height of the distributive lattice 2^P is $|P|$, P a poset). (Cf. [7, p. 59]. 2^P denotes the lattice of dual order ideals of P.)

§ 4. Order dimension.

The order dimension $\dim P$ of a partial order P is the minimum number of linear orders whose intersection is P ([9]).

The basic problem of this section is

PROBLEM 4.1. Given a variety $\mathscr{V}$ of lattices, decide whether there is a constant $c(\mathscr{V})$ such that every finite group G is isomorphic to $\operatorname{Aut} L$ for some $L\in\mathscr{V}$, $\dim L\leqq c(\mathscr{V})$. In other words, are the members of bounded dimension of $\mathscr{V}$ sufficient to represent every finite group as their group of automorphisms?

For $\mathscr{V}=\mathscr{L}$, the answer is positive:

THEOREM 4.2 ([5]). *Every finite group is the automorphism group of a finite lattice of dimension* 4.

CONJECTURE 4.3. *There exist finite groups, not isomorphic to the automorphism group of any finite lattice of dimension* 3.

PROBLEM 4.4. Characterize the automorphism groups of finite lattices of dimension 2. Are all these groups obtained from symmetric groups by repeated application of direct and wreath product?

(These latter groups are the automorphism groups of planar lattices, cf. [2]. Every planar lattice has dimension 2, but not conversely. Still it seems possible that no new groups occur.)

We conjecture that Problem 4.1 has a negative answer for $\mathscr{V}=\mathscr{M}$:

CONJECTURE 4.5. *For every integer N there exists a finite group G such that whenever* $\operatorname{Aut} L\cong G$ *for a modular lattice L then* $\dim L>N$.

REMARK 4.6. This is clearly true for distributive lattices. Indeed, every distributive lattice is the dual ideal lattice of a poset: 2^P and $\operatorname{Aut} P=\operatorname{Aut}(2^P)$. Moreover, the dimension of 2^P is equal to the width of P (size of its largest antichain). (This is folklore, it easily follows from Dilworth's theorem.) Every orbit of $\operatorname{Aut} P$ being an antichain, we have

$$\operatorname{Aut} P \leqq S_W\times\ldots\times S_W$$

where W denotes the width of P and S_W the symmetric group of degree W. Hence, for a distributive lattice L of dimension d,

$$\operatorname{Aut} L \leqq S_d\times\ldots\times S_d,$$

a severe restriction on the possible automorphism groups.

We are able to prove that for some larger modular varieties, the answer to 4.1 is still negative:

THEOREM 4.7 ([5]). *For every k there is an $N(k)$ such that if a modular lattice L of dimension $\leqq k$ has an automorphism of order p, p prime, $p>N(k)$, then M_p has a cover-preserving embedding in L.*

(M_p denotes the modular lattice of height 2 and order $p+2$.)

This implies for every s and k that the members of dimension $\leqq k$ of the variety generated by M_s do not represent every finite group (not even every finite cyclic group) as their group of automorphisms.

It would be nice to find the watershed: an identity f such that the following statements about a lattice variety $\mathcal{V}$ are equivalent: (i) Every finite group is the automorphism group of some member of $\mathcal{V}$ of bounded dimension. (ii) f does not hold in $\mathcal{V}$. (Of course, such an identity may not exist.)

Acknowledgement. My thanks are due to Dwight DUFFUS for discussions on some of these problems. This paper was written while the author was visiting Department of Mathematics, Vanderbilt University, Nashville, Tennessee.

REFERENCES

[1] BABAI, L., On the minimum order of graphs with given group, *Canad. Math. Bull.* **17** (1974), 89—91.

[2] BABAI, L., Automorphism groups of planar graphs II., in *Infinite and Finite Sets, Keszthely 1973* (A. Hajnal et al. eds.), Bolyai — North-Holland, 1975, 29—84

[3] BABAI, L., Finite digraphs with given regular automorphism groups, *Period. Math. Hungar.* **11** (1980), 257—270.

[4] BABAI, L., Vector representable matroids of given rank with given automorphism group, *Discrete Math.* **24** (1978), 119—125.

[5] BABAI, L.—DUFFUS, D., Dimension and automorphism group of lattices, *Alg. Univ.* (to appear).

[6] BIRKHOFF, G., Sobre los grupos de automorfismos, *Rev. Union Mat. Argentina* **11** (1946), 155—157.

[7] BIRKHOFF, G., *Lattice Theory,* American Mathematical Society, Providence, R. I., 1967.

[8] DEMBOWSKI, P., *Finite Geometries,* Springer-Verlag, Heidelberg, 1968.

[9] DUSHNIK B.—MILLER, E. W., Partially ordered sets, *Amer. J. Math.* **63** (1941), 600—610.

[10] FRUCHT, R., Lattices with given abstract group of automorphisms, *Canad. J. Math.* **2** (1950), 417—419.

[11] HERING, C., Eine Bemerkung über Automorphismengruppen von endlichen projektiven Ebenen und Möbiusebenen, *Arch. Math.* **18** (1967), 107—110.

[12] HERING, C., On 2-groups operating on projective planes, *Illinois J. Math.* **16** (1972), 581—595.

[13] MENDELSOHN, E., Every group is the collineation group of some projective plane, *J. of Geometry* **2** (1972), 97—106.

Department of Algebra, Roland Eötvös University
Múzeum krt. 6—8, H—1088 Budapest

(Received May 7, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 143—150.

О ФУНДАМЕНТАЛЬНОЙ ПРИВОДИМОСТИ ПОЛОЖИТЕЛЬНЫХ ОПЕРАТОРОВ В ПРОСТРАНСТВАХ С ИНДЕФИНИТНОЙ МЕТРИКОЙ

Ц. БАЯСГАЛАН

1. Введение

Скалярным произведением на комплексном векторном пространстве H называется комплекснозначная функция $(\cdot, \cdot)$, определенная для всех пар $x, y \in H$, такая что условия

$$(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1 (x_1, y) + \alpha_2 (x_2, y), \quad (y, x) = \overline{(x, y)}$$

выполняются для каждого $\alpha_1, \alpha_2 \in \mathbf{C}$ и $x_1, x_2, x, y \in H$.

Если пространство H представимо в виде прямой суммы двух своих подпространств H^+ и H^- таких, что

а) $(x^+, x^+) > 0$ (соотв. $(x^-, x^-) < 0$) для всех $x^+ \in H^+$, отличных от 0 (соотв. для всех $x^- \in H^-$, отличных от 0),

б) $(x^+, x^-) = 0$ для каждой пары $x^+ \in H^+$ и $x^- \in H^-$, то говорят, что данное разложение $H = H^+ \dotplus H^-$ является фундаментальным. Пространство H может обладать многими фундаментальными разложениями, но может не обладать ни одним.

Пусть $H = H^+ \dotplus H^-$—некоторое фундаментальное разложение, тогда с помощью скалярных произведений

$$(1.1) \qquad \begin{aligned} &[x^+, y^+]_{H^+} = (x^+, y^+) && (x^+, y^+ \in H^+), \\ &[x^-, y^-]_{H^-} = -(x^-, y^-) && (x^-, y^- \in H^-) \end{aligned}$$

можно, естественным образом, определить скалярное произведение на H формулой

$$(1.2) \qquad [x, y] = [x^+, y^+]_{H^+} + [x^-, y^-]_{H^-} = (x^+, y^+) - (x^-, y^-),$$

где $x = x^+ + x^-$, $y = y^+ + y^-$, $x^+, y^+ \in H^+$, $x^-, y^- \in H^-$. Имеет место неравенство $|(x, y)| \leqq \|x\| \, \|y\|$ $(x, y \in H)$, где $\|x\| = \sqrt{[x, x]}$, $\|y\| = \sqrt{[y, y]}$.

Если H обладает некоторым фундаментальным разложением, компоненты которого H^+ и H^- полны относительно скалярных произведений (1.1), то H называется пространством М. Г. Крейна. Тогда легко проверить, что соотвествующее скалярное произведение (1.2) превращает H в гильбертово пространство, причем внутренее произведение $(\cdot, \cdot)$ непрерывно в гильбертовой топологии.

Можно доказать следующие предложения (см. [1]).

1. *Компоненты любого фундаментального разложения пространства Крейна полны по отношению к соотвествующим скалярным произведениям* (1.1).

2. *Любое фундаментальное разложение пространства Крейна определяет описанным выше способом одну и ту же гильбертову топологию.*

3. *Кардинальные числа* $\varkappa^+(H)=\dim H^+$, $\varkappa^-(H)=\dim H^-$ *не зависят от выбора фундаментального разложения.*

Пространство Крейна H с условием $\min(\varkappa^+(H), \varkappa^-(H))<\infty$ называется пространством Понтрягина и обозначается через H_k, где $k=\min(\varkappa^+(H), \varkappa^-(H))$.

В дальнейшим мы рассмотрим только ограниченные линейные операторы в пространстве Крейна H. Каждый ограниченный линейный оператор A определяет ограниченный линейный оператор A^* (сопряженный к оператору A), удовлетворяющий условию

$$(Ax, y) = (x, A^* y) \qquad (x, y\in H).$$

Ограниченный линейный оператор A называется положительным, если $(Ax, x)\geqq 0$ $(x\in H)$. Оператор A называется фундаментально приводимым, если существует такое фундаментальное разложение $H=H^+ \dotplus H^-$, что $AH^+\subset H^+$, $AH^-\subset H^-$.

В этой заметке приводятся некоторые необходимые и достаточные условия фундаментальной приводимости положительных операторов. Фундаментальная приводимость операторов разных классов изучалась в работах Р. Филлипса [2], Е. Песонена [3], Р. Кюне [4], П. Хесса [5].

2. Другие определения и предложения из теории пространства Крейна

2.1. Если $H=H^+ \dotplus H^-$—некоторое фундаментальное разложение пространства Крейна, то проекторы P^+ и P^-, определенные соотношениями $P^+x=x^+$, $P^-x=x^-$, где $x=x^+ +x^-$, $x^+\in H^+$, $x^-\in H^-$, $x\in H$, называются фундаментальными проекторами. Оператор $J=P^+-P^-$ называется фундаментальной симметрией. Очевидно, что $J^2=I$ и $J^*=J$. Скалярное произведение (1.2) называется J-скалярным произведением. Имеет место соотношение $(Jx, y)=$ $=[x, y]$ $(x, y\in H)$. Для каждого ограниченного линейного оператора A, сопряженный относительно J-скалярного произведения обозначается через A^+. Имеем $A^+=JA^*J$.

2.2. Подпространство L пространства Крейна называется равномерно положительным, если $(x, x)\geqq\alpha\|x\|^2$ $(x\in L)$, где $\alpha=\alpha(L)>0$ и равномерно отрицательным, если $(x, x)\leqq -\alpha\|x\|^2$ $(x\in L)$, где $\alpha=\alpha(L)$. Если подпространство L удовлетворяет условию $L+L^\perp=H$, где $L^\perp$—ортогональное дополнение L (ортогональность всегда разумеем в смысле $(\cdot, \cdot)$), то мы говорим что L ортогонально дополняемо. Положительное подпространство L (т. е. $(x, x)\geqq 0$ при $x\in L$) ортогонально дополняемо в том и только в том случае, если оно замкнуто и равномерно положительно (см. [1]).

Подпространство L называется невырожденным если $L\cap L^\perp=0$. В пространстве Понтрягина каждое замкнутое, невырожденное подпространство ортогонально дополняемо.

Положительное подпространство L называется максимальным положительным подпространством, если L не содержится ни в каком другом положительном подпространстве.

2.3. Произвольный самосопряженный оператор в пространстве Понтрягина обладает инвариантным положительным подпространством (см. [1]).

2.4. *Спектральная функция*. Пусть A—положительный линейный оператор в пространстве Крейна H. Тогда существует (см. [1], [6]) единственная функция $E(t)$, определенная для всех вещественных чисел, отличных от нуля, такая что

1. $E(t)$—ортогональный проектор $\big($т. е. $E^2(t)=E(t), \big(E(t)\big)^*=E(t)\big)$ при $t\in\mathbf{R}$, $t\neq 0$;
2. $E(t_1)E(t_2)=E\big(\min(t_1, t_2)\big)$;
3. $E(t)H$ равномерно отрицательно при $t<0$, $\big(I-E(t)\big)H$ равномерно положительно при $t>0$;
4. $E(t)=O$ при достаточно малом t, $E(t)=I$ при достаточно большом t;
5. $E(t-0)=E(t)$, где $E(t-0)$ обозначает сильный предел в точке t;
6. $AE(t)=E(t)A$;
7. в спектральном представлении

$$A = S + \int_{-\infty}^{\infty} t\, dE(t)$$

S является положительным оператором со свойствами $S^2=O$, $AS=SA=O$ и интеграл существует в сильной топологии как несобственный интеграл с особой точкой $t=0$.

2.5. *Несколько обозначений*. Через $N(A)$ обозначается ядро оператора A, через $R(A)$ — область значений A. Знаки $+$ и $(+)$ обозначают прямую и ортогональную прямую сумму соотвественно. $\sigma(A)$ — спектр оператора A

3. О фундаментальной приводимости положительных операторов

Пусть H — пространство Крейна с индефинитным скалярным произведением $(\cdot, \cdot)$. В дальнейшим рассмотрим только ограниченные линейные операторы в пространстве H.

Теорема 1. *Если A — положительный оператор в H, имеющий ограниченный обратный, то A фундаментально приводим.*

Доказательство. Фиксируем некоторое фундаментальное разложение с фундаментальной симметрией J. Поскольку A — положительный оператор, то AJ является положительным оператором относительно J-скалярного произведения $[\cdot, \cdot]$. Действительно, $[AJx, x]=[Jx, A^+x]=[Jx, JAJx]=(x, JAJx)=$ $=(Jx, AJx)\geqq 0$. Далее из непрерывности A^{-1} следует непрерывность оператора $(AJ)^{-1}=JA^{-1}$. Таким образом по известной теореме о квадратных корнях в гильбертовом пространстве существует обратимый J-положительный корень из AJ. Обозначим этот корень через B. Тогда $AJ=B^2$, отсюда следует $AB=$ $=B^2JB=BG$, где мы положили $G=BJB$.

Очевидно, что оператор G обратимый и J-самосопряженный. Следовательно, G есть J-ортогональная сумма строго J-положительного оператора на некотором подпространстве H_1 и строго J-отрицательного оператора на дополнительном подпространстве H_2. Положим $Q_k=BH_k$.

Ввиду обратимости B имеем $Q_1\cap Q_2=0$, $Q_1\dot{+}Q_2=H$. Из $AQ_k=ABH_k=$ $=BGH_k\subset BH_k=Q_k$ следует, что подпространства Q_k инвариантны относительно A. Разложение $Q_1\dot{+}Q_2=H$ фундаментально, так как $(Bx, By)=[JBx, By]=$ $=[BJBx, y]=[Gx, y]$.

Следствие. *Если A является t-дефинизируемым оператором* (см. [6]), *удовлетворяющим неравенству* $\alpha_-(A)<\alpha_+(A)$, *где* $\alpha_-(A)=-\inf\limits_{(x,x)=-1}(Ax, x)$, $\alpha_+(A)=\inf\limits_{(x,x)=1}(Ax, x)$, *то A фундаментально приводим.*

Доказательство. При $\alpha_-(A)<\alpha<\alpha_+(A)$ оператор $A-\alpha I$ будет положительным и существует ограниченный обратный $(A-\alpha I)^{-1}$ (см. [6]). По теореме 1 $A-\alpha I$, и тем самым A, фундаментально приводим.

Теорема 2. *Пусть $H=H_k$ — пространство Понтрягина, где $k=$ $=\min(\varkappa^+(H), \varkappa^-(H))$, и A — положительный оператор в H. Для того, чтобы A был фундаментально приводимым необходимо и достаточно, что ядро оператора A было ортогонально дополняемым.*

Доказательство. *Необходимость.* Вообще, если A — произвольный фундаментально приводимый оператор, тогда его ядро ортогонально дополняемо. Действительно, пусть $H=H^+\dot{+}H^-$ фундаментальное разложение, где $AH_\pm\subset H_\pm$. Обозначим через $A_\pm$ сужения A на подпространства $H_\pm$. Легко видно, что $N(A)=N(A_+)(\dot{+})N(A_-)$. Тогда в силу теоремы V.5.3 из [1] ядро оператора A ортогонально дополняемо.

Достаточность. Пусть $H=N(A)(\dot{+})N(A)^\perp$. Сужение оператора A на $N(A)^\perp$ обозначим через A_1. Ясно, что $N(A_1)=0$. В частности, из $(A_1x, x)=0$ следует $x=0$. Поэтому в пространстве Понтрягина $N(A)^\perp$ оператор A_1 является положительным оператором Песонена. По теореме Понтрягина у A_1 имеются максимальное положительное и максимальное отрицательное инвариантные подпространства, ортогональные между собой (см. 2.3). Так как A_1 является оператором Песонена, по теореме IX.5.2 из [1] A_1 фундаментально приводим. Рассматривая фундаментальное разложение подпространства $N(A)$, мы приходим к фундаментальной приводимости исходного оператора A.

Теорема 3. *Пусть A — положительный оператор в H. Тогда A фундаментально приводим в том и только в том случае, если*

а) *существуют сильные пределы* $E(+0)$ и $E(-0)$ *спектральной функции оператора A в точке* $t=0$ (см. 2.4);

б) *в спектральном представлении* $A=S+\int\limits_{-\infty}^{\infty} t\,dE(t)$ *оператор S равен нулю.*

Доказательство. *Необходимость.* Рассмотрим инвариантное фундаментальное разложение $H=H^+\dot{+}H^-$ и соотвествующие проекторы P^+ и P^-. Поскольку $AP^\pm=P^\pm A$, то (см. конструкцию в [6]) $P^\pm E(t)=E(t)P^\pm$.

Отсюда $E(t)H^{\pm}\subset H^{\pm}$. Теперь из $A=S+\int\limits_{-\infty}^{\infty} t dE(t)$ следует, что $SH^{\pm}\subset H^{\pm}$. Так как S самосопряжен и $S^2=O$ получаем, что $S=O$.

Положим $E(t)|H^{\pm}=E_{\pm}(t)$. При любых $x\in H^+$ и $t_1<t_2$, $t_i\neq 0$, имеем $((E_+(t_2)-E_+(t_1))x, x)=((E_+(t_2)-E_+(t_1))x, (E_+(t_2)-E_+(t_1))x)$, так что по теории гильбертовых пространств $E_+(-0)$ и $E_+(+0)$ существуют. Эти же утверждения верны и для E_-. Стало быть, существуют силные пределы $E(\pm 0)$.

Достаточность. Имеем ортогональное разложение $H=E(-0)H(\dotplus)(\dotplus)(\mathbf{I}-E(+0))H(\dotplus)(E(+0)-E(-0))H$. Здесь $E(-0)H$ — равномерно отрицательное, $(\mathbf{I}-E(+0))H$ — равномерно положительное инвариантные подпространства оператора A. Что касается третьей компоненты, из $S=O$ следует $(E(+0)-E(-0))H=N(A)$ (см. [6]), так что любое фундаментальное разложение подпространства $(E(+0)-E(-0))H$ инвариантно относительно оператора A.

Теорема 4. *Пусть A — положительный оператор в H. Оператор A фундаментально приводим в том и только том случае, когда*

а) *ядро оператора A ортогонально дополняемо;*

б) *существуют пределы $E(+0)$ и $E(-0)$.*

Доказательство. Необходимость вытекает из теоремы 3 (ортогональная дополняемость ядра A содержится в доказательстве теоремы 2). Докажем сейчас, что эти условия достаточны. Представим

$$H=N(A)(\dotplus)\overline{R(A)}. \tag{3.1}$$

Поскольку $AS=SA=0$ (см. 2.4), на $\overline{R(A)}$ оператор S равен нулю. Далее, разложение (3.1) приводит $E(t)$, ибо проекторы на подпространства $N(A)$, $\overline{R(A)}$ коммутируют с A, следовательно и с $E(t)$. Кроме того, S тоже приводится разложением (3.1). Теперь из спектрального разложения оператора A в подпространстве $N(A)$ вытекает, что $S|N(A)=O$. Значит S равен нулю. По теореме 3 доказательство закончено.

Сейчас мы приходим к рассмотрению положительного компактного оператора A в пространстве Крейна. Обозначим через $L_+(L_-)$ замкнутую линейную оболочку всех корневых линеалов, отвечающих положительным (отрицательным) собственным значениям оператора A.

Теорема 5. *Пусть A — положительный компактный оператор в пространстве Крейна H. Тогда A фундаментально приводим тогда и только тогда, когда*

а) *ядро оператора A ортогонально дополняемо;*

б) *$L_{\pm}$ — равномерно определенные подпространства.*

Доказательство. *Необходимость.* Сначала докажем следующее: Если A фундаментально приводим, то все собственные значения оператора A полупросты.

Пусть фундаментальное разложение $H=H^+\dotplus H^-$ приводит A, число t — собственное значение A и для некоторого $x\neq 0$ выполняется $(A-tI)^r x=0$, где $r \geqq 1$. Если $x = x^+ + x^-$, где $x^\pm \in H^\pm$, тогда $(A-tI)^r x^+ = 0$, $(A-tI)^r x^- = 0$. По теории гильбертовых пространств отсюда получим $(A-tI)x^+ = 0$, $(A-tI)x^- = 0$, следовательно $(A-tI)x = 0$. Далее, пусть $Ax = tx$, где $t > 0$. При $x = x^+ + x^-$, $x^\pm \in H^\pm$ имеем $Ax^\pm = tx^\pm$. В частности $x^- = 0$ (если бы $x^- \neq 0$, то t — положительное собственное значение с отрицательным собственным вектором положительного оператора, что невозможно). Значит $x = x^+ \in H^+$ т. е. $L_+ \subset H^+$. Аналогично, доказывается $L_- \subset H^-$. Отсюда $L_\pm$ — равномерно определенные подпространства. Свойство а) следует из доказательства теоремы 2.

Достаточность. Имеем такой факт: если A — положительный компактный оператор в H с нулевым ядром, то $\overline{L_+ + L_-} = H$. Этот факт установлен в работе [7] без использования спектральной функции. Но этого можно легко добиться, используя спектральную функцию. Действительно, так как $A = S + \int\limits_{-\infty}^{\infty} t\,dE(t)$ и по условию $N(A) = 0$ (отсюда $S = O$), то имеем

$$Ax = \sum_{k=1}^{\infty} t_{\pm k}[E(t_{\pm k}+0) - E(t_{\pm k})]x,$$

где $t_{\pm k}$ — положительные, соотв. отрицательные собственные значения оператора A. Далее $A\big(E(t_{\pm k}+0) - E(t_{\pm k})\big)x = t_{\pm k}\big(E(t_{\pm k}+0) - E(t_{\pm k})\big)x$ так что $R(A) \subset \overline{L_+ + L_-}$. По условию $\overline{R(A)} = H$. Отсюда $\overline{L_+ + L_-} = H$.

Докажем достаточность условия. Имеем $H = N(A)(\dotplus)\overline{R(A)}$. Сужение оператора A на подпространство $\overline{R(A)}$ обозначим через A_1. Тогда $N(A_1) = 0$ и A_1 — положительный компактный оператор в $\overline{R(A)}$. С помощью предыдущего факта, имеем $\overline{L_+ + L_-} = \overline{R(A)}$ т. е. $L_+ + L_- = \overline{R(A)}$. Следовательно $H = N(A)(\dotplus)L_+(\dotplus)L_-$. Тем самым A фундаментально приводим.

Замечания. (1) Если A — положительный компактный оператор с нулевым ядром, то единственное возможное фундаментальное разложение приводящее A есть $H = L_+(\dotplus)L_-$.

(2) Теореме можно придать эквивалентную формулировку. Для того, чтобы положительный компактный оператор фундаментально приводим необходимо и достаточно, чтобы а) ядро оператора ортогонально дополняемо, б) линеал $L_+ + L_-$ замкнуто.

4. Примеры

I. Существует положительный компактный оператор в пространстве Крейна H, где $\varkappa^+(H)$, $\varkappa^-(H)$ бесконечные и $N(A) = 0$, который не фундаментально приводим (см. теорему 2). Построение реализовано в двух этапах.

а) Существует такое замкнутое подпространство N_+, которое положительно определенное, максимально положительное, но не равномерно положительное. Действительно, по теореме V. 6.3 [1] существует положительно определенное, замкнутое, но не равномерно положительное подпространство N_+. Но из конструкции доказательства этой теоремы легко видно, что это подпространство можно считать максимальным положительным. Поскольку N_+ положительно определенное, то $\overline{N_+ + N_+^\perp} = H$, причем $N_+^\perp$ — отрицательно определенное подпространство.

Выберем J-ортонормированный базис $\{e_i^+\}_{i=1}^\infty$ в N_+. С помощью ортогонализации Грама-Шмидта (см. [1], теорема IV.3.2) получим ортогональную систему векторов $\{f_i^+\}_{i=1}^\infty \subset N_+$, такую что линейная оболочка этих векторов содержит каждый e_i. Отсюда замкнутая линейная оболочка $\{f_i^+\}_{i=1}^\infty$ совпадает с N_+. При этом можно считать $\|f_i^+\| \leqq C_1$ ($i=1, 2, \ldots$, C_1-постоянное). Аналогично построим ортогональную систему $\{f_i\}_{i=1}^\infty$ в $N_+^\perp$, $\|f_i^-\| \leqq C_2$, замкнутая линейная оболочка которой совпадает с $N_+^\perp$.

После этого, мы выберем числа $\mu_k^+ > 0$, $\mu_k^- < 0$ такие что ряды $\sum\limits_{k=1}^\infty \mu_k^+$, $\sum\limits_{k=1}^\infty \mu_k^-$ сходятся.

б) Сейчас построим вышеупомянутый оператор. Положим

$$Af = \sum_{k=1}^\infty \mu_k^+ (f, f_k^+) f_k^+ - \sum_{k=1}^\infty \mu_k^- (f, f_k^-) f_k^- \qquad (f \in H).$$

1. Существование оператора:

$$\left\| \sum_{k=n}^m \mu_k^+ (f, f_k^+) f_k^+ \right\| \leqq \sum_{k=n}^m \mu_k^+ \|f\| \, \|f_k^+\|^2 \leqq C_1^2 \|f\| \sum_{k=n}^m \mu_k^+ .$$

2. A — компактный оператор: если

$$A_n f \equiv \sum_{k=1}^n \mu_k^+ (f, f_k^+) f_k^+ - \sum_{k=1}^n \mu_k (f, f_k^-) f_k^-,$$

то

$$\|Af - A_n f\| = \left\| \sum_{k=n+1}^\infty \mu_k^+ (f, f_k^+) f_k^+ - \sum_{k=n+1}^\infty \mu_k^- (f, f_k^-) f_k^- \right\| \leqq$$

$$\leqq \max\{c_1^2, c_2^2\} \|f\| \left(\sum_{k=n+1}^\infty \mu_k^+ - \sum_{k=n+1}^\infty \mu_k^- \right).$$

3. Очевидно, A — положительный оператор. Пусть $f \in N(A)$. Тогда $(Af, f) = 0$ или $\sum\limits_{k=1}^\infty \mu_k^+ + |(f, f_k^+)|^2 - \sum\limits_{k=1}^\infty \mu_k^- |(f, f_k^-)|^2 = 0$. Отсюда $(f, f_k^+) = 0$, $(f, f_k^-) = 0$. Поскольку $\overline{N_+ + N_+^\perp} = H$, то $f = 0$. Далее, замкнутая линейная оболочка всех собственных векторов, отвечающих положительным собственным значениям содержит N_+ (потому, что f_k^+ собственный вектор, отвечающий собственному значению $\mu_k \cdot (f_k^+, f_k^+) > 0$).

Тогда ясно, что эта замкнутая линейная оболочка не равномерно положительно. По теореме 4 A не является фундаментально приводимым.

Этот пример показывает, что в непонтрягиновом пространстве для положительного оператора условие тривиальности ядра не достаточно для фундаментально приводимости оператора (см. теорему 2).

С помощью этого примера можно построить такой же оператор как предыдующий, ядро которого не тривиальное, но ортогонально дополняемо.

Эти примеры, вместе с теоремой 4 показывают, что существование пределов $E(+0)$, $E(-0)$ не следует из ортогональной дополняемости подпространства $\sigma_0(A)$, где $\sigma_0(A)$ — корневое подпространство, отвечающее значению $t=0$.

II. Пусть H — индефинитное пространство размерности два. Пусть A — произвольный ненулевой положительный оператор. Тогда легко видно, что A фундаментально приводим в том и только в том случае, если $A^2 \neq O$.

В заключение автор выражает глубокую благодарность проф. Я. Богнару за постоянное внимание к работе.

БИБЛИОГРАФИЯ

[1] Bognár, J., *Indefinite inner product spaces,* Springer, Berlin, 1974.

[2] Phillips, R. S., The extension of dual subspaces invariant under an algebra. In: *Proc. Internat. Sympos. Linear Spaces,* Jerusalem and Oxford, Jerusalem Academic Press and Pergamon Press, 1961, 366—398.

[3] Pesonen, E., Über die Spektraldarstellung quadratischer Formen in linearen Räumen mit indefiniter Metrik, *Ann. Acad. Sci. Fenn. Ser. AI,* no. **227** (1956).

[4] Kühne, R., Über eine Klasse *J*-selbstadjungierter Operatoren, *Math. Ann.* **154** (1964), 56—69.

[5] Hess, P., Zur Theorie der linearen Operatoren eines *J*-Raumes. Operatoren die von kanonischen Zerlegungen reduziert werden, *Math. Z.* **106** (1968), 88—96.

[6] Крейн, М. Г., Шмульян, Ю. Л., *J*-полярное представление плюс-операторов, *Математические исследования* **1** (1966), выпуск 2, 172—210.

[7] Эктов, Ю. С., *J*-неотрицательные вполне непрерывные операторы, *Труды научно-исследовательского института математики Воронежского университета,* 1975, выпуск 17, 77—86.

Математический Институт Венгерской Академии Наук Будапешт

(Поступила: 28-ого мая 1979 г.)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 151—155.

ON SOME PARTITION PROPERTIES OF FAMILIES OF SETS

by

G. ELEKES, P. ERDŐS and A. HAJNAL

§ 0.

This write up contains a list of results and problems concerning questions we have stated and investigated in some earlier papers [1], [2], [3]. Though we do not give proofs the experienced reader will be able to reconstruct most of them, by checking the lemmas we are going to state and following the hints we give after stating some of the theorems.

We deal with questions of the following type. Let S be (a relatively large) family of sets. In most of the questions we will ask S will be of the form $P(\varkappa)$, the set of all subsets of an infinite cardinal $\varkappa$. There will be given a mapping $f\colon P(\varkappa)\to\gamma$. f will be called a partition of $P(\varkappa)$ with γ colors. A subfamily $S'\subset S$ will be called homogeneous for f if $S'\subset f^{-1}(\{\eta\})$ for some $\eta<\gamma$. As usual we will ask for the existence of relatively large homogeneous subfamilies.

DEFINITION 0.1. $\mathcal{A}=\{A_\nu\colon \nu<\varkappa\}$ is a $\varkappa$, *Δ-system* if there is a set D such that $A_\nu\cap A_\mu=D$, $A_\nu\neq A_\mu$ for $\nu\neq\mu$; $\nu,\mu<\varkappa$. D is the *kernel* of the *Δ-system $\mathcal{A}$*.

DEFINITION 0.2. a) Let $\mathcal{A}=\{A_\nu\colon \nu<\varkappa\}$ be a sequence of sets. Put $\mathcal{A}(N)=\bigcup\{A_\nu\colon \nu\in N\}$ for $N\subset\varkappa$.

b) A family $\mathcal{F}$ of sets is said to be a $(\varkappa,\lambda)$-*system* determined by $\mathcal{A}=\{A_\nu\colon \nu<\varkappa\}$ if $\mathcal{F}=\{\mathcal{A}(N)\colon N\in[\varkappa]^{<\lambda}\setminus\{\emptyset\}\}$ and

$$\mathcal{A}(N_0)\neq\mathcal{A}(N_1)\quad\text{for}\quad N_0\neq N_1,\quad N_0,N_1\in[\varkappa]^{<\lambda}.!$$

DEFINITION 0.3. A family $\mathcal{F}$ is said to be a $(\varkappa,\lambda)$, *Δ-system* if there is a $(\varkappa,\lambda)$-system $\mathcal{F}'$ determined by $\mathcal{A}=\{A_\nu\colon \nu<\varkappa\}$ such that $\mathcal{A}$ is a Δ-system with kernel D and $\mathcal{F}=\mathcal{F}'\cup\{D\}$.

To have short notation we introduce relations bearing some resemblance to partition relations investigated earlier.

DEFINITION 0.4. $S\to\Delta(\varkappa)_\gamma$, $S\to([\varkappa]^{<\lambda})_\gamma$, $S\to\Delta([\varkappa]^{<\varkappa})_\gamma$ mean that for all partitions $f\colon S\to\gamma$ of S with γ colors there is a $\varkappa$, Δ-system, a $(\varkappa,\lambda)$-system and a $(\varkappa,\lambda)$, Δ-system homogeneous for f, respectively. $A\nrightarrow$ indicates that the respective statements are false.

§ 1. Positive arrow relations for the first two symbols

DEFINITION 1.1. For a $U\subset P(\varkappa)$, $A, B\subset\varkappa$ we write $A\subset_U B$ if $A\subset B$ and there is a $C\in U$ with $B-A\supset C$.

a) $S\subset P(\varkappa)$ is *dense* in $[A, B]$ for U if $A\subset_U B$ and

$$\forall A'B'(A\subset A'\subset_U B'\subset B\Rightarrow\exists C\in S(A'\subset_U C\subset_U B')),$$

b) $S\subset P(\varkappa)$ is *left (right) dense* in $[A, B]$ for U if $A\subset_U B$ and

$$\forall A'B'\big(A\subset A'\subset_U B'\subset B\Rightarrow\exists C\in S(A'\subset_U C\subset B')$$

$$\big(\forall A'B'(A\subset A'\subset_U B'\subset B\Rightarrow\exists C\in S(A'\subset C\subset_U B'))\big).$$

If $S\subset P(\varkappa)$ is not dense (not left or right dense) for U in any $[A, B]$ then S is *nowhere dense (nowhere left or right dense)* for U in $P(\varkappa)$.

The following sequence of Baire-type lemmas serve as a basis of our proofs. Note that they all imply existence of dense (in a certain sense) sets homogeneous for some partitions.

LEMMA 1.1. *Let $U=[\omega]^\omega$. $P(\omega)$ is not the union of countably many sets nowhere dense for U.*

This lemma has been proved in [1]. The next lemma is due to J. BAUMGARTNER and is included here with his permission.

LEMMA 1.2. *Let $\varkappa>\omega$ be a regular cardinal, and U a normal filter in $P(\varkappa)$. Then $P(\varkappa)$ is not the union of $\varkappa$-sets nowhere left dense (right dense) for U.*

LEMMA 1.3. *Let $\varkappa\geqq\omega$ be a regular cardinal. Let $R(\varkappa)=\big\{g\in{}^\varkappa 2\colon\exists\xi<\varkappa\big(g(\xi)=$ $=1\wedge\forall\xi<\eta<\varkappa(g(\eta)=0)\big)\big\}$, i.e., the well-known Hausdorff set of 0—1-sequences of length $\varkappa$, with a last 1-digit. Let $<_\varkappa$ denote the usual lexicographic ordering of $R(\varkappa)$, and $U_\varkappa$ the set of non-empty open intervals of $\langle R(\varkappa), <_\varkappa\rangle$. Then $P\big(R(\varkappa)\big)$ is not the union of $\varkappa$-sets nowhere dense for $U_\varkappa$.*

COROLLARY 1.1. *Let $\varkappa\geqq\omega$ be regular and $2^{\underset{\smile}{\varkappa}}=\varkappa$. There is a non-empty $U\subset P(\varkappa)$ such that $P(\varkappa)$ is not the union of $\varkappa$ sets nowhere dense for U and U satisfies the following conditions* a) b):

a) *If $\{I_\eta\colon\eta<\varphi\}$ is a decreasing sequence of type $\varphi<\varkappa$ of elements of U then there is an $I\in U$ such that $I\subset I_\eta$ for $\eta<\varphi$.*

b) *Each member of U contains $\varkappa$-pairwise disjoint members of U.*

The next lemma transfers the above results for the case of singular $\varkappa$'s.

LEMMA 1.4. *Let $\lambda=\mathrm{cf}(\varkappa)<\varkappa$ be a singular cardinal. Assume $\varkappa_\alpha<\varkappa$ for $\alpha<\lambda$ and $\varkappa=\sup\{\varkappa_\alpha\colon\alpha<\lambda\}$. Let $\varkappa=\bigcup\{X_\alpha\colon\alpha<\lambda\}$ be a decomposition of $\varkappa$ and assume $U_\alpha\subset P(X_\alpha)$ and $P(X_\alpha)$ is not the union $\varkappa_\alpha$-sets nowhere dense [nowhere left (right) dense] for U_α for $\alpha<\varkappa$. Let*

$$V_\alpha=\{Y\subset\varkappa\colon\forall\alpha\leqq\beta<\lambda(Y\cap X_\beta\in U_\beta)\}\quad\textit{for}\quad\alpha<\lambda.$$

Then for each decomposition $\bigcup\{S_\eta\colon\eta<\varkappa\}=P(\varkappa)$ of $P(\varkappa)$ there are $\eta<\varkappa$ and $\alpha<\lambda$ such that S_η is dense (left or right dense) for V_α in $P(\varkappa)$.

Note that it is consistent with $2^{\aleph_0}=\aleph_2$ that Lemma 1.1 remains true for $\aleph_1$ sets instead of countably-many. This follows from a result of S. SHELAH [4], and from the fact that forcing a Silver real is proper forcing.

THEOREM 1.1. a) *$P(\varkappa)\to\Delta(\lambda)_\varkappa$ for $\lambda<\varkappa\geqq\omega$;*
b) *$P(\varkappa)\to\Delta(\varkappa)_\varkappa$ for all regular $\varkappa\geqq\omega$.*

For $\varkappa=\omega$ use Lemma 1.1. For $\varkappa>\omega$ use the left dense forms of Lemmas 1.4 and 1.2, respectively. For $\varkappa>\omega$ we originally proved this result under the assumption $2^{\underset{\smile}{\varkappa}}=\varkappa$, using Lemma 1.3. The more general theorem stated above is due to Baumgartner.

PROBLEM 1. Does $P(\aleph_\omega)\to\Delta(\aleph_\omega)_{\aleph_\omega}$ hold? We do not know the answer even assuming G.C.H.

THEOREM 1.2. *Assume $\varkappa\geqq\omega$ and $2^{\underset{\smile}{\varkappa}}=\varkappa$. Then*

$$P(\varkappa)\to([\omega]^{<\omega})_\varkappa \quad \textit{for} \quad \varkappa\geqq\omega.$$

Note that this implies $P(\omega)\to([\omega]^{<\omega})_\omega$ without any assumption, as it was announced in [1].

PROBLEM 2. Can one prove $P(\omega_1)\to([\omega]^{<\omega})_{\omega_1}$ without assuming $2^{\aleph_0}=\aleph_1$?

THEOREM 1.3. a) *$P(\varkappa)\to([\lambda]^{<n})_\varkappa$ for $n<\omega, \varkappa\geqq\omega, \lambda<\varkappa$.*
b) *$P(\varkappa)\to([\varkappa]^{<n})_\varkappa$ for $n<\omega$ and for all regular $\varkappa\geqq\omega$.*

As to the proofs of Theorems 1.2 and 1.3 consider a partition $f\colon P(\varkappa)\to\varkappa$ of $P(\varkappa)$. By the density lemmas, there is a $\nu<\varkappa$ such that $S=f^{-1}(\{\nu\})$ is appropriately dense in some $[A, B]$. In cases $\varkappa=\omega$ we use Lemma 1.1 and $U=[\omega]^\omega$ density in both proofs. In case $\varkappa>\omega$ regular, for Theorem 1.2 we use the U described in Corollary 1.1 and for Theorem 1.3 we use the normal filter induced by the clubs. In case $\varkappa$ is singular we apply Lemma 1.4 to obtain appropriate U'-s. We finish the proofs by showing that all S dense in $[A, B]$ for U contain (ω, ω)-systems (λ, η)-systems and $(\varkappa, \eta)$-systems, respectively. In both proofs the (τ, σ)-systems are constructed according to the following pattern. We first define a sequence $\{B_N\colon N\in[\tau]^{<\sigma}\setminus\{\emptyset\}\}\subset S$, and prove later that for $A_\nu=B_{\{\nu\}}, \nu<\tau$ $\mathscr{A}(N)=B_N$ holds for $N\in[\tau]^{<\sigma}\setminus\{\emptyset\}$. The B_N are constructed by induction according to a fixed well-ordering of $[\tau]^{<\sigma}\setminus\{\emptyset\}$. In both proofs we need special tricks to make sure that if $B_M\colon M<N$ is defined there is room enough to find B_N. We omit the details of this constructions.

§ 2. Negative arrow relations for the second symbol. Generalizations

As we have already mentioned in [1] it is easy to see that

THEOREM 2.1. *$P(\omega)\nrightarrow([\omega]^{<\omega_1})_2$ holds.*

PROBLEM 3. a) Does $P(\varkappa)\to([\omega]^{<\omega_1})_2$ hold for any $\varkappa$?
b) Does $[\varkappa]^\omega\nrightarrow([\omega]^{<\omega_1})_2$ hold for any $\varkappa$?

The following result shows that in Theorem 1.2 we cannot get a larger homogeneous $[\lambda]^{<\omega}$ system.

THEOREM 2.2. *Assume* $2^{\varkappa}=\varkappa^+$. *Then*

$$P(\varkappa)\nrightarrow([\omega_1]^{<\omega})_{\varkappa}.$$

PROBLEM 4. a) Can one prove $P(\omega)\nrightarrow([\omega_1]^{<\omega})_{\omega}$ without assuming C.H.?
b) Can one prove $P(\varkappa)\rightarrow([\omega_1]^{<\omega})_2$ for any $\varkappa$?

PROBLEM 5. a) $P(\omega)\rightarrow([\omega_1]^{<3})_2$?
b) $P(\omega)\rightarrow([\omega_1]^{<3})_{\omega}$?

We can neither prove a) nor disprove b) in any reasonable extension of ZFC. Theorem 2.2 is a consequence of the following lemma due to P. KOMJÁTH.

LEMMA 2.1. *Let* $<$ *be a well-ordering of* $T=[\omega_1]^{<\omega}\setminus\{\emptyset\}$. *Then there are* $L<M<N$; $L, M, N\in T$ *such that* $L\cup M=N$.

LEMMA 2.2. *Assume* $\varkappa>\tau\geqq\omega$. *Let* $\{R_\alpha\colon \alpha<\varkappa\}$ *be an increasing continuous sequence of fields of sets, i.e.,* $R_\beta\subset R_\alpha$ *for* $\beta<\alpha<\varkappa$ *and* $R_\alpha=\bigcup\{R_\beta\colon \beta<\alpha\}$ *for limit* α, $\alpha<\varkappa$. *Assume that* $\bigcup\{R_\alpha\colon \alpha<\varkappa\}\supset\mathscr{F}$ *for some* (τ^+,ω)-*system* $\mathscr{F}$. *Then there are* $\alpha<\varkappa$ *and a* (τ,ω)-*system* $\mathscr{F}'$ *with* $\mathscr{F}'\subset R_{\alpha+1}\setminus R_\alpha$.

Let now $\varkappa^{+(\mu)}$ denote the μ-th successor of $\varkappa$. We get

THEOREM 2.3. *Assume* S *is a system of sets,* $|S|=\varkappa^{+(\mu)}$ *for some* $\varkappa\geqq\omega$ *and* $\mu<\varkappa^+$. *Then* $S\nrightarrow([\aleph_\mu]^{<\omega})_{\varkappa}$.

This result actually yields a stronger theorem then 2.2. We also get

COROLLARY 2.1. *Assume* $\varkappa\geqq\omega$ *is regular and* $2^{\varkappa}<\varkappa^{+(\varkappa^+)}$. *Then* $P(\varkappa)\nrightarrow([2^{\varkappa}]^{<\omega})_{\varkappa}$.

For example $P(\omega)\nrightarrow([2^{\omega}]^{<\omega})_{\omega}$ provided $2^{\omega}<\aleph_{\omega_1}$.
There is nothing to prevent this $\nrightarrow$ from being true in ZFC but we cannot prove it.
Finally we state one very special result which only shows how one *cannot* solve Problem 5.

THEOREM 2.4. *Assume* $\varkappa$ *is regular and* $2^{\underset{\smile}{\varkappa}}=\varkappa$. *For every coloring* $f\colon P(\varkappa)\rightarrow\varkappa$, *of* $P(\varkappa)$ *with* $\varkappa$ *colors there is a* $\nu<\varkappa$ *such that* $S=f^{-1}(\{\nu\})$ *contains a set system of the following form* $\mathscr{F}=\{A_\mu\colon \mu<\varkappa\}\cup\{B_\nu\colon \nu<\varkappa^+\}\cup\{A_\mu\cup B_\nu\colon \mu<\varkappa\wedge\nu<\varkappa^+\}$ *where all the sets* A_μ, B_ν, $A_\mu\cup B_\nu$ *are different.*

This could be expressed by the symbol

$$P(\varkappa)\rightarrow([\varkappa,\varkappa^+]^{<2,<2})_{\varkappa}$$

had we defined this in this generality.

§ 3. The third symbol

Here we only mention a result and one problem.

THEOREM 3.1. *Let* $\lambda \geqq \omega$ *then*

$$2^{\aleph_0} \geqq \lambda^{+(n)} \Leftrightarrow P(\omega) \to \Delta([n]^{<n+1})_\lambda \quad \textit{for} \quad 1 \leqq n < \omega.$$

PROBLEM 6. Does $2^{\aleph_0} > \aleph_\omega$ imply

$$P(\omega) \to \Delta([\omega]^{<\omega})_\omega?$$

REFERENCES

[1] ELEKES, G., On a partition property of infinite subsets of a set, *Period. Math. Hungar.* **5** (1974), 215—218.

[2] ERDŐS, P., Problems and results on finite and infinite combinatorial analysis, *Infinite and finite sets* (Colloq. Math. Soc. J. Bolyai **10**), North-Holland Publishing Co., Amsterdam, 1975, 403—424.

[3] ERDŐS, P.—HAJNAL, A., Solved and unsolved problems in set theory, *Proceedings of the Tarski Symposium* (Berkeley Calif., 1971), Amer. Math. Soc., Providence, R. I., 1974, 269—287.

[4] SHELAH, S., (preprint).

[5] ELEKES, G., Colouring of infinite subsets of ω, *Infinite and finite sets* (Colloq. Math. Soc. J. Bolyai 10), North-Holland Publ. Co., Amsterdam—New York, 1975, 393—396.

Department for Analysis I, Roland Eötvös University
H—1088 Budapest, Múzeum krt. 6—8
and
Mathematical Institute of the Hungarian Academy of Sciences
H—1053 Budapest, Reáltanoda u. 13—15

(Received June 4, 1979)

CLOSEST PACKINGS AND CLOSEST COVERINGS BY TRANSLATES OF A CONVEX DISC

by

J. LINHART

By a *disc* we mean a compact convex subset of the Euclidean plane $\mathbf{R}^2$ with non-empty interior. A set M of discs is said to form a *packing*, if every point of the plane is contained in the interior of *at most* one disc of M. Dually, M is said to form a *covering*, if every point is contained in *at least* one disc.

L. FEJES TÓTH [4, 5] introduced the following notion. Let r be the supremum of the radii of the circles disjoint to the discs of a packing. Then $1/r$ is called the *closeness of the packing*. Concerning coverings, Fejes Tóth considered the supremum R of the radii of the circles contained in two discs. R is sometimes called "looseness", but it seems to be more natural to take $1/R$ and call it the *closeness of the covering*. Fejes Tóth pointed out the possibility of other definitions by considering instead of circles and their radii a certain class of convex domains and certain gauges of them.

These definitions (as well as Theorems 1, 2 and 4) can be extended to spaces of higher dimension. The most interesting result concerning these notions is due to BÖRÖCZKY [1], who proved that if in $\mathbf{R}^3$ equal balls form either a closest packing or a closest covering, then the centers of the balls form a body-centered cubic lattice.

In this paper we shall give new definitions of the closeness of packings and coverings with translates of a disc, which will bring into prominence the duality between these notions. Using these definitions, we shall deduce some theorems and we shall scrutinize the relations between our definitions and alternative ones.

DEFINITION 1. Let M be a packing of translates of a disc G, thus $M=\{a_i+G\}$ with $a_i \in R^2$. The *closeness* c of M is defined by

$$\frac{1}{c}=\inf\{\gamma>0\colon\ \{a_i+(1+\gamma)G\} \text{ is a covering}\}.$$

(If this infimum is equal to zero, let $c:=\infty$.)

Roughly speaking, the closeness of a packing measures the extent to which the discs have to be dilated in order to get a covering.

DEFINITION 2. Let M be a covering by translates of a disc G. The *closeness* C of M is then defined by

$$\frac{1}{C}=\inf\{\delta\in(0,1)\colon\ \{a_i+(1-\delta)G\} \text{ is a packing}\}.$$

(If no such δ exists, we set $C:=1$.)

So the closeness of a covering measures the extent to which the discs have to be contracted in order to get a packing.

Let us compare these definitions with the following ones, which resemble more closely the original concepts mentioned in the introduction.

$$\frac{1}{c'}=\sup\{\gamma>0\colon \text{there is a translate of } \gamma G \text{ disjoint to all discs of } M\},$$

$$\frac{1}{C'}=\sup\{\delta>0\colon \text{there is a translate of } \delta G \text{ contained in the intersection of two discs of } M\}.$$

We shall see that $C'=C$ (Theorem 1). But in general $c'\neq c$. For instance, the packing of triangles shown in Fig. 1 has $c=2$ and $c'=4$. For centrally symmetric discs, however, we have $c'=c$. This follows from Theorem 2.

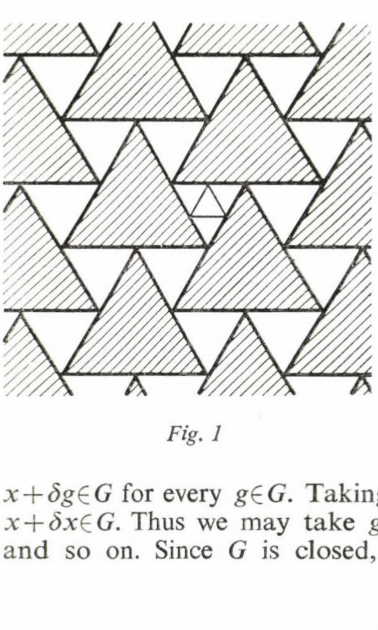


Fig. 1

THEOREM 1. *For the closeness of coverings by translates of a disc G, the two definitions given above coincide, that is*

$$C=C'.$$

PROOF. We assume without restriction of generality that O is an interior point of G. First we show that

$$(1)\qquad \{x\colon x+\delta G\subset G\}=(1-\delta)G$$

for $\delta\in(0,1)$. Because of the convexity of G the right-hand set is contained in the other. The reverse inclusion holds for every closed set G which contains the origin: $x+\delta G\subset G$ means $x+\delta g\in G$ for every $g\in G$. Taking $g=O$ we get $x\in G$. So we may take $g=x$ and get $x+\delta x\in G$. Thus we may take $g=x+\delta x$ and obtain $x+\delta x+\delta^2 x=x+\delta(x+\delta x)\in G$, and so on. Since G is closed, it follows

$$\sum_{n=0}^{\infty}\delta^n x=\frac{1}{1-\delta}x\in G,$$

i.e., $x\in(1-\delta)G$ and the proof of (1) is finished.

For $\delta\in\left(0,\frac{1}{C}\right)$ there is (by definition of C) an $x\in\mathbf{R}^2$ and $j\neq k$, such that $x\in\big(a_j+(1-\delta)G\big)\cap\big(a_k+(1-\delta)G\big)$, i.e., $x+\delta G\subset(a_j+G)\cap(a_k+G)$. On the other hand, let $\delta\in\left(\frac{1}{C},1\right)$. If there existed an x, such that $x+\delta G\subset(a_j+G)\cap(a_k+G)$ for some indices $j\neq k$, we should have $(x-a_j)+\delta G\subset G$. By (1) this yields $x-a_j\in(1-\delta)G$, i.e., $x\in a_j+(1-\delta)G$, and analogously with k instead of j. Therefore $\{a_j+(1-\delta+\varepsilon)G\}$ would not be a packing for every $\varepsilon>0$. So we see that $\frac{1}{C}$ equals the supremum, by which $\frac{1}{C'}$ is defined.

THEOREM 2. *Let M be a packing of translates of a disc G. Then*

$$\frac{1}{c} = \sup\{\gamma > 0: \textit{there is a translate of } -\gamma G \textit{ disjoint to all discs of } M\}.$$

This theorem shows that the difference between the two definitions of the closeness of a packing lies only in the sign of G.

PROOF. Because of the convexity of G we have for $\gamma > 0$:

$$G + \gamma G = (1+\gamma)\,G. \tag{2}$$

For $\gamma \in \left(0, \frac{1}{c}\right)$ there is (by definition of c) an $x \in \mathbf{R}^2$, such that $x \notin a_i + (1+\gamma)G$ for all i. By (2) we see that $x - \gamma G$ is disjoint to all discs $a_i + G$. On the other hand, if $\gamma > \frac{1}{c}$, for every $x \in \mathbf{R}^2$ there is a j, such that $x \in a_j + (1+\gamma)G$, i.e., $(x - \gamma G) \cap \cap (a_j + G) \neq \emptyset$.

By Theorem 2 and a result of Fejes Tóth [4] we obtain immediately

THEOREM 3. *The closeness of a packing of translates of a disc G cannot exceed the closeness of the closest lattice-packing of G.*

($M = \{a_i + G\}$ is called a *lattice-packing,* if the a_i's form a pointlattice.)

Fejes Tóth has proved this theorem for the case that the closeness is measured by the largest member of an arbitrary class of convex subsets of $\mathbf{R}^2$ which is disjoint to all discs. Taking for this class the homothetic images of $-G$ we see that our theorem is merely a special case of Fejes Tóth's. Concerning density there is an analogous theorem due to Fejes Tóth [6] and ROGERS [10] (cf. [4]).

The essential advantage of our definition of closeness lies in the following duality theorem.

THEOREM 4. *Let $c(G)$ be the supremum of the closenesses of all packings of translates of G, and $C(G)$ the corresponding supremum concerning coverings. Then*

$$C(G) = c(G) + 1.$$

PROOF. Let $\{a_i + G\}$ be a covering with closeness C. Since $C(G) > 1$ for every G, we may assume $C > 1$. By definition of C, the discs $a_i + (1-\delta)G$ form a packing for $\delta > \frac{1}{C}$. The closeness of this packing is $\leqq c((1-\delta)G) = c(G)$ (c is invariant under affinities). Thus the discs $a_i + (1+\gamma)(1-\delta)G$ do not cover the plane for $\gamma < \frac{1}{c(G)}$. As $\{a_i + G\}$ is a covering, it follows $(1+\gamma)(1-\delta) < 1$, for all $\gamma < \frac{1}{c(G)}$ and $\delta > \frac{1}{C}$. This yields $\left(1 + \frac{1}{c(G)}\right)\left(1 - \frac{1}{C(G)}\right) \leqq 1$. In a similar manner the reverse inequality may be proved, and the result is

$$\left(1+\frac{1}{c(G)}\right)\left(1-\frac{1}{C(G)}\right)=1,$$

which is equivalent to the assertion.

THEOREM 5. $c(G)\geqq 2$ *and* $C(G)\geqq 3$ *with equality if and only if G is a triangle.*

The proof uses an idea of FÁRY [2], who has proved the corresponding theorem for the density (cf. [3], p. 100). In view of Theorem 4 it is sufficient to consider packings.

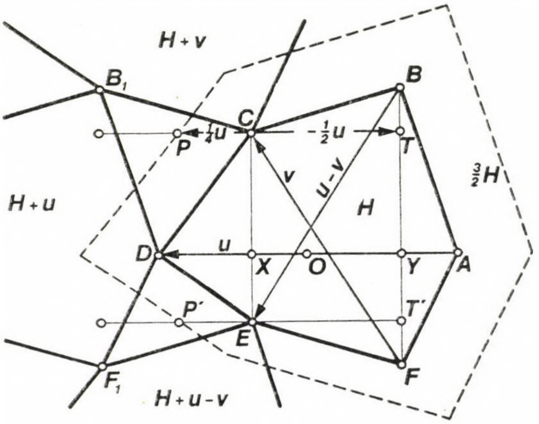


Fig. 2

According to Fáry there is a lattice packing of G with a lattice spanned by two vectors u and v having the following properties: there is a hexagon $H=ABCDEF$ inscribed in G, so that $\overrightarrow{AD}=u$, $\overrightarrow{FC}=v$, $\overrightarrow{BE}=u-v$, $BF\perp u$ and $CE\perp u$ (Fig. 2). To prove the inequality $c(G)\geqq 2$, we have to show that the hexagons $\frac{3}{2}H+$ $+iu+jv$ $(i,\ j\in\mathbf{Z})$ form a covering. We may assume that the origin is the midpoint of AD, so that e.g. $D=\frac{1}{2}u$. Let $B_1:=B+u$ and $F_1:=F+u$. We have to check if the triangles B_1DC and F_1ED are covered. We need only consider one of them, say B_1DC. Let X and Y be the foots of the perpendiculars from C and B to AD. We may assume that $\overline{CX}\leqq\overline{BY}$. In view of the properties of H we see $\overline{XY}=\frac{1}{2}|u|$. Therefore $\overline{CX}\leqq\overline{BY}$ is equivalent to $T:=C-\frac{1}{2}u\in H$. Now let $P:=C+\frac{1}{4}u$.

Since

$$B_1DC\subset DCP\cup B_1DP\cup CB_1P$$

and

$$DC\subset H,\quad B_1D\subset H+u,\quad CB_1\subset H+v,$$

it is sufficient to show that

$$P\in\frac{3}{2}H\cap\left(\frac{3}{2}H+u\right)\cap\left(\frac{3}{2}H+v\right).$$

Now $P\in\frac{3}{2}H$ and $P\in\frac{3}{2}H+v$, as $C+\frac{1}{4}u\in H+\frac{1}{2}H=\frac{3}{2}H$ by (2) and $C=F+v\in H+v$, so $C+\frac{1}{4}u\in H+v+\frac{1}{2}H=\frac{3}{2}H+v$. The remaining relation $P\in\frac{3}{2}H+u$ follows from $T\in H$, since $P=T+u-\frac{1}{4}u$ and $\frac{1}{4}u\in\frac{1}{2}H$.

So the inequality $c(G)\geqq 2$ is proved and we merely have to discuss the case of equality. Let $P':=E+\frac{1}{2}u$. The considered lattice-packing can attain the closeness $c=2$ only if P lies on the boundary of $\frac{3}{2}G$, $\frac{3}{2}G+u$ and $\frac{3}{2}G+v$, or P' lies on the boundary of $\frac{3}{2}G$, $\frac{3}{2}G+u$ and $\frac{3}{2}G+u-v$.

The first of these two conditions can be fulfilled only if the angles of H at A and E are equal to π, i.e., H degenerates into a quadrangle, and G coincides with H. An analogous assertion holds with regard to the second condition. If only one of these conditions is fulfilled, say the first one, we may construct a closer lattice-packing by applying to $G+u$ a slight translation in direction $\overrightarrow{BF}$ and suitable translations to the other discs of M (G is held fixed). So $c=2$ implies both conditions, which means that $G=H$ is a triangle. If G is a triangle, we have indeed $c(G)=2$ (Fig. 1). (According to Theorem 3, it suffices to consider lattice-packings.)

REMARKS. If we replace c by c', Theorem 5 does not hold any more. For a triangle Δ we have $c'(\Delta)=4$; for an equilateral hexagon S with angles alternately equal to 90° and 150°, we have $c'(S)=2+\sqrt{3}<4$ (Fig. 3). Presumably this is the minimal value of $c'(G)$. By the way, this hexagon S has the special property, that the same closeness takes place for all lattice-packings, in which three mutually touching discs occur. We might say, S is a disc with *direction-invariant closeness* (cf. [7] for the corresponding notion concerning density).

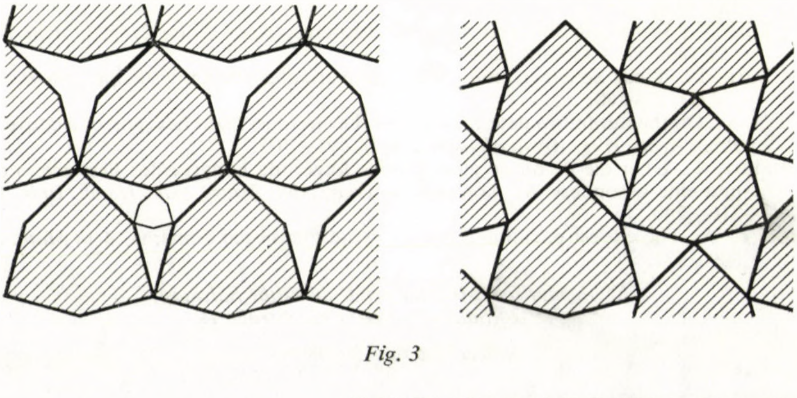


Fig. 3

Restricting ourselves to centrally symmetric discs, the determination of the minimum of $c(G)$ seems to be more difficult.

CONJECTURE. *For a centrally symmetric disc G,*

$$c(G) \geqq 3+2\sqrt{2} \quad \textit{and} \quad C(G) \geqq 4+2\sqrt{2},$$

with equality if and only if G is an affinely regular octagon. (Such octagons are discs with direction-invariant closeness, too.)

Here the statements on $c(G)$ and $C(G)$ are equivalent. Concerning the density, the corresponding problem is solved for coverings only, with the ellipses as extremal discs [3]. REINHARDT [9] and MAHLER [8] (Cf. [3], p. 104) constructed a "smoothed octagon" for which the density of the densest lattice-packing is conjectured to be minimal.

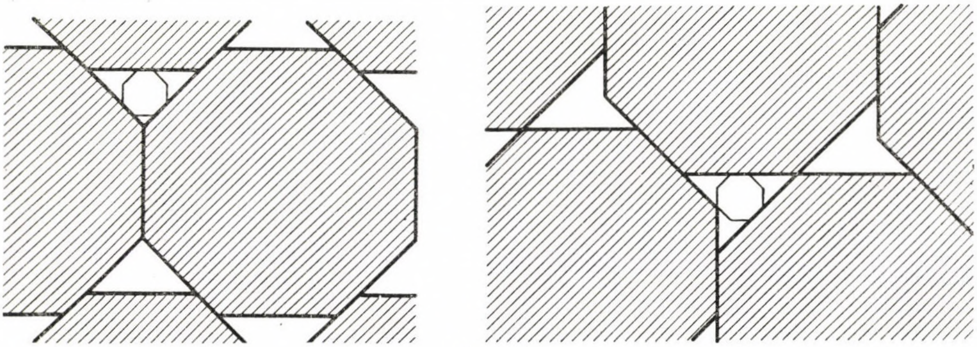


Fig. 4

REFERENCES

[1] BÖRÖCZKY, K., Close packing of spheres, *Acta Math. Acad. Sci. Hungar.* (to appear).

[2] FÁRY, I., Sur la densité des réseaux de domaines convexes, *Bull. Soc. Math. France* **78** (1950), 152—161.

[3] FEJES TÓTH, L., *Lagerungen in der Ebene, auf der Kugel und im Raum*, 2. Aufl., Springer-Verlag, Berlin—Heidelberg—New York, 1972.

[4] FEJES TÓTH, L., Remarks on the closest packing of convex discs, *Comment. Math. Helvetici* **53** (1978), 536—541.

[5] FEJES TÓTH, L., Close packing and loose covering with balls, *Publ. Math. Debrecen* **23** (1976), 323—326.

[6] FEJES TÓTH, L., Some packing and covering theorems, Acta Univ. Szeged, *Acta Sci. Math. (Szeged)* **12**/A (1950), 62—67.

[7] FEJES TÓTH, L., Über Scheiben mit richtungsinvarianter Packungsdichte, *Elem. Math.* **26** (1971), 58—60.

[8] MAHLER, K., On the minimum determinant and the circumscribed hexagon of a convex domain, *Proc. Acad. Wet. Amsterdam* **50** (1947), 695—703.

[9] REINHARDT, K., Über die dichteste gitterförmige Lagerung kongruenter Bereiche in der Ebene und eine besondere Art konvexer Kurven, *Abh. Math. Sem. Hamb. Univ.* **10** (1934), 216—240.

[10] ROGERS, C. A., The closest packing of convex two-dimensional domains, *Acta Math.* **86** (1951), 309—321.

Mathematisches Institut der Universität Salzburg,
A—5020 Salzburg, Petersbrunnstraße 19

(Received June 20, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 163—166.

ON THE DIVERGENCE OF ORTHOGONAL SERIES

by

A. P. SÖVEGJÁRTÓ

1. Let S denote the set of Lebesgue measurable, almost everywhere finite functions on the interval $(0, 1)$. Let $T=\|t_{i,j}\|_0^\infty$ be a matrix such that

$$(1) \qquad |t_{ij}| \leqq K(<\infty) \quad (i, j = 0, 1, \ldots), \qquad \lim_{i\to\infty} t_{i,j} = 1 \ (j = 0, 1, \ldots),$$

and let $f=\{f_k(x)\}_0^\infty$ be a sequence of functions belonging to S. A series

$$(2)' \qquad \sum_{k=0}^{\infty} c_k f_k(x)$$

is said to be *T-summable* in measure (almost everywhere) if the series

$$t_i(x) = \sum_{k=0}^{\infty} t_{i,k} c_k f_k(x) \qquad (i = 1, 0, \ldots)$$

converge in measure (almost everywhere) and the sequence $\{t_i(x)\}_0^\infty$ converges in measure (almost everywhere) to a function belonging to S.

The system f is said to be a *T-convergence system* in measure (T-convergence system) for l_2 if for every $c=\{c_k\}_0^\infty\in l_2$ the series (2) is T-summable in measure (T-summable almost everywhere).

The system f is said to be a *convergence system* in measure (almost everywhere) for l_2 if $c\in l_2$ implies the convergence of the series (2) in measure (almost everywhere).

Joó [2] proved a general theorem which contains the following statement as a special case:

Let T be a matrix satisfying conditions (1). *If the system f is a T-convergence system in measure for l_2, then it is also a convergence system in measure for l_2.*

A natural question is whether a similar statement is true for almost everywhere convergence.

Let $\nu=\{\nu_n\}_0^\infty$ be a strictly increasing sequence of non-negative integers, $\nu_0=0$. We call T_ν the *summation process* generated by a matrix $\|t_{i,k}\|$ of the form

$$t_{i,k} = 1 \quad (k = 0, 1, \ldots, \nu_i), \quad t_{i,k} = 0 \quad (k = \nu_i+1, \nu_i+2, \ldots) \quad (i = 0, 1, \ldots).$$

The T-summation is said to be *equivalent* to T_ν-summation if for every $c\in l_2$ and for every orthonormal system $\varphi=\{\varphi_k(x)\}_0^\infty$ on $(0, 1)$ the orthogonal series

$$(3) \qquad \sum_{k=0}^{\infty} c_k \varphi_k(x)$$

is T-summable almost everywhere if and only if it is T_ν-summable almost everywhere. (We recall the fact that, e.g., $(C, 1)$ summability is equivalent to $T_{\{2^\nu\}}$ summability; see, e.g., ALEXITS [1], p. 118.)

I. Joó and K. TANDORI answered the above-mentioned question in [3]. To state their result let ν be a sequence of indices such that $\overline{\lim\limits_{n\to\infty}}\,(\nu_{n+1}-\nu_n)=\infty$. Let T be a summation process equivalent to T_ν. Then there exists an orthonormal system $\Phi=\{\varphi_k(x)\}_0^\infty$ on $(0, 1)$, which is a T-convergence system for l_2 but is not a convergence system for l_2, indeed, there exists a sequence $c\in l_2$ such that the series (3) diverges almost everywhere.

They remarked that the system Φ is obtained by a rearrangement of the Walsh system $\{w_n(x)\}_0^\infty$. Using ideas of F. MÓRICZ [4] it is easy to see that one can obtain an orthonormal system, with similar properties, also by rearrangement of the trigonometrical system $\{1, \cos 2\pi x, \sin 2\pi x, \ldots\}$.

In this paper we shall prove the following generalization of the theorem of I. Joó and K. Tandori:

THEOREM. *Let $\{\varphi_n\}_0^\infty$ be a complete orthonormal system on $(0, 1)$, and $\{\nu_n\}_0^\infty$ be a sequence of natural numbers, for which $\overline{\lim\limits_{n\to\infty}}\,(\nu_{n+1}-\nu_n)=+\infty$. Then there exists a rearrangement $\{s(r)\}_{r=1}^\infty$ of the natural numbers, such that the system $\{\varphi_{s(r)}\}_0^\infty$ is a T_ν-convergence system for l_2, but is not a convergence system for l_2. Indeed there exists a continuous function f on the interval $(0, 1)$, for which*

$$\overline{\lim_{N\to\infty}}\left|\sum_{r=1}^{N}(f, \varphi_{s(r)})\varphi_{s(r)}(x)\right|=+\infty$$

almost everywhere on $(0, 1)$.

2. For the proof of our theorem we need some known results.

LEMMA 1 (OLEVSKIĬ [5]). *Let $\{\varphi_n\}_0^\infty$ be an arbitrary complete orthonormal system on $(0, 1)$. Then there exist a rearrangement $\{n(l)\}$ of the natural numbers and a function $f\in C(0, 1)$ such that*

$$\overline{\lim_{N\to\infty}}\left|\sum_{l=1}^{N}(f, \varphi_{n(l)})\varphi_{n(l)}(x)\right|=+\infty \tag{4}$$

almost everywhere on $(0, 1)$.

LEMMA 2 (MENŠOV [6]). *For arbitrary orthonormal system $\{\varphi_n\}_0^\infty$ defined on the interval $(0, 1)$ there exists subsequence m_n of natural numbers such that, for any sequence $\{a_n\}\in l_2$ the limit*

$$\lim_{N\to\infty}\sum_{n=1}^{m_N}a_n\varphi_n(x)$$

exists almost everywhere on $(0, 1)$.

From this one can obtain easily the following

LEMMA 3 (Menšov [7]). *From arbitrary infinite orthonormal system* $\{\varphi_n\}_0^\infty$ *on* $(0, 1)$ *we can choose an infinite convergence system* $\{\varphi_{n_k}\}_0^\infty$.

Returning to the proof of the Theorem, first apply Lemma 1. Denote $\{n(l)\}$ and f the resulting rearrangement and function, respectively. Pick a monotone increasing sequence of natural numbers $\{N_k\}$ such that the estimate

$$\max_{N_{k-1}<N<N_k} \left| \sum_{l=N_{k-1}+1}^{N} (f, \varphi_{n(l)})\varphi_{n(l)}(x) \right| \geqq 2^k \tag{5}$$

holds true for $x \in E_k (\subset (0, 1))$, mes $E_k > 1 - \frac{1}{2^k}$. According to Lemma 3 there exists a sequence $\{k_i\}$ of natural numbers with infinite elements such that the system $\{\varphi_{n(N_{k_i})}\}$ is a convergence system. According to Lemma 2 there exists a sequence $\{m_n\}_1^\infty$ of natural numbers such that the limit

$$\lim_{N\to\infty} \sum_{\substack{l=1\\ l\neq N_{k_i}}}^{m_N} a_l \varphi_{n(l)}(x) \tag{6}$$

exists almost everywhere on $(0, 1)$ for every $\{a_l\} \in l_2$. We may suppose that for every n there exists an index k such that

$$m_n \leqq N_k, \quad m_{n+1} \geqq N_{k+1}.$$

After this let n_1 be the smallest number such that

$$p_{n_1+1} - p_{n_1} \geqq m_2 - m_1.$$

Define the rearrangement $\{s(r)\}_1^\infty$ in our theorem by induction, in the following way. Let

$$\varphi_{s(r)} \overset{\text{def}}{=} \varphi_{n(N_{k_r})} \quad \text{for} \quad 1 \leqq r \leqq P_{n_1},$$

$$\varphi_{s(r)} \overset{\text{def}}{=} \varphi_{n(m_1+[r-p_{n_1}])} \quad p_{n_1} < r \leqq p_{n_1} + (m_2 - m_1).$$

Let n_2 be the first $n\ (> n_1)$ such that

$$p_{n_2+1} - p_{n_2} \geqq m_3 - m_2.$$

Define

$$\varphi_{s(r)} \overset{\text{def}}{=} \varphi_{n(N_{k_{p_{n_1}+r}})} \quad \text{for} \quad p_{n_1} + (m_2 - m_1) < r \leqq p_{n_2}$$

further

$$\varphi_{s(r)} \overset{\text{def}}{=} \varphi_{n(m_2+[r-p_{n_2}])} \quad p_{n_2} < r \leqq p_{n_2} + (m_3 - m_2).$$

Continuing this process we obtain the rearrangement $s(r)$ satisfying the requirements of our theorem.

REFERENCES

[1] ALEXITS, G., *Convergence Problems of Orthogonal Series,* Akadémiai Kiadó, Budapest, 1961.

[2] JOÓ, I., A remark on convergence systems in measure, *Acta Sci. Math. (Szeged)* **38** (1976), 301—303.

[3] JOÓ, I.—TANDORI, K., A remark on convergence of orthogonal series, *Acta Sci. Math. (Szeged)* **38** (1976), 305—309.

[4] MÓRICZ, F., On the order of magnitude of the partial sums of rearranged Fourier series of square integrable functions, *Acta Sci. Math. (Szeged)* **28** (1967), 155—167.

[5] OLEVSKIĬ, A. M., *Fourier series with respect to general orthonormal systems,* Springer-Verlag, Berlin, 1975.

[6] MENSHOV, D. E., Суммирование рядов но ортономина функциям линейными методами, *Изв. АН СССР, сер. матем.,* (1937), 203—230.

[7] MENSHOV, D. E., Sur la convergence et la sommation des séries de fonctions orthogonales, *Bull. Soc. Math. France* **64** (1936), 147—170.

H—9700 Szombathely, Szinyei Merse u. 3

(Received June 27, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 167—171.

ANALOGON EINES SATZES VON BAER UND LEVI IN DER KLASSE ALLER HALBGRUPPEN

von

WOLFGANG ROSENOW

Ein Satz von BAER und LEVI besagt, daß eine Gruppe niemals gleichzeitig direktes und freies Produkt von Gruppen sein kann. MÁRKI formulierte 1976 in Szeged das Problem, ob die analoge Aussage für Halbgruppen gültig ist ([2] S. 752). Im folgenden wird dieses Problem positiv gelöst. Zur Erklärung der verwendeten Begriffe wird auf CLIFFORD und PRESTON [1] verwiesen.

SATZ. *Eine Halbgruppe kann nicht gleichzeitig als nichttriviales direktes und freies Produkt von Halbgruppen dargestellt werden.*

BEWEIS. Angenommen es sei $A*B=C\times D$, wobei $A*B$ das freie Produkt der Halbgruppen A und B und $C\times D$ das direkte Produkt der Halbgruppen C und D sind. Dabei gilt $|C|>1$ und $|D|>1$. Im folgenden werden Elemente ab aus $A*B$ mit $a\in A$ und $b\in B$ betrachtet.

Fall 1. Es sei $ab=(c, d)$ mit $(c, d)\in C\times D$ und $c\in C$ ist idempotent $(c^2=c)$.

Fall 1.1. Es gelte auch $d^2=d$.
Dann folgt $ab=(c, d)=(c^2, d^2)=(c, d)^2=abab$. Wegen der Eindeutigkeit der Darstellung der Elemente in $A*B$ ist das ein Widerspruch.

Fall 1.2. Es sei $d^2\neq d$.

Wegen $|C|>1$ existiert ein $c_1\in C$ mit $c\neq c_1$. Jetzt sei $(c_1, d)=\prod_{i=1}^{n}{}^{*} a_ib_i$, wobei $a_i\in A$ und $b_i\in B$ für alle $i=1, \dots, n$ gilt. Außerdem bedeutet *, daß in $\prod_{i=1}^{n}{}^{*} a_ib_i$ die Elemente a_1 und b_n eventuell nicht auftreten.

Fall 1.2.1. $(c_1, d)=a_1$.
Dann gilt wegen

$$(c, d)(c_1, d)(c, d)^2 = (c, d)^2(c_1, d)(c, d) \tag{1}$$

die Gleichung $ab(a_1a)bab=abab(a_1a)b$ und damit $a_1a=a$. (Die Schreibweise $\dots(a_1a)\dots$ bedeutet, daß an der eingeklammerten Stelle bei der Multiplikation von Elementen aus $A*B$ eine Verschmelzung der Randkomponenten der Faktoren auftritt.) Weiterhin folgt $(c_1, d)(c, d)=(a_1a)b=ab=(c, d)$. Damit gilt $d^2=d$ und das ist ein Widerspruch.

AMS (MOS) subject classifications (1979). Primary 20M10; Secondary 20E06.
Key words and phrases. Semigroups, free product, direct product.

Fall 1.2.2. $(c_1, d)=b_1$.

Es gilt wegen (1) die Gleichung $a(bb_1)abab=aba(bb_1)ab$, woraus $bb_1=b$ folgt. Man erhält jetzt $(c, d)(c_1, d)=a(bb_1)=ab=(c, d)$. Das ergibt auch hier den Widerspruch $d=d^2$.

Fall 1.2.3. $(c_1, d)=\prod_{i=1}^{n}{}^{*}\, a_ib_i$ beginnt mit einem Element aus A und endet mit einem Element aus B.

Es sei also $(c_1, d)=a_1b_1\dots a_nb_n$. Dann folgt wegen (1) $aba_1b_1\dots a_nb_nabab=$ $=ababa_1b_1\dots a_nb_nab$. Damit gilt $a=a_i$ und $b=b_i$ für alle $i=1, \dots, n$. Jetzt ergibt sich

$$(c_1, d) = a_1b_1\dots a_nb_n = (ab)^n = (c, d)^n = (c, d^n).$$

Laut Voraussetzung ist $c_1=c$ ein Widerspruch.

Fall 1.2.4. $(c_1, d)=\prod_{i=1}^{n}{}^{*}\, a_ib_i$ beginnt mit einem Element aus B und endet mit einem Element aus A.

Es sei dann $(c_1, d)=b_1a_2b_2\dots a_{n-1}b_{n-1}a_n$. Dann gilt wegen (1)

$$a(bb_1)a_2b_2\dots a_{n-1}b_{n-1}(a_na)bab = aba(bb_1)a_2b_2\dots b_{n-1}(a_na)b.$$

Man erhält $bb_1=b$, $a_na=a$, $a=a_i$, $b=b_i$ für alle $i=2, \dots, n-1$. Somit ist $(c_1, d)=$ $=b_1(ab)^{n-2}a_n$. Weiter gilt $(cc_1c^2, d^4)=(ab)^{n+1}=(c, d)^{n+1}=(c, d^{n+1})$, woraus folgt $d^{n+1}=d^4$. Es ist $(ab)^4=(c, d^4)=(c, d^{n+1})=(ab)^{n+1}$ und damit gilt $n=3$ und schließlich $(c_1, d)=b_1aba_n$.

Es wird jetzt betrachtet $(c, d)(c_1, d^2)=ab\prod_{i=1}^{m}{}^{*}\, x_iy_i$, wobei $x_i\in A$ und $y_i\in B$ gilt. Es ist $(c, d)(c_1, d^2)=(c, d)^2(c_1, d)=abababa_n$. Dann gilt $\prod_{i=1}^{m}{}^{*}\, x_iy_i=b'ababa_n$ mit $bb'=b$ oder $\prod_{i=1}^{m}{}^{*}\, x_iy_i=ababa_n$. $\prod_{i=1}^{m}{}^{*}\, x_iy_{:}=ababa_n$ kann nicht sein, denn es gilt $(c_1, d)(c_1, d^2)=(c_1, d^2)(c_1, d)$, aber es gilt nicht

$$b_1ab(a_na)baba_n = ababa_nb_1aba_n.$$

Aus $(c_1, d^2)=\prod_{i=1}^{m}{}^{*}\, x_iy_i=b'ababa_n$ folgt dann

$$b_1aba_nb'ababa_n = (c_1, d)(c_1, d^2) = (c_1, d^2)(c_1, d) = b'ababa_nb_1aba_n.$$

Damit gilt $b_1=b'=b$ und $a_n=a$, woraus folgt $b^2=b$ und $a^2=a$. Es ist dann $(c_1, d^2)^2=babababababa=(baba)^3=(c_1, d)^3$ und damit folgt $d^4=d^3$. Jetzt gilt $(ab)^3=(c, d^3)=(c, d^4)=(ab)^4$. Das ist ein Widerspruch.

Fall 1.2.5. $(c_1, d)=\prod_{i=1}^{m}{}^{*}\, a_ib_i$ beginnt und endet mit einem Element aus A.

Es sei $(c_1, d)=a_1b_1\dots a_{n-1}b_{n-1}a_n$. Dann gilt wegen (1)

$$aba_1b_1\dots a_{n-1}b_{n-1}(a_na)bab = ababa_1b_1\dots a_{n-1}b_{n-1}(a_na)b.$$

Hieraus folgt $a_n a = a, a = a_i$ und $b = b_i$ für $i = 1, \dots, n-1$. Also ist $(c_1, d) = (ab)^{n-1} a_n$. Weiterhin gilt

$$(cc_1c^2, d^4) = ab(ab)^{n-1}(a_n a)bab = (ab)^{n+2} = (c, d)^{n+2} = (c, d^{n+2}).$$

Es ist damit $d^4 = d^{n+2}$, woraus sich folgendes ergibt: $(ab)^{n+2} = (c, d)^{n+2} = (c, d^{n+2}) = (c, d^4) = (c, d)^4 = (ab)^4$. Wegen der Eindeutigkeit der Darstellung der Elemente in $A * B$ ist jetzt $n = 2$. Das heißt, es ist $(c_1, d) = aba_n$. Es wird jetzt betrachtet $(c, d)(c_1, d^2) = ab \prod_{i=1}^{m}{}^{*} x_i y_i$ mit $x_i \in A$ und $y_i \in B$. Dann gilt

$$(c, d)(c_1, d^2) = (c, d)^2 (c_1, d) = abababa_n.$$

Demnach ist $\prod_{i=1}^{m}{}^{*} x_i y_i = ababa_n$ oder $\prod_{i=1}^{m}{}^{*} x_i y_i = b' ababa_n$ mit $bb' = b$.

$\prod_{i=1}^{m}{}^{*} x_i y_i = b' ababa_n$ ist nicht möglich, denn es gilt $(c_1, d)(c_1, d^2) = (c_1, d^2)(c_1, d)$, aber es gilt nicht $aba_n b' ababa_n = b' abab(a_n a)ba_n$. Jetzt folgt $(c_1, d^2) = ababa_n = ab(a_n a)ba_n = (c_1, d)(c_1, d)$. Damit ist c_1 idempotent $(c_1 = c_1^2)$.

Fall 1.2.6. $(c_1, d) = \prod_{i=1}^{n}{}^{*} a_i b_i$ beginnt und endet mit einem Element aus B.

Zunächst kann eine Rechnung analog zum Fall 1.2.5 erfolgen. In der Darstellung nach (1) tritt dann ebenfalls eine Verschmelzung von zwei Elementen auf, die hier Elemente aus B sind. Schließlich erhält man $(c_1, d) = b_1 ab$ und $(c_1, d^2) = b_1 abab$ mit $bb_1 = b$. Dann folgt weiter $(c_1, d^2) = b_1 abab = b_1 a(bb_1)ab = (c_1, d)^2$ und auch in diesem Fall gilt $c_1 = c_1^2$.

Aus den Fällen 1.2.3 und 1.2.4 ergibt sich, daß die dort angenommenen Elemente c_1 nicht existieren. Also kann ein solches Element c_1 aus C nur zu Darstellungen entsprechend den Fällen 1.2.5 und 1.2.6 führen. Das bedeutet, alle Elemente aus C sind idempotent. Dann folgt für $c_1 \in C$ mit $(c_1, d) = aba_n$ (Fall 1.2.5) und für $c_2 \in C$ mit $(c_2, d) = b_1 ab$ (Fall 1.2.6) die Gleichung $c_1 c_1 c_2 c_2 = c_1^2 c_2^2 = c_1 c_2 = (c_1 c_2)^2 = c_1 c_2 c_1 c_2$.

Man erhält jetzt einen Widerspruch, denn es gilt $(c_1 c_1 c_2 c_2, d^4) = (c_1 c_2 c_1 c_2, d^4)$, aber es gilt nicht $ababa_n b_1 abab = aba_n b_1 ababa_n b_1 ab$. Somit sind alle (c_i, d) mit $c_i \in C$ entweder von der Form $(c_i, d) = aba_n$ (Fall 1.2.5) oder alle (c_i, d) mit $c_i \in C$ sind von der Form $(c_i, d) = b_1 ab$ (Fall 1.2.6). Es wird zunächst der Fall 1.2.5 angenommen.

Wegen $(c_1, d) = aba_n$ gilt $(c_1, d)^2 = ab(a_n a)ba_n = ababa_n = (c, d)(c_1, d)$, woraus mit $c_1 = c_1^2$ folgt

$$c_1 = cc_1. \tag{2}$$

Wegen $(c_1, d)(c, d) = ab(a_n a)b = abab = (c, d)^2 = (c, d^2)$ gilt aber auch

$$c_1 c = c. \tag{3}$$

Die Gleichungen (2) und (3) gelten für alle $c_1 \in C$. Daher gilt für beliebige $c_i, c_j \in C$ die folgende Gleichung:

$$c_j c_i = c_j(cc_i) = (c_j c)c_i = cc_i = c_i.$$

Damit ist C eine Rechtsnullhalbgruppe. Es seien jetzt $a=(c_a, d_a)$ und $b=(c_b, d_b)$. Dann gilt $ab=(c_a c_b, d_a d_b)=(c_b, d_a d_b)=(c, d)$. Also ist $b=(c, d_b)$ und $b^2=(c, d_b)(c, d_b)=(c, d_b)^2$. Ist ferner w ein Element aus $A*B$ mit $w=(c', d_b)$ und $c'\neq c$, dann gilt $wb=(c', d_b)(c, d_b)=(c, d_b^2)=b^2$. Es folgt $w\in B$ mit $w\neq b$. Jetzt ergibt sich der folgende Zusammenhang:

$$bab=(cc_a c, d_b d_a d_b)=(cc, d_b d_a d_b)=(c, d_b d_a d_b)=$$
$$=(c'c, d_b d_a d_b)=(c' c_a c, d_b d_a d_b)=wab.$$

Wegen $b\neq w$ ist das ein Widerspruch.

Damit ist klar, daß auch im Fall 1.2.5 das angenommene Element c_1 nicht existieren kann. Durch analoge Überlegungen läßt sich das genauso für den Fall 1.2.6 nachweisen. Das bedeutet jetzt, daß es in C kein $c_1\neq c$ geben kann, was ein Widerspruch ist.

Fall 2. Es sei $ab=(c, d)$ mit $(c, d)\in C\times D$ und weder c noch d seien idempotent.

Es wird betrachtet $(c, d^2)=\prod\limits_{i=1}^{n}{}^{*} a_i b_i$. Dann gilt $ab\prod\limits_{i=1}^{n}{}^{*} a_i b_i=(c, d)(c, d^2)=(c, d^2)(c, d)=\left(\prod\limits_{i=1}^{n}{}^{*} a_i b_i\right)ab$. Aus der Eindeutigkeit der Darstellung der Elemente in $A*B$ folgt:

$\prod\limits_{i=1}^{n}{}^{*} a_i b_i$ beginnt mit einem Element aus A, nämlich a.

$\prod\limits_{i=1}^{n}{}^{*} a_i b_i$ endet mit einem Element aus B, nämlich b.

Fall 2.1. $n=1$.

Dann ist $(c, d^2)=ab=(c, d)$. Das ist ein Widerspruch, denn d ist nicht idempotent.

Fall 2.2. $n>1$.

Aus $ab\prod\limits_{i=1}^{n}{}^{*} a_i b_i=\left(\prod\limits_{i=1}^{n}{}^{*} a_i b_i\right)ab$ folgt $a=a_i$ und $b=b_i$ für $i=1, \dots, n$. Es gilt weiter $(c, d^2)=(ab)^n=(c, d)^n$ und damit ist $c=c^n$ und $d^2=d^n$.

Fall 2.2.1. $n=2$.

Dann ist c idempotent und das ist ein Widerspruch.

Fall 2.2.2. $n=3$.

Aus $d^2=d^3$ folgt $d^3=d^4$ und damit $d^2=d^4$. Weiter gilt $(ab)^6=((ab)^3)^2=(c, d^2)^2=(c^2, d^4)=(c^2, d^2)=(ab)^2$. In $A*B$ ist aber $(ab)^6=(ab)^2$ nicht möglich.

Fall 2.2.3. $n=4$.

Mit $d^2=d^4$ folgt $(ab)^8=((ab)^4)^2=(c, d^2)^2=(c^2, d^4)=(c^2, d^2)=(ab)^2$. Das ist ebenfalls ein Widerspruch.

Fall 2.2.4. $n>4$.

Aus $d^2=d^n$ folgt $d^2d^{n-4}=d^n d^{n-4}$ und damit $d^{n-2}=d^{2n-4}$. Wegen $d^{2n-4}=(d^{n-2})^2$ ist d^{n-2} idempotent. Dann gilt $(ab)^{n(n-2)}=((ab)^n)^{n-2}=(c, d^2)^{n-2}=$

$=(c^{n-2}, d^{n-2})=(ab)^{n-2}$. Für $n>4$ gilt $n(n-2)\neq n-2$. Das ergibt auch hier einen Widerspruch.

Es wurde nachgewiesen, daß es für das Element $ab\in A*B$ kein Element (c, d) gibt, so daß $ab=(c, d)$ gilt. Dann muß die Annahme $A*B=C\times D$ falsch sein. Also gilt für die Klasse der Halbgruppen, daß sich eine beliebige Halbgruppe niemals gleichzeitig in ein freies und in ein direktes Produkt von Halbgruppen zerlegen läßt.

LITERATUR

[1] Clifford, A. H.—Preston, G. B., *The Algebraic Theory of Semigroups*, Vol. I, Providence, 1964. (Russische Übersetzung, Moskau, 1972.)

[2] Márki, L., Problem 3, in: *Algebraic Theory of Semigroups* (Proc. Confer. Szeged, Hungary, 1976), Colloq. Math. Soc. J. Bolyai 20, North-Holland Publishing Co., Amsterdam, 1979.

Pädagogische Hochschule "Liselotte Herrmann"
Sektion Mathematik/Physik
DDR—26 Güstrow

(Eingegangem am 17. Juli 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 173—180.

SUR CERTAINS CHANGEMENTS DE VARIABLE DES SÉRIES DE FABER

par

L. ALPÁR

§ 1. Introduction

Dans cette note nous allons étendre aux séries de Faber un résultat obtenu pour les séries de Taylor [3].

Les séries de Faber constituent une généralisation des séries de Taylor. On les décrira plus amplement dans le § 2.

Notre point de départ est le problème suivant soulevé et résolu par P. TURÁN [10]. Soit ζ $(0<|\zeta|<1)$ un paramètre complexe, $h(z)=e^{i\tau_0}(z-\zeta)/(1-\bar{\zeta}z)$, τ_0 étant une constante réelle et

$$(1.1)\qquad f_1(z)=\sum_{\nu=0}^{\infty}a_\nu z^\nu$$

une fonction holomorphe pour $|z|<1$. La fonction

$$(1.2)\qquad f_1[h(z)]=f_2(z)=\sum_{n=0}^{\infty}b_n z^n \quad (b_n=b_n(\zeta))$$

est alors également analytique pour $|z|<1$. Admettons de plus qu'il existe un point z_1 $(|z_1|=1)$ où la série (1.1) converge et, par suite, $f_1(z_1)$ a une valeur déterminée. Soit z_2 le point défini par l'égalité $z_1=h(z_2)$, donc $|z_2|=1$ et $f_1(z_1)=f_2(z_2)$. TURÁN a prouvé qu'il existe f_1 telle que malgré la convergence de la série $\sum a_\nu z_1^\nu$, la série $\sum b_n z_2^n$ diverge. En d'autres termes *la convergence locale des séries de puissances sur leur circonférence de convergence n'est pas un invariant conforme.*

Nous avons généralisé ce théorème de TURÁN de différentes manières, d'abord pour les séries de Taylor, ensuite nous avons étendu certains de ces résultats aux séries de Faber. Premièrement nous avons montré [2] qu'une proposition analogue à celle de TURÁN a lieu également pour les séries de Faber. La démonstration de ce théorème a rendu nécessaire de suivre une voie particulière qui ne suit pas de la théorie classique des séries de Faber.

Cependant nous ne sommes pas parvenus, même à l'aide de ces raisonnements nouveaux, à étendre aux séries de Faber notre deux résultats suivants. La sommabilité (C, k) $(k\geqq 0)$ de la série $\sum a_\nu z_1^\nu$ entraîne toujours la sommabilité $(C, k+1/2)$ de la série $\sum b_n z_2^n$ ([1] Théorème 1, p. 100). D'autre part, soient donnés d'avance les paramètres ζ, k et δ $(0\leqq\delta<1/2)$, il existe $f_1=f_1(z;\,\zeta,k,\delta)$ telle que $\sum a_\nu z_1^\nu$ soit sommable (C, k), mais que $\sum b_n z_2^n$ ne soit pas sommable $(C, k+\delta)$ avec le δ donné ([1] Théorème 2, pp. 100—101). Le résultat de TURÁN est un cas particulier de ce dernier théorème pour $k=0$, $\delta=0$.

Néanmoins tout récemment nous avons trouvé [3] une démonstration nouvelle d'un cas particulier du Théorème 2 cité plus haut où $k=0$ et $0\leqq\delta<1/2$.

Les moyens élaborés dans [2] et [3] permettent tout de même de prouver le théorème analogue pour les séries de Faber. C'est l'objet du travail présent.

Dans le § 2 nous récapitulons certains résultats concernant les séries de Faber, dans le § 3 nous citons deux lemmes prouvés dans [3], dans le § 4 nous allons formuler notre théorème d'une façon précise et exposons sa démonstration.

Les c_j $(j=0, 1, 2, \dots)$ désignent des constantes numériques positives.

§ 2. Sur les polynômes et les séries de Faber

Nous rappelons sans démonstration quelques propriétés des polynômes et des séries de Faber (cf. [4], [5], [6], [7], [8], [9], [11]).

Soit C une courbe fermée simple dans le plan des z formée d'un seul arc analytique régulier, avec l'intérieur $\mathrm{I(C)}$, l'extérieur $\mathrm{E(C)}$; $\bar{\mathrm{I}}(\mathrm{C})$ et $\bar{\mathrm{E}}(\mathrm{C})$ dénotent leur fermeture respective. La même notation sera employée aussi dans les cas d'autres courbes fermées.

Il existe une fonction unique $w=\varphi(z)$ qui applique d'une manière conforme et biunivoque $\bar{\mathrm{E}}(\mathrm{C})$ sur $\bar{\mathrm{E}}(\mathrm{K}_R)$ du plan des w où K_R désigne la circonférence $|w|=R$; pour des $|z|$ suffisamment grands $\varphi(z)$ s'écrit sous la forme

$$w = \varphi(z) = z+\alpha_0+\frac{\alpha_1}{z}+\frac{\alpha_2}{z^2}+\dots \tag{2.1}$$

où les quantités $R, \alpha_0, \alpha_1, \dots$ sont déterminées univoquement par les conditions imposées à $\varphi(z)$.

La fonction inverse de $\varphi(z)$ soit notée

$$z = \psi(w) = w+\beta_0+\frac{\beta_1}{w}+\frac{\beta_2}{w^2}+\dots. \tag{2.2}$$

La série (2.2) converge pour $r<|w|<\infty$ avec un $r<R$. Les images des circonférences K_ϱ: $|w|=\varrho$ $(\varrho>r)$ par $\varphi(z)$ sont les courbes de niveau C_ϱ: $|\varphi(z)|=\varrho$; on écrira aussi C_R au lieu de C. Désignons par C_r la frontière de l'ensemble $\bigcup_{\varrho>r} \bar{\mathrm{E}}(\mathrm{C}_\varrho)$. L'anneau circulaire limité par K_R et K_r ainsi que le domaine annulaire de frontière composée de C_R et C_r jouent un rôle important dans la théorie des séries de Faber.

Lorsque $\varphi(z)$ est donnée par la série (2.1), $\Phi_n(z)$, le n-ième polynôme de Faber associé à C_R, est la partie polynomiale de l'expression

$$\varphi^n(z) = \Phi_n(z)+R_n(z), \tag{2.3}$$

$R_n(z)$ ne contient que des puissances négatives de z. Il résulte ainsi de (2.3) que $\Phi_0(z)\equiv 1$ et $\Phi_n(z)\equiv 0$ pour $n<0$.

Toute fonction $F_1(z)$ holomorphe dans $\mathrm{I}(\mathrm{C}_R)$ peut être représentée par sa série de Faber unique:

$$F_1(z) = \sum_{\nu=0}^{\infty} A_\nu \Phi_\nu(z) \tag{2.4}$$

qui est uniformément et absolument convergente dans chaque sous-ensemble fermé de $\mathrm{I}(\mathrm{C}_R)$. Les polynômes Φ_ν, comme on le voit, sont indépendants de F_1 et sont entièrement déterminés par C_R, les coefficients A_ν dépendent aussi de F_1 et s'expriment par l'intégrale

$$(2.5) \qquad A_\nu = \frac{1}{2\pi i} \int_{K_\varrho} \frac{F_1[\psi(w)]}{w^{\nu+1}}\, dw, \quad r < \varrho < R, \ \nu \geqq 0.$$

C_R est la courbe de convergence de la série (2.4), si F_1 a de singularités sur C_R. Inversement, étant donné une suite de nombres $\{A_\nu\}_{\nu=0}^{\infty}$ telle que l'on ait

$$(2.6) \qquad \overline{\lim_{\nu\to\infty}}\, |A_\nu|^{1/\nu} \leqq \frac{1}{R},$$

alors la série $\sum\limits_{\nu=0}^{\infty} A_\nu \Phi_\nu(z)$ converge dans $\mathrm{I}(\mathrm{C}_R)$ et y représente une fonction holomorphe; cette série diverge dans $\mathrm{E}(\mathrm{C}_R)$.

Soit de plus $\varepsilon > 0$ arbitrairement petit, $\varrho \geqq r+\varepsilon$ et $z \in \mathrm{C}_\varrho$, on a

$$(2.7) \qquad \left| \frac{\Phi_n(z)}{\varphi^n(z)} - 1 \right| \leqq c_0 \left(\frac{r}{\varrho} \right)^n.$$

Il en résulte que

$$(2.8) \qquad \lambda |\varphi^n(z)| < |\Phi_n(z)| < \mu |\varphi^n(z)|$$

dans chaque sous-ensemble fermé de $\mathrm{E}(\mathrm{C}_r)$ où $\lambda > 0, \mu > 0$ sont des constantes indépendantes de z.

Désignons par $k(z) \neq z$ une application conforme et biunivoque de $\bar{\mathrm{I}}(\mathrm{C}_R)$ sur lui-même. C_R étant analytique, $k(z)$ est prolongeable au delà de C_R dans un domaine partiel de $\mathrm{E}(\mathrm{C}_R)$. Par conséquent il existe un nombre $R_0 > R$ tel que $k(z)$ est holomorphe dans $\mathrm{I}(\mathrm{C}_{\varrho'})$ pour tout $\varrho' < R_0$, en particulier, si $R \leqq \varrho' < R_0$.

$k(z)$ étant définie pour $z \in \mathrm{C}_R$, considérons les points $z_1 \in \mathrm{C}_R$, $z_2 \in \mathrm{C}_R$ liés par l'égalité $z_1 = k(z_2)$ et posons $w_1 = \varphi(z_1)$, $w_2 = \varphi(z_2)$ où $|w_1| = |w_2| = R$. Quand z_1 et z_2 parcourent C_R le quotient $w_1/w_2 = e^{i\vartheta}$ peut être ou bien une quantité variable ou bien une constante. Il est facile de montrer que ce dernier cas ne se présente que si C_R a un centre de symétrie et simultanément $k(z)$ est une rotation autour de ce centre. Dans ce qui suit nous écartons l'éventualité où $k(z)$ est une telle rotation.

§ 3. Deux lemmes

Dans la note [3] nous avons démontré deux lemmes (là Lemme 1 et Lemme 3) que nous énonçons ici sans preuve.

LEMME 1. — *Soient* $[c_{n\nu}]$ $(n=0, 1, 2, \ldots;\ \nu=0, 1, 2, \ldots)$ *une matrice infinie,* $\{x_\nu\}_{\nu=0}^{\infty}$ *et* $\{y_n\}_{n=0}^{\infty}$ *deux suites infinies liées par la relation*

$$(3.1) \qquad y_n = \sum_{\nu=0}^{\infty} c_{n\nu} x_\nu \qquad (n = 0, 1, 2, \ldots).$$

Pour que la série $\sum y_n$ *soit sommable* (C, δ) $(\delta \geqq 0)$ *chaque fois que la série* $\sum x_\nu$ *converge, il est nécessaire:*

(I) *que chaque colonne de la matrice* $[c_{n\nu}]$ *soit sommable* (C, δ), *c'est-à-dire que les quantités*

$$(3.2)\qquad (C,\delta)\text{-}\sum_{n=0}^{\infty} c_{n\nu} \qquad (\nu = 0, 1, 2, \dots)$$

soient finies;

(II) *qu'il existe une constante* $K>0$ *telle que*

$$(3.3)\qquad \sum_{n=0}^{\infty} |c_{n\nu} - c_{n,\nu+1}| < K(n^{\delta}+1) \qquad (n = 0, 1, 2, \dots).$$

LEMME 2. — *On suppose que la fonction* $h(t)$ *à variable et à valeur réelle jouit des propriétés suivantes:*

(i) $h(t)=h_0(t)+t,\quad h_0(t+2\pi)=h_0(t),\quad h_0(0)=0;$

(ii) $h_0 \in C^2;\quad h'(t)\neq 0$ *resp.* $h_0'(t)\neq -1;$

(iii) $h''(t)=h_0''(t)$ *a un nombre fini de zéros dans* $[0, 2\pi]$.

Soit de plus

$$(3.4)\qquad e^{i\nu h(t)} = \sum_{n=-\infty}^{\infty} a_{n\nu} e^{int} \qquad (\nu = 0, \pm 1, \pm 2, \dots).$$

Il existe alors une constante $\lambda>0$ *telle que*

$$(3.5)\qquad \lambda |n|^{1/2} < \sum_{\nu=-\infty}^{\infty} |a_{n\nu} - a_{n,\nu+1}| \qquad (n = 0, \pm 1, \pm 2, \dots).$$

§ 4. Résultat et preuve

THÉORÈME. — *Soient* C_R *une courbe fermée simple formée d'un seul arc analytique régulier,* $k(z)$ *une application conforme et biunivoque de* $\bar{\mathrm{I}}(\mathsf{C}_R)$ *sur lui-même différente d'une rotation,* $z_1 \in \mathsf{C}_R$ *et* $z_2 \in \mathsf{C}_R$ *deux points tels que* $z_1 = k(z_2)$, *enfin* $\delta \in [0, 1/2)$ *un paramètre donné. Alors il existe une fonction*

$$(4.1)\qquad F_1(z) = \sum_{\nu=0}^{\infty} A_\nu \Phi_\nu(z)$$

holomorphe dans $\mathrm{I}(\mathsf{C}_R)$, *dont la série de Faber converge en* z_1 *et, malgré cela, la série de Faber de la fonction*

$$(4.2)\qquad F_1[k(z)] = F_2(z) = \sum_{n=0}^{\infty} B_n \Phi_n(z),$$

holomorphe également dans $I(\mathsf{C}_R)$, *n'est pas sommable* (C, δ) *en* z_2, *bien que* $F_1(z_1) = F_2(z_2)$ *a une valeur déterminée.*

DÉMONSTRATION. — Comme $k(z)$ n'est pas une rotation, on peut admettre que $k(z)$ a un point double sur C_R, soit $k(z_0) = z_0 \in \mathsf{C}_R$. On suppose de plus que $z_1 = z_2 = z_0$. Ces hypothèses ne restreignent pas la généralité, mais permettent certaines simplifications des calculs. On a ainsi $F_1(z_0) = F_2(z_0)$ et il est à prouver que malgré la convergence de la série (4.1) en z_0, la série (4.2) n'est pas nécessairement sommable (C, δ) en z_0.

Nous commençons par exprimer les B_n à l'aide des A_ν. Nous avons, en tenant compte de (2.5), (4.1) et (4.2),

$$(4.3)\qquad \begin{aligned} B_n &= \frac{1}{2\pi i}\int_{K_\varrho} F_1\big(k[\psi(w)]\big)\frac{dw}{w^{n+1}} = \\ &= \sum_{\nu=0}^{\infty}\frac{A_\nu}{2\pi i}\int_{K_\varrho}\Phi_\nu\big(k[\psi(w)]\big)\frac{dw}{w^{n+1}}, \qquad r<\varrho<R,\ n\geqq 0.\end{aligned}$$

Écrivons $\Phi_\nu\big(k[\psi(w)]\big)=\Phi_\nu(k)$. Comme $|\varphi(z_0)|=R$, il découle de (2.8) que $\Phi_\nu(z_0)\neq 0$ et l'on a, d'après (4.3),

$$(4.4)\qquad B_n\Phi_n(z_0)=\sum_{\nu=0}^{\infty}A_\nu\Phi_\nu(z_0)\frac{\Phi_n(z_0)}{2\pi i}\int_{K_\varrho}\frac{\Phi_\nu(k)}{\Phi_\nu(z_0)}\frac{dw}{w^{n+1}}.$$

En posant donc

$$(4.5)\qquad c_{n\nu}=\frac{\Phi_n(z_0)}{2\pi i}\int_{K_\varrho}\frac{\Phi_\nu(k)}{\Phi_\nu(z_0)}\frac{dw}{w^{n+1}},\qquad x_\nu=A_\nu\Phi_\nu(z_0),\ y_n=B_n\Phi_n(z_0),$$

on peut examiner comment se réalisent les conditions du Lemme 1.

Nous allons établir que la relation (3.3) n'a pas lieu si $0\leqq\delta<1/2$. Il existe donc une série convergente $\sum x_\nu$ qui se transforme par la matrice $[c_{n\nu}]$ définie sous (4.5) en une série $\sum y_n$ qui n'est pas sommable (C,δ). Par conséquent on peut poser $A_\nu=x_\nu/\Phi_\nu(z_0)$ et, selon (2.7),

$$(4.6)\qquad \varlimsup_{\nu\to\infty}|A_\nu|^{1/\nu}=\frac{1}{|\varphi(z_0)|}\varlimsup_{\nu\to\infty}|x_\nu|^{1/\nu}\leqq\frac{1}{R}.$$

La fonction $F_1(z)=\sum A_\nu\Phi_\nu(z)$ est donc holomorphe dans $\mathrm{I}(\mathrm{C}_R)$ (voir (2.6) et la remarque y ajoutée) et cette série de Faber converge en z_0. Les coefficients B_n sont déterminés ensuite par (4.4), $F_2(z)=\sum B_n\Phi_n(z)$ est aussi analytique dans $\mathrm{I}(\mathrm{C}_R)$, mais la série $\sum B_n\Phi_n(z_0)$ n'est pas sommable (C,δ).

Considérons donc les différences $\Delta c_{n\nu}=c_{n\nu}-c_{n,\nu+1}$ et la somme $\sum_{\nu=0}^{\infty}|\Delta c_{n\nu}|$. Nous aurons, vu (4.5),

$$(4.7)\qquad \Delta c_{n\nu}=\frac{\Phi_n(z_0)}{2\pi i}\int_{K_\varrho}\left(\frac{\Phi_\nu(k)}{\Phi_\nu(z_0)}-\frac{\Phi_{\nu+1}(k)}{\Phi_{\nu+1}(z_0)}\right)\frac{dw}{w^{n+1}}=\frac{\Phi_n(z_0)}{2\pi i}\int_{K_\varrho}\frac{D_\nu(w)}{w^{n+1}}dw.$$

Nous allons voir qu'il existe deux matrices $[\gamma^*_{n\nu}]$ et $[\gamma_{n\nu}]$ ($n=0,1,2,\dots;\ \nu=0,1,2,\dots$) telles que d'une part

$$(4.8)\qquad \left|\sum_{\nu=0}^{\infty}(|\Delta c_{n\nu}|-|\gamma_{n\nu}|)\right|\leqq\sum_{\nu=0}^{\infty}|\Delta c_{n\nu}-\gamma^*_{n\nu}|+\sum_{\nu=0}^{\infty}|\gamma^*_{\nu n}-\gamma_{n\nu}|=o(1)\quad(n\to\infty),$$

d'autre part que

$$(4.9)\qquad \sum_{\nu=0}^{\infty}|\gamma_{n\nu}|>\lambda' n^{1/2}\qquad(n=0,1,2,\dots),$$

$\lambda'>0$ étant une constante indépendante de n.

1° Soit

$$\gamma^*_{n\nu} = \frac{\varphi^n(z_0)}{2\pi i}\int_{K_\varrho}\frac{D_\nu(w)}{w^{n+1}}\,dw \tag{4.10}$$

d'où, en vertu de (4.7),

$$\Delta c_{n\nu} - \gamma^*_{n\nu} = \frac{\Phi_n(z_0)-\varphi^n(z_0)}{2\pi i}\int_{K_\varrho}\frac{D_\nu(w)}{w^{n+1}}\,dw \tag{4.11}$$

où, en tenant compte de (2.7),

$$|\Phi_n(z_0)-\varphi^n(z_0)| \leqq c_0\left(\frac{r}{R}\right)^n|\varphi^n(z_0)| = c_0 r^n. \tag{4.12}$$

Pour évaluer $|D_\nu(w)|$, posons

$$R > \max_{|w|=\varrho}\left|\varphi\big(k[\psi(w)]\big)\right| = \varrho_0, \quad \varphi\big(k[\psi(w)]\big) = \varphi(k). \tag{4.13}$$

Ainsi il vient de (2.8)

$$|D_\nu(w)| \leqq \left|\frac{\Phi_\nu(k)}{\Phi_\nu(z_0)}\right| + \left|\frac{\Phi_{\nu+1}(k)}{\Phi_{\nu+1}(z_0)}\right| \leqq \frac{\mu}{\lambda}\left|\frac{\varphi(k)}{\varphi(z_0)}\right|^\nu\left(1+\left|\frac{\varphi(k)}{\varphi(z_0)}\right|\right) \leqq c_1\left(\frac{\varrho_0}{R}\right)^\nu. \tag{4.14}$$

On obtient finalement, à partir de (4.10)—(4.14),

$$\sum_{\nu=0}^{\infty}|\Delta c_{n\nu}-\gamma^*_{n\nu}| \leqq c_2\left(\frac{r}{\varrho}\right)^n = o(1) \quad (n\to\infty). \tag{4.15}$$

2° Soit maintenant

$$\gamma_{n\nu} = \frac{\varphi^n(z_0)}{2\pi i}\int_{K_\varrho}\left(\frac{\varphi^\nu(k)}{\varphi^\nu(z_0)} - \frac{\varphi^{\nu+1}(k)}{\varphi^{\nu+1}(z_0)}\right)\frac{dw}{w^{n+1}} = \frac{\varphi^n(z_0)}{2\pi i}\int_{K_\varrho}\frac{d_\nu(w)}{w^{n+1}}\,dw \tag{4.16}$$

et, par suite,

$$\gamma^*_{n\nu} - \gamma_{n\nu} = \frac{\varphi^n(z_0)}{2\pi i}\int_{K_\varrho}[D_\nu(w)-d_\nu(w)]\frac{dw}{w^{n+1}}. \tag{4.17}$$

Il faut donc majorer les quantités

$$|D_\nu(w)-d_\nu(w)| \leqq \left|\frac{\Phi_\nu(k)}{\Phi_\nu(z_0)} - \frac{\varphi^\nu(k)}{\varphi^\nu(z_0)}\right| + \left|\frac{\Phi_{\nu+1}(k)}{\Phi_{\nu+1}(z_0)} - \frac{\varphi^{\nu+1}(k)}{\varphi^{\nu+1}(z_0)}\right|. \tag{4.18}$$

Écrivons (2.7) sous la forme

$$\Phi_n(z) = \left[1 + c_0\left(\frac{r}{|\varphi(z)|}\right)^n t_n(z)\right]\varphi^n(z) \tag{4.19}$$

où $|t_n(z)| \leqq 1$ et souvenons nous du fait que dans l'expression (4.17) de $\gamma^*_{n\nu}-\gamma_{n\nu}$ on peut remplacer la courbe d'intégration K_ϱ par $K_{\varrho'}$ avec $R<\varrho'<R_0$, $k(z)$ étant encore analytique et univalente dans $I(C_{\varrho'})$ et φ resp. ψ jouit des mêmes propriétés dans $E(C_r)$ resp. $E(K_r)$. C'est ainsi qu'on sait introduire les quantités:

$$R < \min_{|w|=\varrho'}\left|\varphi\big(k[\psi(w)]\big)\right| = \varrho'_* < \max_{|w|=\varrho'}\left|\varphi\big(k[\psi(w)]\big)\right| = \varrho'_0 < R_0. \tag{4.20}$$

Nous aurons alors, en vertue de (2.8), (4.19) et (4.20),

$$(4.21)\quad \left|\frac{\Phi_\nu(k)}{\Phi_\nu(z_0)}-\frac{\varphi^\nu(k)}{\varphi^\nu(z_0)}\right| \leqq \frac{c_0}{\lambda}\left[\left(\frac{r}{\varrho'_*}\right)^\nu |t_\nu(k)|+\left(\frac{r}{R}\right)^\nu |t_\nu(z_0)|\right]\left|\frac{\varphi(k)}{\varphi(z_0)}\right|^\nu \leqq \frac{2c_0}{\lambda}\left(\frac{r\varrho'_0}{R^2}\right)^\nu.$$

Si $\varrho'-R$ est suffisamment petit, on a $R<\varrho'_0<R^2/r$ et $r\varrho'_0/R^2<1$. On conclut donc de (4.18)—(4.21)

$$(4.22)\qquad |D_\nu(w)-d_\nu(w)| \leqq c_0\left(\frac{r\varrho'_0}{R^2}\right)^\nu,$$

d'où l'on tire, en tenant compte de (4.17),

$$(4.23)\qquad \sum_{\nu=0}^{\infty} |\gamma^*_{n\nu}-\gamma_{n\nu}| \leqq c_4\left(\frac{R}{\varrho'}\right)^n = o(1) \qquad (n\to\infty).$$

(4.15) et (4.23) implique (4.8).

3° Il reste à vérifier (4.9). Dans l'expression (4.16) de $\gamma_{n\nu}$ on peut remplacer K_ϱ par K_R et poser pour $|w|=R$

$$(4.24)\quad \varphi(z)=w=Re^{it},\quad \varphi\big(k[\psi(w)]\big)=Re^{ih(t)},\quad \varphi(z_0)=Re^{ih(t_0)}=w_0=e^{it_0}.$$

On voit que $h(t_0)=t_0$ et, pour simplifier le calcul, on admet que $t_0=0$. k, φ, ψ étant des fonctions analytiques et univalentes dans les domaines considérés, $h(t)$ se comporte de la même façon dans un voisinage de l'axe des t. Il s'ensuit que $h(t)\in C^\infty$, $h'(t)>0$ et $h''(t)$ n'a qu'un nombre fini de zéros dans l'intervalle $[0, 2\pi]$. $e^{ih(t)}$ est en outre 2π-périodique, d'où il vient $h(t+2\pi)=h(t)+2\pi$ ou bien $h(t)=$ $=h_0(t)+t$ avec $h_0(t+2\pi)=h_0(t)$ et $h(0)=h_0(0)=0$, car $t_0=0$, par hypothèse. $h(t)$ satisfait donc à toutes les conditions du Lemme 2. On aura ainsi, en raison de (3.4), (4.16) et (4.24),

$$(4.25)\qquad \gamma_{n\nu}=\frac{1}{2\pi i}\int_0^{2\pi}[e^{i\nu h(t)}-e^{i(\nu+1)h(t)}]e^{-int}\,dt = a_{n\nu}-a_{n,\nu+1}.$$

Les $\gamma_{n\nu}$ définis sous (4.25) ont naturellement un sense même pour des $n<0$, $\nu<0$. Nous pouvons donc écrire, grâce à (3.5),

$$(4.26)\qquad \sum_{\nu=-\infty}^{\infty} |\gamma_{n\nu}| > \lambda|n|^{1/2} \qquad (n=0, \pm 1, \pm 2, \ldots).$$

Pour avoir (4.9), on doit montrer encore que

$$\sum_{\nu=-\infty}^{-1} |\gamma_{n\nu}| = O(1) \qquad (n\to+\infty).$$

Remplaçons dans (4.16) K_ϱ par $\mathsf{K}_{\varrho'}$ $(R<\varrho'<R_0)$, nous aurons, vu (4.20), pour $\nu<0$,

$$|\gamma_{n\nu}| \leqq c_5\left(\frac{R}{\varrho'}\right)^n\left(\frac{R}{\varrho'_*}\right)^{-\nu}$$

et de là

$$\sum_{\nu=-\infty}^{-1} |\gamma_{n\nu}| \leqq c_6 \left(\frac{R}{\varrho'}\right)^n = o(1) \quad (n \to +\infty). \tag{4.27}$$

(4.26) et (4.27) vérifient déjà (4.9).

(4.8) prouve donc que la matrice $[c_{n\nu}]$ définie sous (4.5) ne remplit pas la condition (3.3) du Lemme 2, si $0 \leqq \delta < 1/2$. La démonstration du Théorème est ainsi achevée.

RÉFÉRENCES

[1] Alpár, L., Remarques sur la sommabilité des séries de Taylor sur leur cercle de convergence III, *Magyar Tud. Akad. Mat. Kut. Int. Közl.* **5** (1960), 97—152.

[2] Alpár, L., Convergence et représentation conforme, *Magyar Tud. Akad. Mat. Kut. Int. Közl.* **9** (1964), 503—514.

[3] Alpár, L., Sur certains changements de variable des séries de puissances, *Studies in Pure Mathematics, To the Memory of Paul Turán* (à paraître).

[4] Deák, J., Az általánosított Faber-sorok néhány tulajdonsága (Quelques propriétés des séries de Faber généralisées, en hongrois), *Mat. Lapok* **23** (1972), 147—160.

[5] Faber, G., Über polynomische Entwickelungen, *Math. Ann.* **57** (1903), 389—408.

[6] Faber, G., Über polynomische Entwickelungen II., *Math. Ann.* **64** (1907), 116—135.

[7] Montel, P., *Leçons sur les séries de polynômes à une variable complexe,* Gauthier-Villars, Paris, 1910.

[8] Smirnov, V. L.—Lebedev, N. A., *Functions of a Complex Variable,* Iliffe Books Ltd., London, 1968.

[9] Tietz, H., Faber series and the Laurent decomposition, *Mich. Math. J.* **4** (1957), 175—179.

[10] Turán, P., A remark concerning the behaviour of a power series on the periphery of its convergence-circle, *Publ. Inst. Math. Acad. Serbe Sci.* **12** (1958), 19—26.

[11] Ullman, J. L., On Faber series, 1. A problem of transfer, *Mich. Math. J.* **2** (1953—54), 109—114.

Institut Mathématique de l'Académie Hongroise des Sciences
H—1053 Budapest, Reáltanoda u. 13—15

(Reçu le 13 août 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 181—185.

SEMISIMPLE CLASSES AND H-RELATIONS

by

RICHARD WIEGANDT

In a recent paper [9] ROSSA and TANGEMAN introduced the notion of H-relations and investigated such relations in the context of radical classes. The purpose of this note is to deal with H-relations and semisimple classes. In the proofs we shall apply recent results of [1], [2], [3], [8] and [11]. Radical and semisimple classes are meant in the sense of KUROSH and AMITSUR, for details we refer to [13].

Let $\mathbf{A}$ be a universal class (that is a *subring hereditary* and homomorphically closed class) of not necessarily associative rings. Following Rossa and Tangeman [9], a relation σ on $\mathbf{A}$ will be called an *H-relation*, if σ satisfies the properties:

(1) $L\sigma A$ implies L is a subring in A.

(2) If $L\sigma A$ and φ is a homomorphism of A, then $\varphi(L)\sigma\varphi(A)$.

(3) If $L\sigma A$ and $J\lhd A$, then $\big((L\cap J)\sigma J\big)$.

A large class of examples for H-relations was provided in [9]. In particular, $I\lhd A$ (I is an ideal of A), $I\blacktriangleleft A$ (I is a left ideal of A) and $I<A$ (I is a subring of A) are H-relations. We mention also a further example. A subring S of a ring A is called a *bi-ideal*, if $SAS\subseteq S$ holds (cf. [4]). Being a bi-ideal, is an H-relation. All these examples for H-relations satisfy also the following additional condition

(4) If $I\lhd A$, then $I\sigma A$.

In what follows σ will always denote an H-relation (not necessarily satisfying condition (4)).

PROPOSITION 1. *If $L\sigma A$ and $K\lhd A$, then $\big((L+K)/K\big)\sigma(A/K)$.*

PROOF. Put $K=\operatorname{Ker}\varphi$. Then condition (2) yields the assertion.

As usual, we shall use the *upper radical operator* $\mathscr{U}$ and the *semisimple operator* $\mathscr{S}$ acting on a subclass $\mathbf{X}$ of $\mathbf{A}$ and defined by

$\mathscr{U}\mathbf{X}=\{A\in\mathbf{A}\colon A$ has no non-zero homomorphic image in $\mathbf{X}\}$,

$\mathscr{S}\mathbf{X}=\{A\in\mathbf{A}\colon A$ has no non-zero ideal in $\mathbf{X}\}$.

A subclass $\mathbf{M}$ of $\mathbf{A}$ is called *σ-regular*, if $\mathbf{M}$ satisfies the following condition: If $L\sigma A\in\mathbf{M}$ and $L\neq 0$, then L can be mapped homomorphically onto a non-zero ring in $\mathbf{M}$. A subclass $\mathbf{P}$ of $\mathbf{A}$ is called *σ-strong*, if $L\sigma A$ and $L\in\mathbf{P}$ imply $L\in\mathbf{P}(A)$ where $\mathbf{P}(A)$ is defined as

$$\mathbf{P}(A)=\sum(J\lhd A\colon J\in\mathbf{P}).$$

PROPOSITION 2. *Let $\mathbf{S}$ be a semisimple class. If $\mathbf{S}$ is σ-regular, then the radical class $\mathbf{R}=\mathscr{U}\mathbf{S}$ is σ-strong.*

PROOF. Suppose that $L\sigma A$, $L\in\mathbf{R}$ and $L\nsubseteq\mathbf{R}(A)$. Now by Proposition 1 we get

$$0\neq L/(\mathbf{R}(A)\cap L)\cong(\mathbf{R}(A)+L)/\mathbf{R}(A)\sigma\big(A/\mathbf{R}(A)\big)\in\mathbf{S}.$$

Since $\mathbf{R}$ is homomorphically closed, it follows

$$0 \neq (\mathbf{R}(A)+L)/\mathbf{R}(A) \cong L/(\mathbf{R}(A)\cap L)\in\mathbf{R},$$

contradicting the assumption that $\mathbf{S}$ is σ-regular.

PROPOSITION 3. *If* $\mathbf{R}$ *is the upper radical* $\mathbf{R}=\mathcal{U}\mathbf{M}$ *of a* $\lhd$*-regular class* $\mathbf{M}$ *and* $\mathbf{R}$ *is* σ*-strong, then* $\mathbf{M}$ *is* σ*-regular.*

PROOF. Suppose that $L\sigma A\in\mathbf{M}$ and L has no non-zero homomorphic image in $\mathbf{M}$. Then $L\in\mathcal{U}\mathbf{M}=\mathbf{R}$. Since $\mathbf{M}\subseteqq\mathcal{S}\mathcal{U}\mathbf{M}=\mathcal{S}\mathbf{R}$, we have $\mathbf{R}(A)=0$. Since $\mathbf{R}$ is σ-strong, we get $L\subseteqq\mathbf{R}(A)=0$ implying $L=0$.

Propositions 2 and 3 yield

COROLLARY 4. *A radical class* $\mathbf{R}$ *is* σ*-strong if and only if its semisimple class* $\mathcal{S}\mathbf{R}$ *is* σ*-regular.*

A subclass $\mathbf{X}$ of $\mathbf{A}$ is called *σ-hereditary*, if $L\sigma A\in\mathbf{X}$ implies $L\in\mathbf{X}$.

PROPOSITION 5. *Assume that the H-relation* σ *satisfies the following condition:*
(*) *If* $L\lhd J$ *and* $J\sigma A$, *then also* $L\sigma A$.
A radical class $\mathbf{R}$ *is* σ*-strong if and only if the semisimple class* $\mathcal{S}\mathbf{R}$ *is* σ*-hereditary.*

PROOF. Since a σ-hereditary class is always σ-regular, Proposition 2 yields the sufficiency.

Suppose that $\mathbf{R}$ is σ-strong and that $L\sigma A\in\mathcal{S}\mathbf{R}$. If $L\notin\mathcal{S}\mathbf{R}$, then $0\neq\mathbf{R}(L)\lhd L\sigma A$ and by the assumption upon σ, we get $\mathbf{R}(L)\sigma A$. Since $\mathbf{R}$ is σ-strong, it follows $0\neq\mathbf{R}(L)\subseteqq\mathbf{R}(A)=0$, a contradiction.

Condition (*) is rather restrictive, and it will not be used throughout the paper. As in [9], we shall use condition

(σH) If $0\neq L\sigma A$ and L has a non-zero homomorphic image in $\mathbf{M}$, then also A has a non-zero homomorphic image in $\mathbf{M}$.

In the next assertion we paraphrase Theorem 5 of [9].

PROPOSITION 6. *For a* $\lhd$*-regular class* $\mathbf{M}$ *the upper radical* $\mathcal{U}\mathbf{M}$ *is* σ*-hereditary if and only if* $\mathbf{M}$ *satisfies condition* (σH).

For a subclass $\mathbf{X}$ of $\mathbf{A}$, let $(A)\mathbf{X}$ denote the ideal

$$(A)\mathbf{X}=\bigcap_{\alpha}(J_\alpha\lhd A\colon A/J\in\mathbf{X}).$$

COROLLARY 7. *A subclass* $\mathbf{S}$ *is a semisimple class of a* σ*-hereditary radical class if and only if* $\mathbf{S}$ *is a* $\lhd$*-regular class,* $\mathbf{S}$ *is closed under subdirect sums and ideal extensions,* $\mathbf{S}$ *satisfies condition* (σH) *and* $((A)\mathbf{S})\mathbf{S}\lhd A$ *for every ring* $A\in\mathbf{A}$.

The assertion is an immediate consequence of Proposition 6 and [8] Theorem 9.

COROLLARY 8. *A subclass* $\mathbf{S}$ *is a semisimple class of a* σ*-hereditary radical class* $\mathbf{R}$ *if and only if* $\mathbf{S}$ *is closed under subdirect sums, satisfies condition* (σH) *and* $\mathbf{R}(A)=(A)\mathbf{S}$ *holds for every ring* $A\in\mathbf{A}$.

The proof is a straightforward application of Proposition 6 and [8] Theorem 10.

Corollary 4 and Proposition 6 yield the following assertion which can be considered as a generalization of [12] Theorem 2 and [6] Theorem 1.

COROLLARY 9. *Let* **M** *be a* $\triangleleft$*-regular class in* **A**. *The class* $\mathcal{U}$**M** *is a* σ*-hereditary and* ϱ*-strong radical if and only if* **M** *is* ϱ*-regular and satisfies condition* (σH). *In particular, a radical class* **R** *is* σ*-hereditary and* ϱ*-strong if and only if the semisimple class* $\mathcal{S}$**R** *is* ϱ*-regular and satisfies* (σH).

Next, let us suppose that **A** *is the class of all not necessarily associative rings.*

PROPOSITION 10. *If* **R** *is a radical class such that* $\mathcal{S}$**R** *is* $\triangleleft$*-hereditary, then* $\mathcal{S}$**R** *is* σ*-hereditary for any H-relation* σ.

PROOF. By GARDNER's [3] Corollary 2,5 **R** is $<$-strong. Hence in view of Proposition 3 $\mathcal{S}$**R** is $<$-hereditary and so by (1) $\mathcal{S}$**R** is also σ-hereditary.

PROPOSITION 11. *Let* σ *be an H-relation satisfying condition* (4). *If* **S** *is a* σ*-hereditary semisimple class, then* **S** *is* ϱ*-hereditary for every H-relation* ϱ.

PROOF. In view of (4), S is $\triangleleft$-hereditary. Hence Proposition 10 yields the statement.

COROLLARY 12. *Let* σ *be an H-relation satisfying condition* (4). *Then the following four conditions are equivalent for a semisimple class* **S**:

(i) **S** *is* $\triangleleft$*-hereditary;*
(ii) **S** *is* σ*-hereditary;*
(iii) **S** *is* $<$*-hereditary;*
(iv) **S** *is* ϱ*-hereditary for every H-relation* ϱ.

A class **X** of rings is said to be *weakly homomorphically closed,* if $I \triangleleft A \in \mathbf{X}$ and $I^2=0$ imply $A/I \in \mathbf{X}$.

THEOREM 13. *Let* σ *be an H-relation satisfying condition* (4) *and let* **S** *be a semisimple class such that* **S** *is* σ*-hereditary and weakly homomorphically closed. Then either* **S**=**A** *or* **S** *consists of one-element rings.*

PROOF. In view of Corollary 12 S is $\triangleleft$-hereditary, and so by Gardner's [3] Corollary 2,5 $\mathcal{U}$**S** is an A-radical. Note that by definition an A-radical containing all zero-rings, must be **A**, and a subidempotent A-radical coincides with the class of one-element rings. On the other hand, by [1] Corollary 2 either $\mathcal{U}$**S** contains all zero-rings, or $\mathcal{U}$**S** is subidempotent. Hence either **S** consists of one element rings, or **S**=**A**.

From here onwards we assume that the *universal class* **A** *consists of associative or alternative rings.*

THEOREM 14. *Let* σ *be an H-relation satisfying condition* (4). *The following three con'itions are equivalent:*

(i) **S** *is the semisimple class of a* σ*-hereditary radical class* **R**=$\mathcal{U}$**S**;
(ii) **S** *is* $\triangleleft$*-regular, closed under subdirect sums and ideal extensions and* **S** *satisfies condition* (σH);
(iii) **S** *is* $\triangleleft$*-regular, closed under subdirect sums and ideal extensions, and* **R**=$\mathcal{U}$**S** *is* σ*-hereditary.*

PROOF. By [11] and [2] $\mathbf{S}$ is a semisimple class of associative or alternative rings if and only if $\mathbf{S}$ is $\lhd$-regular, and closed under subdirect sums and ideal extensions. Hence an application of Proposition 6 yields the assertions.

An ideal L of a ring A is said to be *large in* A if $L\cap I\neq 0$ for every non-zero ideal I of A. We shall need the following condition:

(λ) If L is a large ideal in A and $L\in\mathbf{S}$, then $A\in\mathbf{S}$.

The next theorem is a generalization of VAN LEEUWEN's Theorem (cf. [5] and [7]).

THEOREM 15. *Let σ be an H-relation satisfying condition* (4). *A class $\mathbf{S}$ is the semisimple class of a σ-strong and $\lhd$-hereditary radical if and only if $\mathbf{S}$ is σ-regular, closed under subdirect sums, and satisfies condition* (λ).

In view of Corollary 4 and Theorem 14 the proof is straightforward; it is to be taken into consideration that a $\lhd$-regular class which is closed under subdirect sums and satisfies condition (λ), is a semisimple class of a $\lhd$-hereditary radical (cf. [2] Corollary 2).

Confining ourselves to the *variety of all associative or alternative* rings [2] Corollary 3 asserts that a proper subclass is the semisimple class of a supernilpotent radical if and only if it is $\lhd$-regular, subdirectly closed, weakly homomorphically closed and satisfies condition (λ). Hence for an H-relation σ satisfying (4), Theorem 15 implies immediately the following

COROLLARY 16. *A class $\mathbf{R}$ is a supernilpotent σ-strong radical if and only if $\mathbf{S}=\mathscr{S}\mathbf{R}$ is a proper subclass which is σ-regular, weakly homomorphically closed, closed under subdirect sums, and satisfies condition* (λ).

SANDS [10] called a radical class an *N-radical* if it is $\blacktriangleleft$-hereditary, $\blacktriangleleft$-strong and it contains all zero-rings. Theorem 14 and Corollary 16 yield the following characterization of semisimple classes of N-radicals.

COROLLARY 17. *A proper subclass $\mathbf{S}$ is the semisimple class of an N-radical if and only if $\mathbf{S}$ is $\blacktriangleleft$-regular, weakly homomorphically closed, closed under subdirect sums and ideal extensions, and satisfies condition* ($\blacktriangleleft$H).

REFERENCES

[1] ANDERSON, T.—WIEGANDT, R., Weakly homomorphically closed semisimple classes, *Acta Math. Acad. Sci. Hungar.* **14** (1979), 329—336.

[2] ANDERSON, T.—WIEGANDT, R., Semisimple classes of alternative rings, *Proc. Edinburgh Math. Soc.* (to appear).

[3] GARDNER, B. J., Some degeneracy and pathology in non-associative radical theory, *Annales Univ. Sci. Budapest. Sect. Math.* **22—23** (1979—80), 65—74.

[4] LAJOS, S.—SZÁSZ, F., Bi-ideals in associative rings, *Acta Sci. Math. (Szeged)* **32** (1971), 185—193.

[5] LEEUWEN, L. C. A. VAN, Properties of semisimple classes, *J. Nat. Sci. and Math. Lahore* **15** (1975), 59—67.

[6] LEEUWEN, L. C. A. VAN—JENKINS, T. L., *N*-radicals and simple rings, *J. Nat. Sci. and Math. Lahore* **17** (1977), 95—103.

[7] LEEUWEN, L. C. A. VAN—ROOS, C.—WIEGANDT, R., Characterizations of semisimple classes, *J. Austral. Math. Soc. Ser. A* **23** (1977), 172—182.

[8] LEEUWEN, L. C. A. VAN—WIEGANDT, R., Radicals, semisimple classes and torsion theories, *Acta Math. Acad. Sci. Hungar.* **36** (1980), 37—47.

[9] ROSSA, R. F.—TANGEMAN, R. L., General heredity for radical theory, *Proc. Edinburgh Math. Soc.* **20** (1976—77), 333—337.

[10] SANDS, A. D., Radicals and Morita contexts, *J. Algebra* **24** (1973), 335—345.

[11] SANDS, A. D., A characterization of semisimple classes, *Proc. Edinburgh Math. Soc.* (to appear).

[12] WIEGANDT, R., On N-radicals, *J. Nat. Sci. and Math. Lahore* **13** (1973), 255—262.

[13] WIEGANDT, R., *Radical and semisimple classes of rings,* Queen's papers in pure and appl. math., no. 37, Kingston, Ontario, 1974.

Mathematical Institute of the Hungarian Academy of Sciences
H—1053 Budapest, Reáltanoda u. 13—15

(Received August 13, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 187—192.

ON A TRANSFORMATION OF GRONWALL CONCERNING THE THREE-BODY PROBLEM OF QUANTUM MECHANICS

by

E. MAKAI

1. The Schrödinger equation of the three-body problem is

$$(1.1)\qquad -\sum_{i=1}^{3}\mu_i\Delta_i\psi+U\psi=E\psi$$

where $\Delta_i=\partial^2/\partial x_i^2+\partial^2/\partial y_i^2+\partial^2/\partial z_i^2$, μ_i and E are constants. We suppose that

$$(1.2)\qquad \mu_i\geqq 0,\quad M=\mu_1\mu_2+\mu_2\mu_3+\mu_3\mu_1>0$$

and U is a function of the quantities r_1, r_2, r_3 only, which are defined by

$$r_1=[(x_2-x_3)^2+(y_2-y_3)^2+(z_2-z_3)^2]^{1/2}\quad \text{(cycl.)}$$[1].

It is known that (1.1) has then solutions ψ depending only on r_1, r_2, r_3 and satisfying the equation

$$(1.3)\qquad -\sum_{i=1}^{3}\mu_i\Delta^i\psi+U\psi=E\psi,$$

where

$$\Delta^1=\frac{\partial^2}{\partial r_2^2}+\frac{-r_1^2+r_2^2+r_3^2}{r_2r_3}\frac{\partial^2}{\partial r_2\partial r_3}+\frac{\partial^2}{\partial r_3^2}+\frac{2}{r_2}\frac{\partial}{\partial r_2}+\frac{2}{r_3}\frac{\partial}{\partial r_3}\quad \text{(cycl.)},$$

cf. [2]. Introducing the bilinear forms

$$B_1(u,v)=\frac{\partial u}{\partial r_2}\frac{\partial v}{\partial r_2}+\frac{-r_1^2+r_2^2+r_3^2}{2r_2r_3}\left(\frac{\partial u}{\partial r_2}\frac{\partial v}{\partial r_3}+\frac{\partial u}{\partial r_3}\frac{\partial v}{\partial r_2}\right)+\frac{\partial u}{\partial r_3}\frac{\partial v}{\partial r_3}\quad \text{(cycl.)}$$

in the first derivatives of u and v and the domain D_r defined by

$$(1.4)\qquad D_r\colon\ r_1+r_2-r_3>0,\ r_2+r_3-r_1>0,\ r_3+r_1-r_2>0$$

equation (1.3) is the Euler equation of the variational problem

$$\delta Q(\psi)=0,$$

where

$$(1.5)\quad Q(\psi)=\iiint_{D_r}\left[\sum_{i=1}^{3}\mu_iB_i(\psi,\psi)+U\psi^2\right]r_1r_2r_3\,dr_1\,dr_2dr_3\Big/\iiint_{D_r}\psi^2r_1r_2r_3\,dr_1dr_2dr_3,$$

a circumstance noted by E. A. HYLLERAAS [7].

[1] The notation "cycl." will be used whenever a formula holds after cyclic interchanges of the indices 1, 2, 3, too.

In the particular case $\mu_1=\mu_2>0, \mu_3=0$ T. H. GRONWALL [5], [6] has shown that introducing a certain set of new variables ξ_1, ξ_2, ξ_3 instead of r_1, r_2, r_3 the variational quotient $Q(\psi)$ can be transformed into

$$(1.6)\quad Q(\psi)=\iiint\limits_{D_\xi}\left[4\varrho\sum_{i=1}^{3}\left(\frac{\partial\psi}{\partial\xi_i}\right)^2+V\psi^2\right]\frac{\xi_3}{\varrho}\,d\xi_1\,d\xi_2\,d\xi_3\Big/\iiint\limits_{D_\xi}\psi^2\frac{\xi_3}{\varrho}\,d\xi_1\,d\xi_2\,d\xi_3$$

where

$$(1.7)\qquad \varrho=(\xi_1^2+\xi_2^2+\xi_3^2)^{1/2},\quad V(\xi_1,\xi_2,\xi_3)=U(r_1,r_2,r_3)$$

and the domain of integration is the extremely simple domain

$$(1.8)\qquad D_\xi\colon\ \xi_3>0.$$

He remarked further that the Euler equation of (1.6) is

$$(1.9)\qquad -4\varrho\left[\sum_{i=1}^{3}\frac{\partial^2\psi}{\partial\xi_i^2}+\frac{1}{\xi_3}\frac{\partial\psi}{\partial\xi_3}\right]+V\psi=E\psi,$$

a simple form of the Schrödinger equation (1.3).

Since the publishing of Gronwall's papers several facts became known about eigenfunctions of this variational problem. Of these we first mention the result of T. KATO [9], that if U has only certain "mild" singularities, e.g., $U=\sum e_i r_i^{-1}$, e_i $=$const., then eigenfunctions are continuous on the closure $\bar{D}_\xi$ of D_ξ, that is on $\xi_3\geqq 0$. Further results originate from M. S. BAOUENDI and C. GOULAOUIC [1] and from V. H. FROIM [4], who independently investigated a family of partial differential equations of which (1.9) is an element. Their results applied to (1.9) state that if ξ_{10}, ξ_{20} are fixed quantities, $[V(\xi_1,\xi_2,\xi_3)-E]/\varrho$ and $f(\xi_1,\xi_2)$ analytic on the domain D^0: $|\xi_1-\xi_{10}|<R_1+\varepsilon$, $|\xi_2-\xi_{20}|<R_2+\varepsilon$, $|\xi_3|<\varepsilon$ $(\varepsilon>0)$ then there exists a positive ε_1 and a *unique* solution of (1.9) analytic on D^1: $|\xi_1-\xi_{10}|<R_1$, $|\xi_2-\xi_{20}|<R_2$, $|\xi_3|<\varepsilon_1$, not necessarily an eigenfunction, for which the initial condition $\psi(\xi_1,\xi_2,0)=$ $=f(\xi_1,\xi_2)$ holds. Froim showed further that if $[V(\xi_1,\xi_2,\xi_3)-E]/\varrho$ is analytic on D^0, then any solution of (1.9) analytic on the domain D^2: $|\xi_1-\xi_{10}|<R_1$, $|\xi_2-\xi_{20}|<R_2, 0<|\xi_3|<\varepsilon_1$ (ε_1 small enough) is of the form $\varphi(\xi_1,\xi_2,\xi_3)+$ $+\chi(\xi_1,\xi_2,\xi_3)\log\xi_3$, where φ and χ are analytic on D^1, thus clearing up the nature of the singularities of solutions to (1.9) on the boundary of D_ξ.

The purpose of this paper is to show that Gronwall's transformation can be generalized to the case when the μ_i's are subject only to the conditions (1.2).

If we introduce the notations

$$(1.10)\qquad \varrho_i=r_i^2,\quad \sigma_1=-\varrho_1+\varrho_2+\varrho_3\quad \text{(cycl.)}$$

and define the real variables ξ_1, ξ_2, ξ_3 *by*

$$(1.11_{1,2})\quad \xi_1+i\xi_2=\frac{(\mu_2+iM^{1/2})^2\sigma_1+(\mu_1-iM^{1/2})^2\sigma_2+(\mu_1+\mu_2)^2\sigma_3}{2(\mu_1+\mu_2)M}\qquad (i=\sqrt{-1})$$

$$(1.11_3)\qquad \xi_3=\left(\frac{\sigma_1\sigma_2+\sigma_2\sigma_3+\sigma_3\sigma_1}{M}\right)^{1/2}\quad (\geqq 0)$$

then the variational quotient (1.5) *is transformed into* (1.6) *where* ϱ, V *and* D_ξ *are defined by* (1.7) *and* (1.8), *respectively.*

The corresponding Euler equation is, of course, again (1.9). We remark that by a general theorem [2, vol. I, Ch. IV, § 8] equation (1.3) can be directly transformed into (1.9) by introducing the new variables ξ_i.

If r_1, r_2, r_3 satisfy inequalities (1.4), then ξ_3 is real since [2]

$$w = M\xi_3^2 = (r_1+r_2+r_3)(-r_1+r_2+r_3)(r_1-r_2+r_3)(r_1+r_2-r_3) \tag{1.12}$$

and one recognizes the connection of the right-hand side with the area of a triangle of sides r_1, r_2, r_3. Note that if the point (r_1, r_2, r_3) is on the boundary of the domain D_r (corresponding to a degenerate triangle of sides r_1, r_2, r_3) then $\xi_3=0$.

The quantity ξ_3 depends symmetrically on the σ_i's and μ_i's. Though this is not true for ξ_1 and ξ_2, yet it holds for $\varrho'=(\xi_1^2+\xi_2^2)^{1/2}$. One has namely by $M+\mu_1^2= =(\mu_1+\mu_2)(\mu_1+\mu_3)$ (cycl.) that

$$\begin{aligned}4M^2(\xi_1^2+\xi_2^2) &= (\mu_2+\mu_3)^2\sigma_1^2+(\mu_3+\mu_1)^2\sigma_2^2+(\mu_1+\mu_2)^2\sigma_3^2+\\ &+2(\mu_1^2-M)\sigma_2\sigma_3+2(\mu_2^2-M)\sigma_3\sigma_1+2(\mu_3^2-M)\sigma_1\sigma_2.\end{aligned}$$

An easy and useful consequence of this is

$$\begin{aligned}\varrho = (\xi_1^2+\xi_2^2+\xi_3^2)^{1/2} &= \frac{1}{2M}[(\mu_2+\mu_3)\sigma_1+(\mu_3+\mu_1)\sigma_2+(\mu_1+\mu_2)\sigma_3] =\\ &= \frac{1}{M}\sum \mu_i\varrho_i,\end{aligned} \tag{1.13}$$

from which we have that each of the quantities ϱ, ξ_1, ξ_2 depends linearly on $\sigma_1, \sigma_2, \sigma_3$ (or on $\varrho_1, \varrho_2, \varrho_3$).

Our statement will be proved if we show (i)

$$\sum \mu_i B_i(\psi, \psi) = 4\varrho \sum \left(\frac{\partial\psi}{\partial\xi_i}\right)^2; \tag{1.14}$$

(ii) the value of the Jacobi determinant $\partial(r_1, r_2, r_3)/\partial(\xi_1, \xi_2, \xi_3)$ is $c\xi_3(\varrho r_1 r_2 r_3)^{-1}$ where c is a constant differing from 0;

(iii) the domain D_r is transformed by (1.11) onto the domain D_ξ.

2. Introducing the notations

$$B_0(u, v) = B(u, v) = \sum_{i=1}^{3} \mu_i B_i(u, v) = \sum_{i,j=1}^{3} b_{ij}\frac{\partial u}{\partial r_i}\frac{\partial v}{\partial r_j} \tag{2.1}$$

with

$$b_{11} = \mu_2+\mu_3, \quad b_{12} = b_{21} = \mu_3\frac{\sigma_3}{2r_1r_2} \quad \text{(cycl.)} \tag{2.2}$$

[2] Here and in the following algebraic identities will not be proved, if their verification involves only elementary operations.

the left-hand side of (1.14) is $B(\psi, \psi)$. We recall the elementary relations

(2.3) $B_i(u, v_1+v_2) = B_i(u, v_1)+B_i(u, v_2), \quad B_i(u, v_1 v_2) = v_1 B_i(u, v_2)+v_2 B_i(u, v_1)$

for $i=0, 1, 2, 3$. We have

$$B(\psi, \psi) = \sum_{ij} b_{ij} \frac{\partial\psi}{\partial r_i} \frac{\partial\psi}{\partial r_j} = \sum_{ij} b_{ij} \sum_{kl} \frac{\partial\psi}{\partial\xi_k} \frac{\partial\xi_k}{\partial r_i} \frac{\partial\psi}{\partial\xi_l} \frac{\partial\xi_l}{\partial r_j} = \sum_{kl} B(\xi_k, \xi_l) \frac{\partial\psi}{\partial\xi_k} \frac{\partial\psi}{\partial\xi_l},$$

thus (1.14) is equivalent to the statements

(2.4) $$B(\xi_k, \xi_l) = 4\varrho\delta_{kl} \qquad (k, l = 1, 2, 3)$$

where δ_{kl} is Kronecker's delta.

A simple calculation shows that the quantities σ_i satisfy the equalities

$$B_1(\sigma_1, \sigma_1) = 4(2\sigma_1+\varrho_1), \quad B_1(\sigma_2, \sigma_2) = B_1(\sigma_3, \sigma_3) = 4\varrho_1 \quad \text{(cycl.)}$$

$$B_1(\sigma_1, \sigma_2) = -B_1(\sigma_1, \sigma_3) = 4(\varrho_3-\varrho_2), \quad B_1(\sigma_2, \sigma_3) = -4\varrho_1 \quad \text{(cycl.)}$$

thus by (2.1), (1.13) and (2.3) we get

(2.5) $$B(\sigma_1, \sigma_1) = 4(2\mu_1\sigma_1+M\varrho), \quad B(\sigma_1, \sigma_2) = 4(\mu_1\sigma_2+\mu_2\sigma_1-M\varrho) \quad \text{(cycl.)}.$$

Further one has by (2.3) for any u and for $w=\sigma_1\sigma_2+\sigma_2\sigma_3+\sigma_3\sigma_1$

$$B(u, w) = (\sigma_2+\sigma_3)B(u, \sigma_1)+(\sigma_3+\sigma_1)B(u, \sigma_2)+(\sigma_1+\sigma_2)B(u, \sigma_3).$$

Substituting $u=\sigma_i$ we get from (2.5) and (1.13)

(2.6) $$B(\sigma_i, w) = B(w, \sigma_i) = 8\mu_i w \qquad (i = 1, 2, 3)$$

and if $u=w$ then by (2.6) and (1.13)

$$B(w, w) = 16M\varrho w.$$

On the other hand by (1.12) and (2.3)

$$B(w, w) = M^2 B(\xi_3^2, \xi_3^2) = 4M^2\xi_3^2 B(\xi_3, \xi_3) = 4MwB(\xi_3, \xi_3)$$

and the last two formulas yield $B(\xi_3, \xi_3)=4\varrho$, i.e., (2.4) in the case $k=l=3$.

Further, introducing the quantities

$$q_1 = \mu_3\sigma_2-\mu_2\sigma_3 \quad \text{(cycl.)}$$

(for which $\sum \mu_i q_i=0$ holds) (2.6) yields

(2.7) $$B(w, q_j) = 0 \qquad (j = 1, 2, 3).$$

On the other hand ξ_1 and ξ_2 depend linearly on the quantities q_i:

(2.8) $$\xi_1 = \frac{(\mu_1+\mu_2)(q_2-q_1)+(\mu_2-\mu_1)q_3}{2(\mu_1+\mu_2)M}, \quad \xi_2 = \frac{q_3}{(\mu_1+\mu_2)M^{1/2}}$$

[alternatively $2(\mu_1+\mu_2)\mu_3 M\xi_1=(\mu_1^2-M)q_1-(\mu_2^2-M)q_2$] and so by (2.7), (2.3) and (1.12)

$$B(w,\xi_l)=2M\xi_3 B(\xi_3,\xi_l) \qquad (l=1,2)$$

verifying thus (2.4) for $k=3, l=1$ and $k=3, l=2$.

It rests to show (2.4) in the cases $k<3, l<3$. Using (2.5) we get

$$B(q_3,q_3)=4(\mu_1+\mu_2)^2 M\varrho, \quad B(q_1,q_2)=4(\mu_3^2-M)M\varrho \quad \text{(cycl.)}.$$

The last equalities combined with (2.8) allow us to calculate $B(\xi_1,\xi_1)$, $B(\xi_1,\xi_2)$ and $B(\xi_2,\xi_2)$.

3. To calculate the Jacobian mentioned at the end of Section 1 we factorize it in the following way:

$$\frac{\partial(r_1,r_2,r_3)}{\partial(\xi_1,\xi_2,\xi_3)}=\frac{\partial(r_1,r_2,r_3)}{\partial(\varrho_1,\varrho_2,\varrho_3)}\cdot\frac{\partial(\varrho_1,\varrho_2,\varrho_3)}{\partial(\varrho,\xi_1,\xi_2)}\cdot\frac{\partial(\varrho,\xi_1,\xi_2)}{\partial(\xi_1,\xi_2,\xi_3)}.$$

The value of the first factor is $(8r_1r_2r_3)^{-1}$ and that of the last factor is ξ_3/ϱ by (1.7). The middle factor is a constant different from 0. To calculate its value we solve equations (1.11_1), (1.11_2) and (1.13) with respect to the unknowns $\sigma_1,\sigma_2,\sigma_3$. Thus, by (1.10) we have

(3.1) $$2\varrho_1=\sigma_2+\sigma_3=(\mu_2+\mu_3)(\varrho+c_2\xi_1-s_2\xi_2),$$

(3.2) $$2\varrho_2=\sigma_3+\sigma_1=(\mu_3+\mu_1)(\varrho+c_1\xi_1+s_1\xi_2),$$

(3.3) $$2\varrho_3=\sigma_1+\sigma_2=(\mu_1+\mu_2)(\varrho-\xi_1)$$

where

(3.4) $$c_i=\frac{M-\mu_i^2}{(\mu_1+\mu_2)(\mu_i+\mu_3)}, \quad s_i=\frac{2\mu_i\sqrt{M}}{(\mu_1+\mu_2)(\mu_i+\mu_3)} \qquad (i=1,2).$$

From the system (3.1), (3.2), (3.3) we get $\partial(\varrho_1,\varrho_2,\varrho_3)/\partial(\varrho,\xi_1,\xi_2)=M^{3/2}/2$ and hence (ii) at the end of Section 1 holds.

4. The transformation (1.11) maps D_r into D_ξ and, conversely, to each point of D_ξ corresponds one and only one point in D_r. For showing this we first remark that for the quantities defined in (3.4) $c_i^2+s_i^2=1$ holds. Thus there is a real β_i such that $c_i=\cos\beta_i$, $s_i=\sin\beta_i$ $(i=1,2)$. Let now (ξ_1,ξ_2,ξ_3) be an interior point of D_ξ, that is $\xi_3>0$. Then, introducing the notations

$$\xi_1=\varrho'\cos\varphi, \quad \xi_2=\varrho'\sin\varphi, \quad \text{where} \quad 0\leqq\varrho'=(\xi_1^2+\xi_2^2)^{1/2}<\varrho,$$

we have

$$2\varrho_1=(\mu_2+\mu_3)[\varrho+\varrho'\cos(\varphi+\beta_2)],$$

$$2\varrho_2=(\mu_3+\mu_1)[\varrho+\varrho'\cos(\varphi-\beta_1)],$$

$$2\varrho_3=(\mu_1+\mu_2)[\varrho-\varrho'\cos\varphi].$$

Since by (1.2) at most one of the non-negative quantities μ_i can vanish, we see that to each triplet $(\xi_1, \xi_2, \xi_3) \in D_\xi$ correspond positive quantities $\varrho_1, \varrho_2, \varrho_3$ or a unique point (r_1, r_2, r_3) in the domain D_r^+: $r_1 > 0, r_2 > 0, r_3 > 0$ of the real (r_1, r_2, r_3) space.

Let now (r_1^1, r_2^1, r_3^1) be a point in D_r and its image in D_ξ be $(\xi_1^1, \xi_2^1, \xi_3^1)$. Suppose there exists a point $(\xi_1^2, \xi_2^2, \xi_3^2)$ in D_ξ such that it is the image of some point (r_1^2, r_2^2, r_3^2) in $D_r^+ \setminus D_r$. Join $(\xi_1^1, \xi_2^1, \xi_3^1)$ and $(\xi_1^2, \xi_2^2, \xi_3^2)$ by a straight line segment in D_ξ. This straight line is the image of some continuous curve in D_r^+. This curve intersects the boundary of D_r in some point (r_1^0, r_2^0, r_3^0) the image of which lies by (1.4) and (1.12) on the boundary $\xi_3 = 0$ of D_ξ: a contradiction.

5. Finally we show in an elementary way a consequence of the theorems of Baouendi—Goulaouic and Froim, namely that any solution of (1.9) which is an analytic function of ξ_3 on the real point set D^1: $|\xi_1 - \xi_{10}| < R_1$, $|\xi_2 - \xi_{20}| < R_2$, $|\xi_3| < \varepsilon_1$ is uniquely determined on D^1 by its values taken on the rectangle R: $|\xi_1 - \xi_{10}| < R_1$, $|\xi_2 - \xi_{20}| < R_2$, $\xi_3 = 0$ provided $(V-E)/(4\varrho)$ can be expanded in D^1 into a power series $\sum \alpha_n(\xi_1, \xi_2)\xi_3^n$, where the coefficients $\alpha_n(\xi_1, \xi_2)$ are infinitely many times differentiable in R.

Indeed for each solution ψ of (1.9) which can be expanded into a power series $\psi = \sum \psi_n(\xi_1, \xi_2)\xi_3^n$ $[(\xi_1, \xi_2, \xi_3) \in D^1]$ one has the recursion

$$n^2\psi_n = \sum_{k=0}^{n-2} \alpha_k \psi_{n-k-2} - \left(\frac{\partial^2 \psi_{n-2}}{\partial \xi_1^2} + \frac{\partial^2 \psi_{n-2}}{\partial \xi_2^2}\right) \qquad (n = 1, 2, \ldots),$$

$(\psi_{-1} = 0)$ which shows that ψ_n $(n > 0)$ is uniquely determined by $\psi_0 = \psi(\xi_1, \xi_2, 0)$. To express this in somewhat looser terms and with the help of the variables r_1, r_2, r_3: any sufficiently smooth solution of (1.3) is uniquely determined by its values taken on the boundary of D_r. The boundary points of D_r correspond to degenerate triangles of sides r_1, r_2, r_3, thus the three bodies with coordinates (x_i, y_i, z_i), $i = 1, 2, 3$ lie in these cases on a straight line segment in the Euclidean (x, y, z) space.

REFERENCES

[1] Baouendi, M. S.—Goulaouic, C., Cauchy problems with characteristic initial hypersurface, *Comm. Pure Appl. Math.* **26** (1973), 455—475.

[2] Bethe, H. A.—Salpeter, E. E., Quantum Mechanics of one- and two-electron systems, *Encyclopedia of Physics,* Vol. 35, Springer-Verlag, Berlin—Göttingen—Heidelberg, 1957.

[3] Courant, R.—Hilbert, D., *Methods of Mathematical Physics,* Interscience Publishers, New York, 1953.

[4] Froim, V., Linear scalar partial differential equations with regular singularities on a hyperplane, *Diff. Uravn.* **9** (1973), 533—541.

[5] Gronwall, T. H., A special conformally euclidean space of three dimensions occurring in wave mechanics, *Ann. of Math.* **33** (1932), 279—293.

[6] Gronwall, T. H., The helium wave equation, *Phys. Rev.* **51** (1937), 655—660.

[7] Hylleraas, E. A., Neue Berechnung der Energie des Heliumatoms im Grundzustande, *Zeitschr. f. Phys.* **54** (1929), 347—366.

[8] Kato, T., On the eigenfunctions of many-particle systems in quantum mechanics, *Comm. Pure Appl. Math.* **10** (1957), 151—177.

Mathematical Institute of the Hungarian Academy of Sciences
H—1053 Budapest, Reáltanoda u. 13—15

(Received September 10, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 193—200.

INTERPOLATION IN WEAKLY ASSOCIATIVE LATTICES

by

E. FRIED

It is well-known that no non-trivial lattice enjoys the interpolation property (further on IP). In fact, for $a<b$ no function with $f(a)=b, f(b)=a$ can be interpolated (by a polynomial function).

However, if we weaken the lattice axioms such that the "partial order" need not be transitive we get a class of algebras containing, also, functionally complete algebras. The class we get in this way is the variety of weakly associative lattices (further on WAL) where all the lattice axioms hold but the associative laws, which are weakened in the following way:

$$[(x\vee y)\wedge(x\vee z)]\wedge x = [(x\wedge y)\vee(x\wedge z)]\vee x = x.$$

As an important example, we mention the "triangle" T consisting of three elements $0, 1, 2$ endowed with the relation $0<1<2<0$ ($a\leqq b$ is equivalent to any of the relations $a=a\wedge b$ and $b=a\vee b$). In fact, the triangle was the first WAL which was proved to be functionally complete (i.e., finite and satisfying IP) (see [7]).

There are two generalizations of the triangle, which are, in many respects, natural.

We shall say that a WAL $\mathfrak{A}$ satisfies the unique bound property (further on UBP) if any two-element subset of the underlying set A of $\mathfrak{A}$ has exactly one upper bound and exactly one lower bound. (For example, the two-element lattice or the triangle satisfies UBP.) It was proven in [4] that if a WAL satisfies UBP and it contains more than two elements then it enjoys IP. The reason for this result is that there exists a term which represents the dual discriminator in each WAL satisfying UBP. (A ternary function $f(x, y, z)$ is called the dual discriminator if $f(a, a, c)=a$ and $f(a, b, c)=c$, whenever a and b differ.) This fact gives a reason why the variety **U** (generating by the WALs satisfying UBP) has the congruence extension property (further on CEP). An algebra $\mathfrak{A}$ has CEP if for each subalgebra $\mathfrak{B}$ of $\mathfrak{A}$ and congruence Θ of $\mathfrak{B}$ there exists a congruence Φ of $\mathfrak{A}$ such that $\Theta=(\mathfrak{B}\times\mathfrak{B})\cap\Phi$. A variety satisfies CEP if each member of the variety has CEP. It was proven in [2] that **U** is the greatest subvariety of WALs satisfying CEP.

The other generalization of the triangle is the class of tournaments. One can define a tournament as a relational system $\mathcal{T}$ with an antisymmetrical relation $<$ where for each pair a, b of its elements exactly one of the relations $a<b, a=b, b<a$ holds. Each tournament becomes a WAL when one defines for each $a\in\mathcal{T}$: $a\vee a=a\wedge a=a$, and $a\vee b=b, a\wedge b=a$ whenever $a, b\in\mathcal{T}$ satisfy $a<b$.

Since IP obviously implies simplicity we have to deal with simple algebras only. In [6], there was given a sufficient condition for a simple tournament to have IP. We prove:

THEOREM 1. *Any simple tournament with more then two elements has IP.*

For this end we need the following

LEMMA. *A WAL $\mathfrak{A}$ has IP iff it contains at least three elements and it satisfies: For each $a, b, c \in \mathfrak{A}$ with $a \neq b$ there exists a unary polynomial-function p such that $p(a)=a, p(b)=c$.*

PROOF. The condition is, clearly, necessary. For the sufficiency we prove first the so-called 2-interpolation property for $\mathfrak{A}$. Let $a \neq b, u, v \in A$ and the function f be given such that $f(a)=u$, $f(b)=v$. In case $u \neq b$ there exist, by the condition of the Lemma, polynomial-functions p, q such that $p(a)=u, p(b)=b$ and $q(u)=u$, $q(b)=v$. Then, the polynomial-function $qp(x)$ interpolates f on $\{a, b\}$. We have, by symmetry, the same result if $v \neq a$. Suppose, finally, $u=b$ and $v=a$. Since $\mathfrak{A}$ has at least three elements it contains some $c \notin \{a, b\}$. By the first case we have polynomial-functions p, q, r such that $p(a)=a, p(b)=c$; $q(a)=b, q(c)=c$; $r(b)=b$, $r(c)=a$ and so $rqp(x)$ interpolates f on $\{a, b\}$.

Since different n-tuples differ at least in one place, if $(a_1, \dots, a_n) \neq (b_1, \dots, b_n)$ there exists a polynomial-function $f(a_1, \dots, a_n)=u$, $f(b_1, \dots, b_n)=v$, where u and v are given at will.

It is easy to see that the polynomial-function $m(x, y, z)=[(x \vee y) \wedge (x \vee z)] \wedge (y \vee z)$ is a majority function for WALs. Using this fact one can prove the IP of $\mathfrak{A}$ by induction on the number of places. The proof is left to the reader.

(For a generalization of the idea of the Lemma see [3].)

PROOF of Theorem 1. A subset A of an underlying set T of a tournament is convex if for $a, b \in A$ the condition $a \leqq x \leqq b$ implies $x \in A$, for each $x \in T$. The singletons and T itself are trivial convex subsets, all others are proper ones. A partition consisting of convex subsets is clearly a congruence relation of the tournament if considered as a WAL. This implies that a simple tournament has no proper convex subsets.

Now, let $a \neq b$ be elements of T and consider the subset B, consisting of all $p(b)$ where p is a polynomial-function sending a to a. Take for p the constant map with the value a and the identity function. Both these functions send a to a and b to a and b, respectively. This implies $a, b \in B$. Suppose $u, v \in B$ and $u \leqq w \leqq v$. Since we have a tournament, a and w are comparable, e.g., $a \leqq w$. Since $v \in B$, there exists a unary polynomial-function p with $p(a)=a$ and $p(b)=v$. Then, for $q(x)=$ $=w \wedge p(x)$ we have $q(a)=w \wedge a=a$ and $q(b)=w \wedge v=w$, yielding $w \in B$. Thus, B is convex and contains at least two elements. Hence, $B=T$ and the application of the Lemma finishes the proof.

These results give rise to the question: has every simple WAL IP if it is not a lattice?

The answer of this question is in the negative, moreover the following result holds:

THEOREM 2. *Each WAL is contained in a simple one which does not have IP. (For a finite WAL the presented construction gives a finite extension.)*

In spite of Theorem 2 one must not think that only very special WALs have IP.

THEOREM 3. *Each (finite) WAL is contained in a (finite) WAL which satisfies IP.*

It would be easy to prove this theorem in the following way. By adding new elements one can get a partial WAL where a single partial function $\varphi(a)=c$, $\varphi(b)=d$ can be interpolated. This process would give a WAL having IP in infinitely many steps. The disadventage of this method is that the extension is always infinite.

For the proofs of both theorems we need some remarks. Let us consider elements of $\mathfrak{A}\times\mathfrak{A}$ where $\mathfrak{A}$ is a WAL. We shall denote by $(a, b)\to(c, d)$ the existence of a unary polynomial-function f such that $c=f(a)$ and $d=f(b)$. Clearly, $(a, b)\to\to(c, d)$ means that (c, d) belongs to the diagonal subalgebra of $\mathfrak{A}\times\mathfrak{A}$ generated by (a, b). By the Lemma, $\mathfrak{A}$ has IP iff the following holds:

For each (c, d) and offdiagonal (a, b) the relation $(a, b)\to(c, d)$ is valid; and $|\mathfrak{A}|>2$.

CLAIM 1. If $a<b<c$ then $(a, c)\to(a, b)$ and $(a, c)\to(b, c)$.

CLAIM 2. Let $a<b<c<a$ and $u, v, u', v'\in\{a, b, c\}$ such that $u\neq v$. Then one has $(u, v)\to(u', v')$.

CLAIM 3. $(a, b)\to(a, a\vee b)$ and $(a, b)\to(a, a\wedge b)$.

Claims 1 and 3 are obvious. Claim 2 follows from the result in [7] (but the reader can compute it easily).

PROOF of Theorem 3. Let $\mathfrak{A}=\langle A;\vee,\wedge\rangle$ be any non-trivial WAL. Let A' be a set disjoint from A such that A' and A have the same cardinality. Let us fix a bijection $A\to A'$ and let a' denote the image of $a\in A$ under this bijection. Now, we define a tournament $\mathfrak{A}'=\langle A';\vee,\wedge\rangle$ as follows: If $a<b$ then $b'<a'$, while for incomparable a and b we define $a'<b'$ or $b'<a'$ at will.

Let, finally, $A^*=A\cup A'\cup\{0, 1\}$ where neither 0 nor 1 belong to $A\cup A'$. We extend the relation $<$ to A^* such that $a<a'$ and $0<x<1<0$ for all $x\in A\cup A'$.

We claim the following:

(i) $\mathfrak{A}^*=\langle A^*;\vee,\wedge\rangle$ is a WAL,
(ii) $\mathfrak{A}$ is a subalgebra of $\mathfrak{A}^*$,
(iii) $\mathfrak{A}^*$ satisfies IP.

First of all, we mention that comparable elements have, clearly, l.u.b. and g.l.b. Suppose, x and y are not comparable. In case $x, y\in A$ then their only common upper (lower) bound in $A^*\setminus A$ is 1 and 0, respectively. Then, considering $x\vee y$ and $x\wedge y$ in $\mathfrak{A}$, we have, $x\vee y<1$ and $0<x\wedge y$. This proves that they have the same l.u.b. and g.l.b. in $\mathfrak{A}$ and in $\mathfrak{A}^*$. This yields, immediately, (ii). If either x or y does not belong to A then their incomparability gives that one of them is of the form a' and the other is $b\in A$ with $a\neq b$. Let $U(H)$ and $L(H)$ the set of upper and lower bounds of a set $H\subseteq A^*$, respectively. Now,

$$U(\{a', b\})\subseteq U(a')\cap U(b)\subseteq(A'\cup\{1\})\cap(\{b'\}\cup A\cup\{1\})=\{b', 1\},$$

$$L(\{a', b'\})\subseteq L(a')\cap L(b)\subseteq(\{a\}\cup A'\cup\{0\})\cap(A\cup\{0\})=\{a, 0\}$$

yield the existence of $a'\vee b$ and $a'\wedge b$, because of $b'<1$ and $0<a$. Thus, $\mathfrak{A}^*$ is a WAL.

Now, we prove that $\mathfrak{A}$ has IP. By Claim 3 and by the remark before Claim 1 it is enough to prove $(x, y)\to(u, v)$ for each $x<y$ and u, v.

We consider the case $0\notin\{x, y\}, 1\notin\{x, y\}$. Then, we have three possibilities:

$$x=a, y=b \quad (\text{in } A);\ x=a', y=b' \quad (\text{in } A');\quad \text{and}\quad x=a, y=a'.$$

In the first case $(a, b)\vee(a', a')=(a', 1)$ yields $(a, b)\to(a', 1)$. In the second case $(a', b')\vee(a, a)=(a', 1)$ yields $(a', b')\to(a', 1)$. In the third case $(a, a')\vee(b, b)=(b, 1)$ yields $(a, a')\to(b, 1)$, for $a<b$, and $(a, a')\wedge(b, b)=(b, 0)$ yields $(a, a')\to(b, 0)$, for $b<a$.

Thus, we have always either $(x, y)\to(u, 1)$ or $(x, y)\to(u, 0)$ with $u\notin\{0, 1\}$ (if $x<y$), since no element of $\mathfrak{A}$ is both the least and the greatest element of $\mathfrak{A}$. Applying Claim 2 we get both $(x, y)\to(0, 1)$ and $(x, y)\to(1, 0)$, for $x<y$. Now, using Claim 1, we get $(x, y)\to(u, v)$ whenever both (x, y) and (u, v) are comparable. Thus, in order to prove IP it is enough to show that any diagonal subalgebra $\mathfrak{B}$ of $\mathfrak{A}^*\times\mathfrak{A}^*$ which contains all comparable pairs of elements is equal to $\mathfrak{A}^*\times\mathfrak{A}^*$.

(i) $(a', b')\in\mathfrak{B}$ and $(a', b')\wedge(a\vee b, a\vee b)=(a, b)$ implies $(a, b)\in\mathfrak{B}$.

(ii) If $a<b$, we have $(b, b')\vee(a, a)=(b, a')$ and $(b, a')\wedge(a', b')=(a, b')$. Thus, $(b, a'), (a, b')\in\mathfrak{B}$ and, by Claim 2 also (a', b), $(b', a)\in\mathfrak{B}$.

(iii) Let, finally, a and b incomparable elements of A. We may suppose $a'<b'$. By the previous result $(a', a\vee b)$, $(b', a\vee b)\in\mathfrak{B}$. Thus, by $(a', b)=(a', a\vee b)\wedge(b', b)$ and $(a, b')=(a', (a\wedge b)')\wedge(a\vee b, b')$ we have $(a', b), (a, b')\in\mathfrak{B}$, proving Theorem 3.

Before proving Theorem 2 we need the following

PROPOSITION 1. *Let $\mathfrak{A}=\langle A;\ \vee, \wedge\rangle$ be a simple WAL containing a triangle $a<b<c<a$. Let B be the disjoint union of A and $\{0, 1, p, q\}$. We extend the relation $<$ from A to B such that*

(i) *0 is the least and 1 is the greatest element of B,*

(ii) *$q<a<p$.*

Then, we get a simple WAL $\mathfrak{B}=\langle B;\ \vee, \wedge\rangle$ containing $\mathfrak{A}$ as a subalgebra.

PROOF. The straightforward verification that $\mathfrak{B}$ is a WAL and contains $\mathfrak{A}$ as a subalgebra we leave to the reader.

Now, let $\Theta>\omega$ be any congruence of $\mathfrak{B}$, and x, y be different elements of a congruence class C of Θ.

Case 1. $x, y\in A$. Then, by the simplicity of $\mathfrak{A}$, we have $A\subseteqq C$. Thus, $(p, p)\wedge(a, c)=(a, 0)$ and $(q, q)\vee(a, b)=(a, 1)$ yields $0, 1\in C$. Conditions $0<p<1, 0<q<1$ give us $B\subseteqq C$, i.e., $\Theta=\iota$.

Case 2. $x\in A, y\in\{p, q\}$. By duality, we may suppose $y=p$. If $x\neq a$, then by $(p, x)\wedge(a\vee x, a\vee x)=(a, x)$ we get $a, x\in C$, and so Case 1 applies. In case $x=a$ we get $(a, p)\vee(c, c)=(a, 1)$ and $(a, p)\vee(b, b)=(b, 1)$, i.e., $a\equiv 1\equiv b$, yielding $a, b\in C$ so we can apply Case 1 again.

Case 3. (x, y) and A are disjoint. By duality and Claim 3 we may suppose that $x=0$ and, by Claim 1, $y=q$. Then $(0, q)\vee(c, c)=(c, a)$ yields $a\equiv c$, i.e., $\Theta=\iota$ by Case 1.

Case 4. $x\in A, y\in\{0, 1\}$. We may suppose $y=0$, by duality. If $x\neq a$ we get $(0, x)\vee(p, p)=(p, 1)$ yielding $\Theta=\iota$ by Case 3. The case $x=a$ gives, e.g., $0\equiv q$, by Claim 1, and we can apply Case 3 again.

PROPOSITION 2. *Let $\mathfrak{A}=\langle A; \vee, \wedge\rangle$ and $\mathfrak{B}=\langle B, \vee, \wedge\rangle$ be simple WALs with a least and a greatest element. Suppose* 0 *is the least element of $\mathfrak{A}$ and* 1 *is the greatest element of $\mathfrak{B}$, further $A\cap B=\{p\}$ where p is the greatest element of $\mathfrak{A}$ as well as the least element of $\mathfrak{B}$. Let $\leqq$ denote the following relation on $C=A\cup B$: $a_1\leqq a_2$ iff $a_1=a_1\wedge a_2$ $(a_1, a_2\in A)$, $b_1\leqq b_2$ iff $b_1=b_1\wedge b_2$ $(b_1, b_2\in B)$, $a<1$ for all $a\in A$ and $0<b$ for all $b\in B$.*

With this relation we get a WAL $\mathfrak{C}=\langle C; \vee, \wedge\rangle$ containing both $\mathfrak{A}$ and $\mathfrak{B}$ as subalgebras. If $|A|, |B|>2$ then $\mathfrak{C}$ is simple and does not satisfy IP.

PROOF. The easy computation that $\mathfrak{C}$ is a WAL containing both $\mathfrak{A}$ and $\mathfrak{B}$ as subalgebras is left to the reader.

First we prove that $\mathfrak{C}$ is simple. We consider a congruence $\Theta>\omega$, therefore there are elements $x\neq y$ such that $(x, y)\in\Theta$. For simplicity, it is enough to prove, by Claim 1, that $(0, 1)\in\Theta$.

We may suppose, by duality, that $x\in A$ but $x\notin B$. Now, we have five cases:

Case 1. $x\neq 0, y\in B, y\notin\{p, 1\}$. Then $x\vee y=1, x\wedge y=0$ finishes the proof of Case 1.

Case 2. $x\neq 0, y=1$. Let $z\in B, z\notin\{p, 1\}$. We have $x<p<1$ and $p<z<1$. Claim 1 implies $(x, p), (p, 1), (p, z)\in\Theta$, hence, by transitivity $(x, z)\in\Theta$. Thus, Case 1 applies.

Case 3. $x\neq 0, y=p$. Choose any $z\in B, z\notin\{p, 1\}$. Then, $(z, z)\vee(x, p)=(1, z)$ yields $(1, z)\in\Theta$, implying $(p, 1)\in\Theta$, by the simplicity of $\mathfrak{B}$. Thus, we have $(x, 1)\in\Theta$ and we can apply Case 2.

Case 4. $x, y\in A$. Since $\mathfrak{A}$ is simple, there exist $z\neq 0, z\in A$ such that $(z, p)\in\Theta$. Then, we can apply Case 3.

Case 5. $x=0, y\in A$. If $y=1$ we are done. Otherwise, we have Case 2, by duality.

Now, we prove that $\mathfrak{C}$ does not satisfy IP. By the Lemma it is enough to prove that $\mathfrak{C}\times\mathfrak{C}$ contains a proper diagonal subalgebra.

We shall prove that the elements of $A\times A\cup B\times B$ form a subalgebra of $\mathfrak{C}\times\mathfrak{C}$. By duality, it is enough to prove that this set is closed under $\vee$. Suppose, $(x_1, y_1), (x_2, y_2)\in A\times A\cup B\times B$, and let $(x, y)=(x_1, y_1)\vee(x_2, y_2)$ such that $x\in A$ and $y\in B$. Then, $x_1\vee x_2=x$ yields $x_1, x_2\in A$ implying $y_1, y_2\in A$, i.e., $y\in A$ as well. Since this algebra is obviously proper and diagonal, the proposition is proven.

PROOF of Theorem 2. Let $\mathfrak{A}$ be any non-trivial WAL. By Theorem 3 it is contained in a simple WAL $\mathfrak{A}_1$ which contains a triangle. Then, by Proposition 1, it is contained in a simple WAL $\mathfrak{A}_2$ having a least and a greatest element. Now, we can use Proposition 2. We choose $\mathfrak{A}_2$ for $\mathfrak{A}$ and, say, the five-element modular simple lattice for $\mathfrak{B}$. Thus, we have embedded $\mathfrak{A}$ into a simple WAL which does not satisfy IP. Clearly, if starting with a finite one we get finite WAL in each step. Hence, the theorem is proved.

REMARK. The construction of Proposition 2 can be generalized as follows: Let $\mathfrak{A}$ be a WAL. We put in each interval $a<b$ a WAL $\mathfrak{A}_{a,b}$ with the least element a and the greatest element b. In $\mathfrak{A}$ and in each $\mathfrak{A}_{a,b}$ we have the original relation and we add the following relations:

For $x\in\mathfrak{A}_{a,b}$ let $c\leqq x\leqq d$ whenever $c\leqq a, b\leqq d$ $(c, d\in\mathfrak{A})$. In this way we get a WAL $\mathfrak{A}^*$. One can give conditions on $\mathfrak{A}$ when the simplicity of all $\mathfrak{A}_{a,b}$ implies the simplicity of $\mathfrak{A}^*$.

Finally, we give a connection between CEP and a weaker form of IP.

As we mentioned **U** is the largest variety of WALs satisfying CEP. This need not imply that any WAL which satisfy CEP belongs to **U**. However, we shall prove that if there is a WAL not belonging to **U** which satisfies CEP then there exists some other WAL with the same properties which contains a very special simple WAL as subdirect component. (I conjecture that such a special simple WAL does not exist.)

We shall say that a WAL belongs to AIP (a single interpolation) if it contains a three-element chain $a<b<c$ such that for some unary polynomial-function f either $\{f(a), f(b)\}=\{b, c\}$ or $\{f(b), f(c)\}=\{a, b\}$ hold. The class of all other WALs will be denoted by BIP.

PROPOSITION 3. *Let $\mathfrak{A}$ satisfy CEP and let $\mathfrak{A}_1$ be a subdirect irreducible component of $\mathfrak{A}$. Let, further, $a, b, c, d, x_0, \dots, x_n$ be elements of $\mathfrak{A}_1$ and $f_0, \dots, f_{n-1}$ be unary polynomial-functions on $\mathfrak{A}_1$ such that $c=x_0$, $d=x_n$ and for all $i=0, \dots, n-1$ we have $\{f_i(a), f_i(b)\}=\{x_i, x_{i+1}\}$. Then $c\equiv d(\Theta(a, b))$ in the subalgebra generated by $\{a, b, x_0, \dots, x_n\}$.*

PROOF. Consider elements $A, B\in\mathfrak{A}$ whose first components are a, b, respectively. Let us, furthermore, enumerate the elements $f_i(A), f_i(B)$ in the following way $(i=0, \dots, n-1)$:

We denote by Y_0 the element in $\{f_0(A), f_0(B)\}$, which has $x_0(=c)$ as first component. Suppose $Y_0, \dots, Y_{2i}$ is already defined, then let Y_{2i+1} be that element of $\{f_i(A), f_i(B)\}$ which was not chosen to be Y_{2i} and we define Y_{2i+2} as that element of $\{f_{i+1}(A), f_{i+1}(B)\}$ which has as first component x_{i+1} (i.e., the first component of Y_{2i+1}). Thus, we have $Y_0\equiv Y_{2n+1}(\Phi)$ where Φ denotes the congruence of $\mathfrak{A}$ generated by $(A, B), (Y_1, Y_2), (Y_3, Y_4), \dots, (Y_{2n-1}, Y_{2n})$. By CEP the same relation holds in the subalgebra of $\mathfrak{A}$ generated by $\{A, B, Y_0, Y_1, \dots, Y_{2n}, Y_{2n+1}\}$ and another application of Mal'cev's lemma gives that the same holds for the first projection. This proves our proposition.

PROPOSITION 4. *If $c<d$ and $c\equiv d(\Theta(a, b))$ holds in a WAL $\mathfrak{A}$, then there exists a Mal'cev sequence $c=x_0, x_1, \dots, x_n=d$ satisfying $c\leqq x_i\leqq d$, for $i=0, 1, \dots, n$.*

PROOF. Let $f_0, \dots, f_{n-1}$ be unary algebraic functions such that $\{f_i(a), f_i(b)\}=\{y_i, y_{i+1}\}$ and $c=y_0, d=y_n$. Define $g_i(x)=(c\vee f_i(x))\wedge d$. Then, for $x_i=(c\vee y_i)\wedge d$ the proposition is satisfied.

PROPOSITION 5. *If $\mathfrak{A}$ satisfies CEP and $\mathfrak{A}_1$ is a subdirect irreducible component of $\mathfrak{A}$ then $\mathfrak{A}_1$ belongs to BIP.*

PROOF. Let $a<b<c$ a chain in $\mathfrak{A}_1$ such that for a unary algebraic function f we have, say, $\{f(b), f(c)\}=\{a, b\}$. By Proposition 3 we have $a\equiv b(\Theta(b, c))$ in the three element chain $\{a, b, c\}$ which is a contradiction.

THEOREM 4. *Suppose $\mathfrak{A}$ is a WAL satisfying CEP and $\mathfrak{A}_1$ is a subdirect irreducible component of $\mathfrak{A}$ not belonging to* **U** *and Θ is the smallest non-ω congruence of $\mathfrak{A}_1$. Then each Θ-class is a simple WAL which does not belong to* **U** *unless it is a singleton. Let $\mathfrak{B}$ denote the subalgebra of $\mathfrak{A}$ consisting of those elements whose first components belong to a given non-singleton Θ-class. Then $\mathfrak{B}$ satisfies CEP and the first component of $\mathfrak{B}$ is simple, does not belong to* **U** *and belongs to BIP.*

PROOF. The Θ-classes are simple by Propositions 3 and 4, using the fact that $c\equiv d(\Theta(a, b))$ yields $c\equiv c\vee d(\Theta(a, b))$ and $d\equiv c\vee d(\Theta(a, b))$.

Now, let $[a]\Theta$ be a non-trivial class. Then there is, clearly, an element $b\in[a]\Theta$ such that, say, $a<b$. In case $\Theta=\iota$ we are done. Thus, we may suppose that there exists a $c\notin[a]\Theta$ such that $b<c$ (or $c<a$ which is the dual case). We claim that we can change, if necessary, these elements such that $a<c$ is satisfied, too. $a\wedge c\equiv b\wedge c=b(\Theta)$ and $(a\wedge c)\vee b\equiv b\vee b=b(\Theta)$ implies that $a\wedge c, (a\wedge c)\vee b\in[a]\Theta$. Thus, $a\wedge c\neq c$, i.e., $a\wedge c<c$. Since $b<c$, by definition, and $(a\wedge c)\vee b\equiv c(\Theta)$, we have $(a\wedge c)\vee b<c$. The relation $a\wedge c\leqq a<b$ implies $a\wedge c\neq b$, i.e., $(a\wedge c)\vee b$ differs either from $a\wedge c$ or from b, yielding that either $\{a\wedge c, (a\wedge c)\vee b, c\}$ or $\{b, (a\wedge c)\vee b, c\}$ is a chain of the desired form. By definition of Θ, we have $a\equiv b(\Theta(b, c))$. Then, by Proposition 4, there exists a Mal'cev sequence $a=x_0, x_1, \dots, x_n=d$ satisfying $a\leqq x_i\leqq b$ $(i=0, 1, \dots, n)$. Suppose, $[a]\Theta\in$ **U**. Then, by simplicity, it satisfies the UBP, yielding $x_i\in\{a, b\}$ for each i in question. Thus, for one of the unary polynomial functions determining the Mal'cev sequence we have $\{a, b\}=\{f(b), f(c)\}$. This implies that this component is AIP which contradicts Proposition 5. Taking into consideration that CEP is hereditary to subalgebras, the theorem is proved.

As a consequence of Theorem 4 we mention that no WAL which does not belong to **U**, satisfies CEP if the following conjecture is true:

CONJECTURE. *Each simple WAL is either AIP or belongs to* **U**.

We remark that by Theorem 4 this conjecture is equivalent to the following "stronger" one: each subdirect irreducible WAL is either AIP or belongs to **U**.

All the simple WALs which were mentioned in this paper are AIP. Moreover, the simple lattices which do not belong to **U** are AIP, too. Indeed, these lattices contain one of the five-element non-distributive lattices, for which the existence of the chain in question can be given by an easy computation.

REFERENCES

[1] Fried, E., Subdirectly irreducible weakly associative lattices with the congruence extension property, *Ann. Univ. Sci. Budapest. Eötvös Sect. Math.* **17** (1974), 59—65.

[2] Fried, E.—Grätzer, G.—Quackenbush, R. W., The equational class generated by weakly associative lattices with the unique bound property, *Ann. Univ. Sci. Budapest. Eötvös Sect. Math.* **22—23** (1979—1980), 205—211.

[3] Fried, E.—Kaiser, H.—Márki, L., An elementary way for polynomial interpolation in universal algebras (to appear in *Algebra Universalis*).

[4] Fried, E.—Pixley, A. F., The dual discriminator function in universal algebra, *Acta Sci. Math. (Szeged)* **41** (1979), 83—100.

[5] Grätzer, G., *Universal Algebra,* Van Nostrand, Princeton, 1968.

[6] Park, R. E., Equational classes of Non-Associative Ordered Algebras, Ph. D. Thesis, UCLA, 1976.

[7] Quackenbush, R. W., The triangle is functionally complete, *Algebra Universalis* **2** (1972), 128.

Department of Algebra and Number Theory, Roland Eötvös University
H—1088 Budapest, Múzeum krt. 6—8

(Received September 14, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 201—205.

ON THE CESÀRO SUMMABILITY OF POWER SERIES ON THE CIRCLE OF CONVERGENCE

by

J. DEÁK

ALPÁR [1] has proved the following two theorems for $\alpha \geqq 0$:

THEOREM 1. *Suppose that*

$$f(z) = \sum_{p=0}^{\infty} a_p z^p \tag{1}$$

is (C, α) summable at the point ζ, $|\zeta|=1$. Let φ be a conformal univalent mapping of the unit disc onto itself. Then

$$f[\varphi(z)] = \sum_{p=0}^{\infty} b_p z^p \tag{2}$$

is $(C, \alpha+1/2)$ summable at ζ^* defined by $\varphi(\zeta^*)=\zeta$.

THEOREM 2. *Let φ be a conformal univalent mapping of the unit disc onto itself and suppose that φ is not a rotation. Let $0 \leqq \delta < 1/2$ and $|\zeta|=1$. Then there is a function f of the form* (1) (C, α) *summable at ζ such that* (2) *is not $(C, \alpha+\delta)$ summable at ζ^* defined by $\varphi(\zeta^*)=\zeta$.*

We are going to extend these theorems by proving that they hold for each $\alpha > -1$. In view of Alpár's results, it would be enough to give proofs for $-1 < \alpha < 0$. We shall show, however, that the validity of the theorems for $\alpha = 0$ involves their validity for all $\alpha > -1$. Our proof may be, at least in the case of Theorem 2, of some interest even for $\alpha > 0$, since Alpár [2] has found a new proof simpler than the original one for the special case $\alpha=0$ of Theorem 2. The two theorems will be proved simultaneously.

NOTATIONS. $(\mathcal{K}, \alpha)$ is the family of those functions of the form (1) for which a_p/A_p^α is convergent $(p \to \infty)$ where $A_p^\alpha = \binom{p+\alpha}{p}$ [in other words: a_p/p^α is convergent]. $(\mathcal{C}, \alpha)$ denotes the family of those functions of the form (1) which are (C, α) summable at 1 [i.e., $f(z)/(1-z)^{\alpha+1} \in (\mathcal{K}, \alpha)$]. $\mathcal{R}_{\mathrm{D}}$ is the system of functions regular in the closed unit disc D [i.e., in an open set containing D].

The following simple lemma proved in [3] will be used:

LEMMA. *If $M \in (\mathcal{K}, \alpha)$, $\alpha > -1$ and $N \in \mathcal{R}_{\mathrm{D}}$, then $MN \in (\mathcal{K}, \alpha)$.*

PROOF of both theorems for $\alpha > -1$. Let $\alpha > -1$ and $0 \leqq \delta \leqq 1/2$ and φ be fixed. We may suppose without loss of generality that $\zeta = \zeta^* = 1$, so $\varphi(1) = 1$.

Let us consider the following statement (which is, depending on the choice of α and δ, either true or false):

$\mathbf{P}(\alpha, \delta)$ $\qquad f \in (\mathscr{C}, \alpha) \Rightarrow f[\varphi] \in (\mathscr{C}, \alpha+\delta).$

Take now an $f \in (\mathscr{C}, \alpha)$ and put

$$L(z) = \frac{f(z)}{(1-z)^{\alpha+1}} = \sum_{p=0}^{\infty} l_p z^p \tag{3}$$

and

$$H_\delta(z) = \frac{L[\varphi(z)]}{(1-z)^\delta} = \sum_{p=0}^{\infty} h_p^\delta z^p. \tag{4}$$

We have

$$\frac{f[\varphi(z)]}{(1-z)^{\alpha+1+\delta}} = H_\delta(z) \left(\frac{1-\varphi(z)}{1-z} \right)^{\alpha+1}. \tag{5}$$

As $\varphi(1)=1$ and $\varphi'(1) \neq 0$:

$$\left(\frac{1-\varphi(z)}{1-z} \right)^{\alpha+1} \in \mathscr{R}_{\mathrm{D}}, \quad \left(\frac{1-z}{1-\varphi(z)} \right)^{\alpha+1} \in \mathscr{R}_{\mathrm{D}}. \tag{6}$$

Thus, by the Lemma and (5):

$$\frac{f[\varphi(z)]}{(1-z)^{\alpha+\delta+1}} \in (\mathscr{K}, \alpha+\delta) \quad \text{iff} \quad H_\delta \in (\mathscr{K}, \alpha+\delta)$$

and so $\mathbf{P}(\alpha, \delta)$ is equivalent to

$\mathbf{P}_1(\alpha, \delta)$ $\qquad L \in (\mathscr{K}, \alpha) \Rightarrow H_\delta \in (\mathscr{K}, \alpha+\delta).$

It is easy to see that

$$\frac{h_p^\delta}{A_p^{\alpha+\delta}} = \sum_{n=0}^{\infty} \gamma_{np}^{\alpha\delta} \frac{l_n}{A_n^\alpha} \qquad (p = 0, 1, 2, \dots) \tag{7}$$

where

$$\gamma_{np}^{\alpha\delta} = B_{np}^\delta \frac{A_n^\alpha}{A_p^{\alpha+\delta}}$$

and the B_{np}^δs are defined by

$$\frac{\varphi^n(z)}{(1-z)^\delta} = \sum_{p=0}^{\infty} B_{np}^\delta z^p$$

(remember that $1/(1-z)^{\alpha+1} = \sum_{p=0}^{\infty} A_p^\alpha z^p$). Now, according to (3), (4) and (7), $\mathbf{P}_1(\alpha, \delta)$ [and so $\mathbf{P}(\alpha, \delta)$] is equivalent to the following:

$\mathbf{P}_2(\alpha, \delta) \left\{ \begin{array}{l} \textit{The sequence-to-sequence transformation defined by the matrix} \\ [\gamma_{np}^{\alpha\delta}]_{np} \textit{ is convergence preserving.} \end{array} \right.$

By the KOJIMA—SCHUR theorem, $[\gamma^{\alpha\delta}_{np}]_{np}$ is convergence preserving iff

a) for each n fixed, $\gamma^{\alpha\delta}_{np}$ converges to a finite limit as $p\to\infty$;

b) the sequence $\left\{\sum_{n=0}^{\infty}\gamma^{\alpha\delta}_{np}\right\}_p$ converges to a finite limit as $p\to\infty$ and

c) $\sum_{n=0}^{\infty}|\gamma^{\alpha\delta}_{np}| = O(1) \quad (p\to\infty).$

We shall show that a) and b) always hold; thus $[\gamma^{\alpha\delta}_{np}]_{np}$ is convergence preserving iff c) is satisfied, so $\mathbf{P}_2(\alpha,\delta)$ [i.e., $\mathbf{P}(\alpha,\delta)$] holds iff

$$\mathbf{P}_3(\alpha,\delta) \qquad \sum_{n=1}^{\infty} n^{\alpha}|B^{\delta}_{np}| = O(p^{\alpha+\delta})$$

[remember that $A^{\alpha}_{p}\sim p^{\alpha}/\Gamma(1+\alpha)$].

PROOF of a). We have

$$B^{\delta}_{np} = \frac{1}{2\pi i}\oint\frac{\varphi^n(z)}{z^{p+1}(1-z)^{\delta}}\,dz.$$

In order to get an estimation for B^{δ}_{np}, we take the integral along the curve $\bigcup_{j=0}^{3}\mathsf{E}_j$ (with $0<\varepsilon<1$ depending on n,p and δ) where E_0 is the circular arc round 1 connecting $e^{-i\varepsilon}$ with $e^{i\varepsilon}$ inside D; E_1 is the segment $[e^{i\varepsilon}, Re^{i\varepsilon}]$; E_2 is the arc $Re^{i\tau}$ $(\varepsilon\leqq\tau\leqq 2\pi-\varepsilon)$; finally, E_3 is the segment $[Re^{-i\varepsilon}, e^{-i\varepsilon}]$. Here $R>1$ is chosen such that φ is regular in the disc $|z|\leqq R$ and there is a constant $c>1$ with

$$|\varphi(z)| < 1+c(|z|-1) \qquad (1\leqq|z|\leqq R).$$

Our estimations for the integrals on E_1, E_2 and E_3 will be independent of ε; for n and p fixed, the integral on E_0 can be made arbitrarily small by choosing a small enough ε. Thus the integral on E_0 can be disregarded in what follows. Now it will be shown that

$$(8) \qquad B^{\delta}_{np} = O(p^{\delta-1}) \qquad (n<p/2c)$$

uniformly in n. [To prove a), it would be enough to verify (8) for n fixed, but later we shall need this more general statement.] Now

$$B^{\delta}_{np} = \frac{1}{2\pi i}\sum_{j=0}^{3}\int_{\mathsf{E}_j}\frac{\varphi^n(z)}{z^{p+1}(1-z)^{\delta}}\,dz = \frac{1}{2\pi i}\sum_{j=0}^{3}I_j;$$

here, as $|1-z|>|z|-1$ on E_1, we have for $n<p/2c$:

$$|I_1| < \int_1^{\infty}\frac{[1+c(x-1)]^n}{x^{p+1}(x-1)^{\delta}}\,dx = \int_0^{\infty}\frac{(1+cy)^n}{(1+y)^{p+1}y^{\delta}}\,dy <$$

$$< \int_0^{\infty}\frac{(1+y)^{cn}}{(1+y)^{p+1}y^{\delta}}\,dy < \int_0^{\infty}\frac{dy}{(1+y)^{p/2}y^{\delta}} = 2\int_0^{\infty}\frac{t^{1-2\delta}}{(1+t^2)^{p/2}}\,dt.$$

Since

$$\int_0^{\infty}\frac{t^{1-2\delta}}{(1+t^2)^m}\,dt = \frac{\Gamma(1-\delta)\Gamma(m+\delta-1)}{2\Gamma(m)} \qquad (m>1),$$

(see [4], 3.241.4), we have $I_1=O(p^{\delta-1})$. Similarly, $I_3=O(p^{\delta-1})$. Further, $I_2=O(p^{\delta-1})$ is evident, so (8) has been proved and $\gamma_{np}^{\alpha\delta}=O(1/p^{\alpha+1})=o(1)$, i.e., a) holds.

Proof of b). Take $f(z)\equiv 1$. Then evidently $f[\varphi]\in(\mathscr{C},\alpha)$, hence, according to (5), (6) and the Lemma, $H_\delta\in(\mathscr{K},\alpha+\delta)$. But now $l_n=A_n^\alpha$ [see (3)], so the convergence of $h_p^\delta/A_p^{\alpha+\delta}$ gives b) [see (7)]. Thus we have proved that $\mathbf{P}(\alpha,\delta)$ and $\mathbf{P}_3(x,\delta)$ are equivalent.

By (8),

$$\sum_{n\leqq p/2c} n^\alpha |B_{np}^\delta| = O(p^{\delta-1}) \sum_{n\leqq p/2c} n^\alpha = O(p^{\delta-1})\,O([p/2c]^{\alpha+1}) = O(p^{\alpha+\delta}),$$

thus $\mathbf{P}_3(\alpha,\delta)$ [i.e., $\mathbf{P}(\alpha,\delta)$] is equivalent to

$$\mathbf{P}_4(\alpha,\delta) \qquad \sum_{n>p/2c} n^\alpha |B_{np}^\delta| = O(p^{\alpha+\delta}).$$

Starting from the equality

$$B_{np}^\delta = \frac{1}{2\pi i}\int_{|z|=\varrho} \frac{\varphi^n(z)}{z^{p+1}(1-z)^\delta}\,dz$$

with some $0<\varrho<1$, one can readily check that

$$|B_{np}^\delta| = O(r^n) \qquad (n > c^\circ p)$$

holds uniformly in p for some $0<r<1$ and $c^\circ>0$. c° is clearly independent of α. Thus

$$\sum_{n\geqq c^\circ p} n^\alpha |B_{np}^\delta| = O(1) \sum_{n\geqq c^\circ p} n^\alpha r^n =$$

$$= \begin{cases} O(1) \displaystyle\sum_{n\geqq c^\circ p} r^n = O(1)O(r^p) = O(p^{\alpha+\delta}) & (\alpha\leqq 0), \\ O(1) \displaystyle\sum_{n=0}^{\infty} A_n^\alpha r^n = O(1)\frac{1}{(1-r)^{\alpha+1}} = O(1) = O(p^{\alpha+\delta}) & (\alpha\geqq 0). \end{cases}$$

So $\mathbf{P}_4(\alpha,\delta)$ [i.e., $\mathbf{P}(\alpha,\delta)$] is equivalent to

$$\sum_{\substack{n \\ p/2c<n<c^\circ p}} n^\alpha |B_{np}^\delta| = O(p^{\alpha+\delta})$$

which is evidently equivalent to

$$\mathbf{P}_5(\alpha,\delta) \qquad \sum_{\substack{n \\ p/2c<n<c^\circ p}} |B_{np}^\delta| = O(p^\delta).$$

Theorem 1 means $\mathbf{P}(\alpha, 1/2)$ and Theorem 2 means that $\mathbf{P}(\alpha,\delta)$ does not hold if $\delta<1/2$ and φ is not a rotation. But $\mathbf{P}_5(\alpha,\delta)$ [and so $\mathbf{P}(\alpha,\delta)$] is independent of α, so if the theorems hold for some α (for example: $\alpha=0$), then they hold for each $\alpha>-1$ as well.

REFERENCES

[1] Alpár, L., Remarque sur la sommabilité des séries de Taylor sur leurs cercles de convergence III, *Magyar Tud. Akad. Mat. Kut. Int. Közl.* **5** (1960), 97—152.

[2] Alpár, L., Sur certains changements de variable des séries de puissances, *Studies in pure mathematics, to the memory of Paul Turán* (to appear).

[3] Deák, J., The influence of certain transformations of the independent variable on the Cesàro-summability of power series, *Period. Math. Hungar.* **3** (1973), 247—254.

[4] Gradstein, I. S.—Ryzhik, I. M., *Tables of integrals, sums, series and products,* Moscow, 1963 (in Russian).

Mathematical Institute of the Hungarian Academy of Science,
H—1053 Budapest, Reáltanoda u. 13—15

(Received September 20, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 207—211.

EINE UNGLEICHUNG FÜR KONVEXE FUNKTIONEN

von

WILHELM FLEISCHER

Gegeben sei eine streng monoton wachsende Folge reeller Zahlen $x_0, x_1, x_2, \dots$ mit $x_0=0, x_1=1$. Für $j \geqq 1$ werde mit δ_j der Abstand von x_j zum nächstgelegenen Punkt der Folge bezeichnet: $\delta_j = \min(x_j - x_{j-1}, x_{j+1} - x_j)$. Im Jahre 1968 zeigte R. E. L. Turner [1], daß stets $\sum_{j=1}^{\infty} x_j^{-2}\delta_j \leqq \pi^2/6$ gilt. A. Florian [2] bewies 1972 die Ungleichung

$$\sum_{j=1}^{\infty} x_j^{-s}\delta_j \leqq \zeta(s) = \sum_{k=1}^{\infty} k^{-s} \quad (s>1)$$

mit dem Zusatz, daß das Gleichheitszeichen nur für die Folge der natürlichen Zahlen ($x_j = j$ für $j=0, 1, 2, \dots$) richtig ist. Hier soll dies verallgemeinert werden durch den

Satz. *Ist $f(x)$ eine positive konvexe Funktion über dem Intervall $[1, +\infty)$ mit $\int_1^{\infty} f(x)\,dx < +\infty$, dann gilt mit den obigen Bezeichnungen stets*

$$\sum_{j=1}^{\infty} f(x_j)\delta_j \leqq \sum_{k=1}^{\infty} f(k) \tag{1}$$

und bei streng konvexem $f(x)$ ist das Gleichheitszeichen wieder nur für die Folge der natürlichen Zahlen richtig.

Hilfssatz 1. *Sei $g(x)$ eine monoton abnehmende Funktion über dem Intervall $[1, +\infty)$ und linear zwischen je zwei aufeinanderfolgenden natürlichen Zahlen: $g(x) = g([x])(1-\{x\}) + g([x]+1)\{x\}$ für alle $x \geqq 1$ ($[x] \leqq x < [x]+1$ mit natürlichem $[x]$ und $\{x\} = x - [x]$). Sind dann $x_0, x_1, \dots, x_n$ reelle Zahlen mit $x_0 = 0$ und $x_j - x_{j-1} \geqq 1$ für $1 \leqq j \leqq n$, so gilt stets*

$$\sum_{j=1}^{n} g(x_j)(x_j - x_{j-1}) \leqq \sum_{k=1}^{[x_n]} g(k) + g([x_n]+1)\{x_n\}.$$

Beweis (Induktion nach n). Für $x \geqq 1$ gilt

$$g(x)x = g(x) + g(x)(x-1) = g([x])(1-\{x\}) + g([x]+1)\{x\} + g(x)(x-1) \leqq$$
$$\leqq g([x])(1-\{x\}+x-1) + g([x]+1)\{x\} = g([x])[x] + g([x]+1)\{x\} \leqq$$
$$\leqq \sum_{k=1}^{[x]} g(k) + g([x]+1)\{x\},$$

was die Ungleichung für $n=1$ beweist. Sind jetzt $x_0, x_1, \dots, x_n$ reelle Zahlen mit $x_0=0, x_j-x_{j-1}\geqq 1$ für $1\leqq j\leqq n$ $(n>1)$ und ist die Ungleichung richtig für $x_0, x_1, \dots, x_{n-1}$, so gilt

$$\sum_{j=1}^{n} g(x_j)(x_j-x_{j-1}) = \sum_{j=1}^{n-1} g(x_j)(x_j-x_{j-1})+g(x_n)+g(x_n)(x_n-x_{n-1}-1) \leqq$$

$$\leqq \sum_{k=1}^{[x_{n-1}]} g(k)+g([x_{n-1}]+1)\{x_{n-1}\}+g([x_n])(1-\{x_n\})+g([x_n]+1)\{x_n\}+$$

$$+g([x_n])(x_n-x_{n-1}-1) = \sum_{k=1}^{[x_{n-1}]+1} g(k)-g([x_{n-1}]+1)(1-\{x_{n-1}\})+$$

$$+g([x_n])([x_n]-x_{n-1})+g([x_n]+1)\{x_n\} \leqq \sum_{k=1}^{[x_{n-1}]+1} g(k)+$$

$$+g([x_n])([x_n]-[x_{n-1}]-1)+g([x_n]+1)\{x_n\} \leqq \sum_{k=1}^{[x_n]} g(k)+g([x_n]+1)\{x_n\},$$

was Hilfssatz 1 beweist.

HILFSSATZ 2. *Sei $g(x)$ eine Funktion wie in Hilfssatz* 1 *und die Folge der Differenzen $g(n)-g(n+1)$ monoton abnehmend. Ist dann $1\leqq x\leqq y<y+\varepsilon<[y]+1$, so gilt*

$$g(x)+\varepsilon^{-1}\big(g(y+\varepsilon)-g(y)\big)(y-x) \geqq g(y).$$

BEWEIS. Es ist

$$\varepsilon^{-1}\big(g(y+\varepsilon)-g(y)\big) =$$

$$= \varepsilon^{-1}\big(g([y])(1-\{y\}-\varepsilon)+g([y]+1)(\{y\}+\varepsilon)-g([y])(1-\{y\})-g([y]+1)\,\{y\}\big) =$$

$$= \varepsilon^{-1}\big(-g([y])\varepsilon+g([y]+1)\varepsilon\big) = g([y]+1)-g([y]).$$

Somit bleibt $g(x)-g(y)\geqq\big(g([y])-g([y]+1)\big)(y-x)$ zu zeigen.

Ist $[x]=[y]$, so gilt

$$g(x)-g(y) = g([y])(1-\{x\})+g([y]+1)\,\{x\}-g([y])(1-\{y\})-g([y]+1)\,\{y\} =$$

$$= g([y])(\{y\}-\{x\})-g([y]+1)(\{y\}-\{x\}) =$$

$$= \big(g([y])-g([y]+1)(y-x).$$

Ist $[x]<[y]$, so gilt

$$g(x)-g(y) = g([x])(1-\{x\})+g([x]+1)\,\{x\}-g([y])(1-\{y\})-g([y]+1)\,\{y\} =$$

$$= \big(g([x])-g([x]+1)\big)(1-\{x\})+g([x]+1)+$$

$$+\big(g([y])-g([y]+1)\big)\,\{y\}-g([y]) \geqq \big(g([y])-g([y]+1)\big)(1-\{x\}+[y]-[x]-1+\{y\}) =$$

$$= \big(g([y])-g([y]+1)\big)(y-x).$$

Damit ist Hilfssatz 2 gezeigt.

HILFSSATZ 3. *Sei $g(x)$ eine Funktion wie in Hilfssatz 2, $g(x)>0$ für alle $x\geqq 1$ und $\sum_{k=1}^{\infty} g(k)<+\infty$. Seien weiters $x_0, x_1, \ldots, x_n$ reelle Zahlen mit $x_0=0, x_1=1$ und $x_1<x_2<\ldots<x_n$. Wird $\delta_j=\min(x_j-x_{j-1}, x_{j+1}-x_j)$ für $1\leqq j\leqq n-1$ und $\delta_n=x_n-x_{n-1}$ gesetzt, so gilt stets*

$$\sum_{j=1}^{n} g(x_j)\delta_j \leqq \sum_{k=1}^{\infty} g(k).$$

BEWEIS. Es werde für beliebige $n+1$-Tupel $\omega=(x_0, x_1, \ldots, x_n)$ mit obiger Gestalt $F(\omega)=\sum_{j=1}^{n} g(x_j)\delta_j$ mit den eben erklärten δ_j gebildet. Da

$$F(\omega)\leqq \sum_{j=1}^{n} g(x_j)(x_j-x_{j-1})\leqq g(1)+\int_1^{x_n} g(x)\,dx \leqq g(1)+\int_1^{\infty} g(x)\,dx =$$

$$= (3/2)g(1)+\sum_{k=2}^{\infty} g(k),$$

ist $\sup F(\omega)<+\infty$.

Da $g(x)x$ wegen $1/2\; g(x)x\leqq \int_{x/2}^{x} g(t)\,dt$ für x nach $+\infty$ gegen Null konvergiert, folgt mit dem Satz von Bolzano—Weierstraß, daß das Supremum für ein $\overline{\omega}$ angenommen wird:

$$F(\overline{\omega}) = \max_{\omega} F(\omega). \tag{2}$$

Sei $\overline{\omega}=(x_0, x_1, \ldots, x_n)$ und $l_j=x_j-x_{j-1}$ für $1\leqq j\leqq n$. Es gilt $1=l_1\leqq l_2\leqq\ldots\leqq l_n$. Angenommen dies wäre falsch. Dann gäbe es ein k, sodaß $l_{k-1}>l_k$. Sei dieses k maximal: $l_k\leqq l_{k+1}\leqq\ldots\leqq l_n$. Ist $l_k<l_{k+1}$, so gilt mit einer hinreichend kleinen positiven Zahl ε und $\overline{\omega}_\varepsilon=(x_0, x_1, \ldots, x_{k-1}, x_k+\varepsilon, x_{k+1}, \ldots, x_n)$ die Beziehung

$$\varepsilon^{-1}\big(F(\overline{\omega}_\varepsilon)-F(\overline{\omega})\big)=\varepsilon^{-1}\big(g(x_{k-1})(l_k+\varepsilon)+g(x_k+\varepsilon)(l_k+\varepsilon)+g(x_{k+1})(l_{k+1}-\varepsilon)-$$

$$-g(x_{k-1})l_k-g(x_k)l_k-g(x_{k+1})l_{k+1}\big)=g(x_{k-1})+\varepsilon^{-1}\big(g(x_k+\varepsilon)-g(x_k)\big)(x_k-x_{k-1})+$$

$$+g(x_k+\varepsilon)-g(x_{k+1})\geqq g(x_k)+g(x_k+\varepsilon)-g(x_{k+1})$$

wegen Hilfssatz 2 (bei $k=n$ ist $-g(x_{k+1})$ wegzulassen).

Der letzte Ausdruck ist aber positiv und damit wäre $F(\overline{\omega}_\varepsilon)>F(\overline{\omega})$ im Widerspruch zu (2).

Ist $l_k=l_{k+1}=\ldots=l_{k+p}<l_{k+p+1}\leqq\ldots\leqq l_n$, so gilt wieder mit einer hinreichend kleinen positiven Zahl ε und

$$\overline{\omega}_\varepsilon = (x_0, x_1, \ldots, x_{k-1}, x_k+\varepsilon, x_{k+1}+\varepsilon, \ldots, x_{k+p}+\varepsilon, x_{k+p+1}, \ldots, x_n)$$

die Beziehung

$$\varepsilon^{-1}\big(F(\bar{\omega}_\varepsilon)-F(\bar{\omega})\big) = \varepsilon^{-1}\big(g(x_{k-1})(l_k+\varepsilon)+g(x_k+\varepsilon)\,l_k+g(x_{k+1}+\varepsilon)\,l_k+\dots$$
$$\dots+g(x_{k+p}+\varepsilon)\,l_k+g(x_{k+p+1})(l_{k+p+1}-\varepsilon)-g(x_{k-1})\,l_k-g(x_k)\,l_k-$$
$$-g(x_{k+1})\,l_k-\dots-g(x_{k+p})\,l_k-g(x_{k+p+1})\,l_{k+p+1}\big) =$$
$$= g(x_{k-1})+\varepsilon^{-1}\big(g(x_k+\varepsilon)-g(x_k)\big)\,l_k+\varepsilon^{-1}\big(g(x_{k+1}+\varepsilon)-g(x_{k+1})\big)\,l_k+\dots$$
$$\dots+\varepsilon^{-1}\big(g(x_{k+p}+\varepsilon)-g(x_{k+p})\big)\,l_k-g(x_{k+p+1}) \geqq g(x_{k+p})-g(x_{k+p+1})$$

wegen Hilfssatz 2 (bei $k+p=n$ ist $-g(x_{k+p+1})$ wegzulassen).

Wieder ist der letzte Ausdruck positiv, was widersprüchlich ist. Somit gilt $1=l_1\leqq l_2\leqq\dots\leqq l_n$ und daher ist

$$F(\bar{\omega}) = \sum_{j=1}^{n} g(x_j)(x_j-x_{j-1}) \quad \text{mit} \quad x_j-x_{j-1}\geqq 1 \quad \text{für} \quad 1\leqq j\leqq n.$$

Wegen Hilfssatz 1 folgt

$$F(\bar{\omega}) \leqq \sum_{k=1}^{[x_n]} g(k)+g([x_n]+1)\{x_n\} \leqq \sum_{k=1}^{\infty} g(k)$$

und daraus Hilfssatz 3.

Beweis des Satzes. Sei $f(x)$ eine positive konvexe Funktion über $[1,+\infty)$ mit $\int_1^\infty f(x)\,dx<+\infty$ und $g(x)$ ebenfalls definiert für $x\geqq 1$ durch $g(n)=f(n)$ für alle natürlichen Zahlen n sowie $g(x)=g([x])(1-\{x\})+g([x]+1)\{x\}$ für alle $x\geqq 1$. Es gilt dann stets $f(x)\leqq g(x)$ und $g(x)$ erfüllt wegen $\sum_{k=1}^{\infty} g(k)=\sum_{k=1}^{\infty} f(k)\leqq$ $\leqq f(1)+\int_1^\infty f(x)\,dx$ die Voraussetzungen von Hilfssatz 3. Ist jetzt $x_0, x_1, x_2, \dots$ eine streng monoton wachsende Folge reeller Zahlen mit $x_0=0$, $x_1=1$ und wird $\delta_j=\min(x_j-x_{j-1}, x_{j+1}-x_j)$ für $j\geqq 1$ gesetzt, dann gilt für beliebiges natürliches n wegen Hilfssatz 3 die Beziehung

$$\sum_{j=1}^{n} f(x_j)\delta_j \leqq \sum_{j=1}^{n} g(x_j)\delta_j \leqq \sum_{k=1}^{\infty} g(k) = \sum_{k=1}^{\infty} f(k)$$

und daher auch

$$\sum_{j=1}^{\infty} f(x_j)\delta_j \leqq \sum_{k=1}^{\infty} f(k).$$

Bei streng konvexem $f(x)$ ist $f(x)<g(x)$ für alle x, die keine natürlichen Zahlen sind. Enthält in diesem Fall die Folge (x_n) nicht nur natürliche Zahlen, so gilt

$$\sum_{j=1}^{\infty} f(x_j)\delta_j < \sum_{j=1}^{\infty} g(x_j)\delta_j \leqq \sum_{k=1}^{\infty} g(k) = \sum_{k=1}^{\infty} f(k).$$

Sind die $x_j = n_j$ lauter natürliche Zahlen, so gilt

$$\sum_{j=1}^{\infty} f(n_j)\delta_j = \sum_{j=1}^{\infty} g(n_j)\delta_j \leqq \sum_{j=1}^{\infty} g(n_j)(n_j - n_{j-1}).$$

Der letzte Ausdruck ist wegen der strengen Monotonie von $g(x)$ stets kleiner als $\sum_{k=1}^{\infty} g(k)$, wenn mindestens eine Differenz $n_j - n_{j-1} > 1$ ist. Somit ist bei streng konvexem $f(x)$ das Gleichheitszeichen in (1) nur für die Folge der natürlichen Zahlen erfüllt. Damit sind beide Aussagen des Satzes bewiesen.

LITERATUR

[1] TURNER, R. E. L., An extremal problem, *Acta Math. Acad. Sci. Hungar.* **19** (1968), 437—440.

[2] FLORIAN, A., Über lineare Punktverteilungen, *Monatsh. Math.* **76** (1972), 31—42.

Mathematisches Institut der Universität Salzburg
A—5020 Salzburg, Petersbrunnstrasse 19

(Eingegangen am 27. September 1979)

REDUCTION OF THE NUMERICAL DIFFUSION IN MULTI-PHASE FLOW

by

GYÖRGY ADLER

Introduction

The numerical diffusion inherent to the usual finite difference formulae for the numerical solution of diffusion and convection problems is well-known: the numerical solutions of diffusion problems show higher diffusivity than the exact solution, while in the case of purely convection problems without any diffusion a diffusivity appears. In the case of a strong physical diffusion there is an easy way to overcome the problem: reducing the physical diffusion coefficient by the amount of the numerical diffusion, hence the numerical solution will have the required diffusivity (see e.g. [3]). In the case of a purely convective transport, or if the convection is preponderant over the physical diffusion, the remedy is not so straightforward since in this case the numerical solution will produce unwanted oscillations and will violate essential physical conditions (automatically satisfied by the exact solution), e.g., densities would become negative. Boris and Book in their paper [2] present a method to solve this problem.

In the present paper we shall discuss the case of multi-phase flow in a porous medium without any diffusion. (E.g., the immiscible flow of gas, oil and water in a rock.) This is a very strongly non-linear problem formulated in the form of a system of equations.

The method of Boris and Book is a multi-step method. The so-called "anti-diffusion" is applied in a separate step. In the case of multi-phase flow it seems that the anti-diffusion has to be incorporated in the first step where the basic flow is calculated, otherwise the algebraic condition (3) for the saturations cannot be ensured in a natural way. (The usually followed "natural" way is to use a convenient equation for the pressure to satisfy (3). But a purely diffusive transport is independent of pressure hence condition (3) could not be satisfied by manipulating the pressure in an anti-diffusion step.)

The problem of numerical diffusion is a great obstacle in the simulation (by the finite difference method) of certain oil recovery process where the position of a small amount of non diffusive chemical should be traced in the reservoir. While it is known from practice that this chemical moves through the reservoir always occupying a narrow strip, in the numerical solution the concentration spreads out on an unrealistically large area. (The use of finite element methods solves the problem in those cases only where the problem can be formulated in terms of elliptic or parabolic equations, which is not the case in multi-phase flow problems.) It is hopeful that the method presented in this paper — with proper adjustments — can be used to attack this problem.

Formulation of the study cases

We shall use the following formulation of the problem of multi-phase flow in a porous medium:

$$\Phi \frac{\partial(\varrho_l S_l)}{\partial t} = \frac{\partial}{\partial x}\left(\varkappa \frac{k_L}{\mu_L} \varrho_l S_l \frac{\partial p}{\partial x}\right) + F_l \tag{1}$$

$$0 \leqq S_l \leqq 1 \qquad (l = 1, 2, \dots, m) \tag{2}$$

$$\sum_{l=1}^{m} S_l = 1. \tag{3}$$

The physical meaning of the symbols is the following:

S	— saturation
p	— pressure
ϱ	— density
Φ	— porosity of the medium
$\varkappa$	— absolute permeability of the medium
k	— relative permeability of the phase
μ	— viscosity
F	— injection rate
x	— length
t	— time
l	— index of the component
$L=L(l)$	— index of the phase containing the component l.

We denote by S_l the saturation of component l defined by

$$S_l = \frac{V_l}{V_{\text{pore}}},$$

where V_l is the volume of component l and V_{pore} is the pore volume of the medium; moreover, $\hat{S}_L$ will denote the saturation of phase L:

$$\hat{S}_L = \sum_{l_L} S_l$$

where $\sum\limits_{l_L}$ denotes summation for all components belonging to phase L.

In general $\varrho_l = \varrho_l(p)$ is a function of pressure, while the relative permeabilities (taking care of the capillary effects) depend on the phase saturations:

$$k_L = k_L(\hat{S}_1, \hat{S}_2, \dots, \hat{S}_M).$$

The strong non-linearity of the problem results from the dependence of k_L on the saturations. By definition we have:

$$0 \leqq k_L \leqq 1,$$

$$k_L = 0 \quad \text{for} \quad \hat{S}_L \leqq \hat{S}_L^{\text{residual}} \quad (0 \leqq \hat{S}_L^{\text{residual}} < 1),$$

$$k_L = 1 \quad \text{for} \quad \hat{S}_L = 1.$$

The system (1), (3) consists of $m+1$ equations for the $m+1$ unknowns $p, S_1, S_2, \ldots, S_m$.

A usual way to solve numerically the system (1), (3) is to substitute equation (3) by an equation for the pressure. Then, in each time step, we compute first the pressure from this equation, and afterwards the saturations are computed by using the previously computed pressure. With the help of the identities

$$\frac{\partial(\varrho S)}{\partial t} = \frac{d\varrho}{dp}\frac{\partial p}{\partial t} S + \varrho\frac{\partial S}{\partial t},$$

$$\sum_{l=1}^{m}\frac{\partial S_l}{\partial t} = 0,$$

equations (1) can be combined to give

$$\Phi\left(\sum_{l=1}^{m}\frac{1}{\varrho_l}\frac{d\varrho_l}{dp}S_l\right)\frac{\partial p}{\partial t} = \sum_{l=1}^{m}\frac{1}{\varrho_l}\frac{\partial}{\partial x}\left(\varkappa\frac{k_L}{\mu_L}\varrho_l S_l\frac{\partial p}{\partial x}\right) + \sum_{l=1}^{m}\frac{F_l}{\varrho_l}. \tag{4}$$

This is an equation of the parabolic type for the pressure.

In order to get a more transparent study case we shall assume that Φ and $\varkappa$ are constant and that the liquids are incompressible, i.e., $\varrho_l=\text{const}$. With a proper choice of the constants the equations will be simplified to give

$$\frac{\partial S_l}{\partial t} = \frac{\partial}{\partial x}\left(\psi_L S_l\frac{\partial p}{\partial x}\right) + F_l \qquad (l = 1, 2, \ldots, m), \tag{5}$$

$$\sum_{l=1}^{m}\frac{\partial}{\partial x}\left(\psi_L S_l\frac{\partial p}{\partial x}\right) + \sum_{l=1}^{m} F_l = 0, \tag{6}$$

where

$$\psi_L = \frac{k_L}{\mu_L}.$$

Equation (6) does not contain time derivatives, the dependence of the pressure on time enters through the coefficients of this equation.

In the special case of a single-phase flow $(M=1)$ ψ_1 is constant; assuming this constant as unity, we then have

$$\frac{\partial S_l}{\partial t} = \frac{\partial}{\partial x}\left(S_l\frac{\partial p}{\partial x}\right) + F_l \quad (l = 1, 2, \ldots, m) \tag{5'}$$

$$\frac{\partial^2 p}{\partial x^2} = -\sum_{l=1}^{m} F_l. \tag{6'}$$

In those regions where injection is not present (i.e., $F_l=0$), we have

$$\frac{\partial p}{\partial x} = -v = \text{constant}$$

and (5′) is reduced to the form

$$(5'') \qquad \frac{\partial S}{\partial t} = -v\frac{\partial S}{\partial x}.$$

This is the case studied by Boris and Book.

In order to make the calculations easier for eventual manual checks, we shall assume that the relative permeability k_L for each phase L depends on the respective saturation $\hat{S}_L$ only: $k_L = k_L(\hat{S}_L)$. This does not change the basic problem arising from the non-linearity.

The numerical method

A comparison of results obtained with various possible finite difference formulations shows that the so-called up-stream weighted formulae give the best correspondence of the numerical solution of equations (1) with empirical results. Hence in oil reservoir simulations exclusively up-stream weighted versions are used, therefore we shall follow this practice.

Let us assume that the flow is oriented in the direction of the positive x axis, i.e., $\partial p/\partial x < 0$.

In our study case (5) we shall use the simplest, fully up-stream weighted and mass-conserving formula (explicit in the saturations and implicit in the pressure):

$$(7) \qquad \frac{S_j^{n+1}-S_j^n}{\Delta t} = -\frac{1}{\Delta x}(v_{j+1/2}S_j^n - v_{j-1/2}S_{j-1}^n)+F_j^n,$$

where

$$v_{j+1/2} = -\psi_{j+1/2}\frac{1}{\Delta x}(p_{j+1}^{n+1}-p_j^{n+1}),$$

$$\psi_{j+1/2} = \psi(S_j^n).$$

(Here we have dropped the component and phase indices and we are using equidistant nodes in both space and time directions. Space node points are denoted by the subscripts j and time levels by the superscripts n. $\psi(S_j^n)$ means the value of ψ calculated from saturations at the time level n at point j. Half integer indices $j \pm 1/2$ refer to the block boundaries.) We shall denote this as our "conventional" formula.

The corresponding finite difference form of the pressure equation (6) will take the form

$$(8) \qquad \frac{1}{(\Delta x)^2}[(p_{j+1}^{n+1}-p_j^{n+1})\sum(S_j\psi_{j+1/2})-(p_j^{n+1}-p_{j-1}^{n+1})\sum(S_{j-1}\psi_{j-1/2})]+\sum F_j = 0,$$

where summation refers to all components and all values — where not explicitely indicated — are taken at the time level $t=t_n$.

The conventional formula will be modified by the addition of an "anti-diffusion" term:

$$(9) \qquad \begin{aligned} S_j^{n+1}-S_j^n = -\frac{\Delta t}{\Delta x}(v_{j+1/2}S_j^n - v_{j-1/2}S_{j-1}^n)+F_j+ \\ +\mu_{j+1/2}(S_{j+1}^n-S_j^n)+\mu_{j-1/2}(S_{j-1}^n-S_j^n). \end{aligned}$$

As we shall see, μ will have a negative value, hence the name anti-diffusion. The corresponding pressure equation will have the same form (8) with the only difference that the source term $\sum F_j$ will be substituted by $\sum \tilde{F}_j$ where

$$\tilde{F}_j = F_j + \sum [\mu_{j+1/2}(S^n_{j+1} - S^n_j) + \mu_{j-1/2}(S^n_{j-1} - S^n_j)].$$

In order to find a proper corrective anti-diffusion we start with the constant velocity case (5″) when (9) will have the form

$$S^{n+1}_j = S^n_j - \varepsilon(S^n_j - S^n_{j-1}) + \mu(S^n_{j+1} - 2S^n_j + S^n_{j-1}) \tag{10}$$

with

$$\varepsilon = v\frac{\Delta t}{\Delta x}.$$

Let us develop the discret function S^n_j $(j=1, 2, ..., N)$ defined at the node points x_j in a discret (complex) Fourier-series of which the general term is

$$a^n_k e^{ikj\Delta x} \qquad \left(i = \sqrt{-1},\ j = 1, 2, ..., N\right),$$

where k is the wave number. The application of the finite difference operator appearing on the right-hand side of equation (10) transforms this function into

$$a^{n+1}_k e^{ikj\Delta x},$$

where

$$a^{n+1}_k = A(k)a^n_k,$$

$$A(k) = 1 - (\varepsilon + 2\mu)[1 - \cos(k\Delta x)] - i\varepsilon \sin(k\Delta x). \tag{11}$$

$A(k)$ is called the (complex) amplification factor.

The general term in the "ideal" Fourier-series solution of equation (5″) (describing a translatory motion with constant velocity v) is

$$a_k e^{ik(j\Delta x - vt)}$$

which corresponds to the amplification factor

$$B(k) = e^{-ikv\Delta t} = e^{-i\varepsilon k\Delta x}$$

for a time step Δt.

Evidently $|B(k)| = 1$, independently of the wave number k and time step Δt. This expresses the lack of diffusion. The same property for $A(k)$ cannot be assured with any value of μ. One possible "good" choice for μ is to select it in such a way that the difference

$$|A(k)|^2 - 1$$

be of the smallest possible order in k for $k \to 0$. By expanding the trigonometric functions in (11) into a Taylor-series at $k=0$ and equating the coefficient of k^2 to 0, we get

$$\mu = -\frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon^2. \tag{12}$$

This is the anti-diffusion we propose to correct the conventional up-stream formula. With this value of μ, the amplification factor becomes

$$A(k) = 1-\varepsilon^2[1-\cos(k\Delta x)]-i\varepsilon\sin(k\Delta x).$$

We have the identity

$$|A(k)|^2 = 1-(\varepsilon^2-\varepsilon^4)[1-\cos(k\Delta x)]^2, \tag{13}$$

wherefrom

$$|A(k)| \leqq 1 \quad \text{for} \quad \varepsilon = v\frac{\Delta t}{\Delta x} \leqq 1.$$

This is the well-known stability condition for the solution of equation (9).

We generalize formula (12) to the case of non-constant velocity by putting

$$\mu_{j+1/2} = -\frac{1}{2}\varepsilon_{j+1/2}+\frac{1}{2}\varepsilon^2_{j+1/2},$$

$$\varepsilon_{j+1/2} = v_{j+1/2}\frac{\Delta t}{\Delta x}.$$

It results from (13) that certain higher harmonics will be damped faster than the lowest harmonics (with $k\sim 0$). This fact will produce the Gibbs-phenomenon in solutions presenting sharp fronts or discontinuities. Therefore we must take into consideration the fact that the numerical solution may violate condition (2) which we want to satisfy in any case. Hence we have to perform a correction on the solution at each time step; the correction will consist of the following steps (here the time and space indices will be omitted):

i) if $S_l<0$ then we set $S_l=0$ $(l=1, 2, \ldots, m)$;

ii) if
$$\left|\sum_{l=1}^{m} S_l-1\right| \leqq \delta$$
where δ is a given small tolerance limite, then the correction is terminated;

iii) we put
$$S_l^{\text{corrected}} = S_l+S_l\cdot\Delta S_l\frac{1-\sum\limits_{l=1}^{m} S_l}{\sum\limits_{l=1}^{m} S_l\cdot\Delta S_l},$$
i.e., we distribute the error $1-\sum S_l$ according to the weights $S_l\cdot\Delta S_l$ where ΔS_l denotes the latest increase of S_l:

$$\Delta S_l = S_l^{n+1}-S_l^n.$$

Then we iterate the cycle i)—ii)—iii) starting at step i) until the correction is terminated in step ii).

This correction procedure has been introduced by us in our paper [1], where we observed that the correction procedure is quite a delicate step. E.g., the simplest idea of distributing the error uniformly among the phases introduces new sources of instability for the numerical solution.

Numerical results

In all numerical examples we shall use 60 grid points. Injection rate will be unity so that the global flow velocity will be unity, as well. We shall use a time step $\Delta t = 0.05$, i.e., saturation fronts will move — on the average — 1 grid point in 20 time steps. This is a choice within the range of most practical applications. Higher values of Δt would introduce too high errors due to the non-linearities.

a) *Single-phase flow*

First we consider the case of a single-phase flow with 3 components. Initially, the distribution S_j of the component № 2 will be

$$S_j = \begin{cases} 1 & \text{for} \quad j = 2, 3, \dots, 11, \\ 0 & \text{for} \quad j = 1, 12, 13, \dots, 60. \end{cases}$$

The component № 3 will occupy the region to the right of component № 2, i.e.,

$$S_j = \begin{cases} 1 & \text{for} \quad j = 12, 13, \dots, 60, \\ 0 & \text{for} \quad j = 1, 2, \dots, 11. \end{cases}$$

Finally, component № 1 will only be present in the grid point $j=1$ with a saturation $S_1=1$. We inject the component № 1 at the point $j=1$ with intensity 1.

We shall trace the movement of component № 2 pushed forward by component № 1 and pushing component № 3. The exact solution of the original differential problem is known: the square shape of the initial values should undergo a translatory motion with constant velocity $v=1$.

Figure 1.a shows the movement of component № 2 computed with the conventional formula without anti-diffusion. Figure 1.b presents the same movement computed with anti-diffusion but without the correction step. Finally, Figure 1.c shows the solution of the same problem with anti-diffusion and correction.

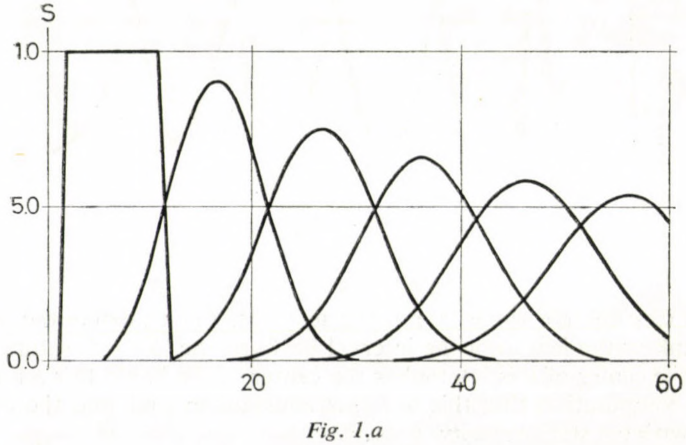


Fig. 1.a

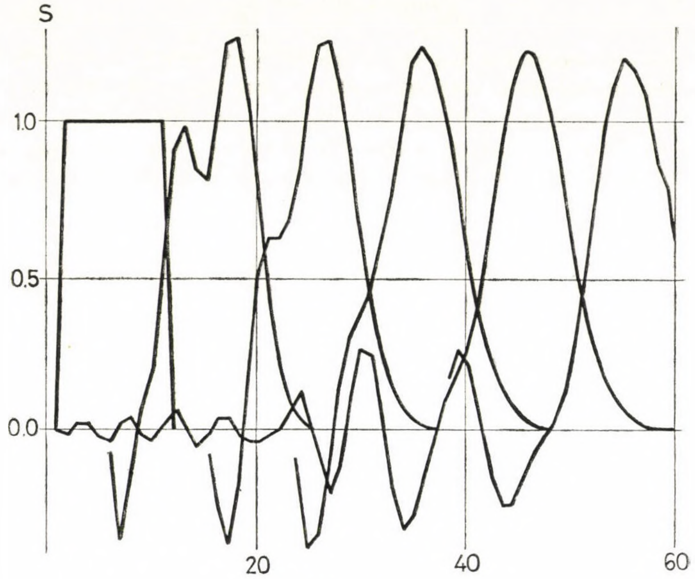


Fig. 1.b

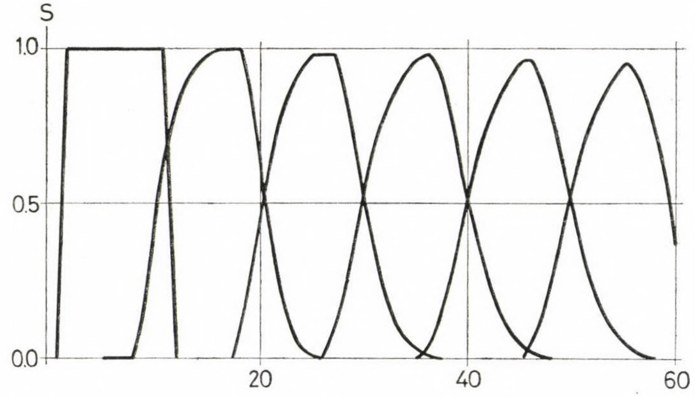


Fig. 1.c

Here we can observe some interesting features. Let us define the width of the saturation bank as the length of the interval between the $S=0.5$ saturation points; we shall call the center of this interval as the center of the bank. It is a nice empirical result of the computation that this width is constant in time and the center of the bank moves with constant velocity $v=1$.

b) *3-phase flow*

Here we suppose that each phase consists of one single component. We shall discuss two cases differing in the ratio of the viscosities. In both cases we assume that the relative permeability is described by the same function $k_L=\hat{S}_L$ for each phase L.

We shall use in both cases the same initial conditions, identical to the initial conditions discussed in the previous point a), with the only difference that the numbering № 1, № 2 and № 3 will refer to different phases instead of denoting components within a single phase. Again, the material № 1 will be injected.

In the first case the ratio of the viscosities will be chosen as

$$\mu_1:\mu_2:\mu_3 = 1:1:1,$$

while in the second case we shall use

$$\mu_1:\mu_2:\mu_3 = 100:10:1.$$

Hence, in the second case the phase fronts should be very sharp, the phases performing a piston-like motion.

Figures 2.a and 2.b show the saturation of phase № 2 without and with anti-diffusion, respectively, in the first case. Figures 3.a and 3.b represent the results in the second case, without and with anti-diffusion, as well.

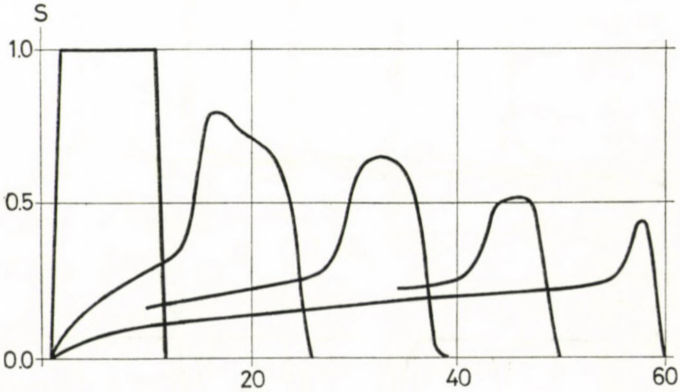


Fig. 2.a

It should be observed that in the calculation the functions $k(S)$ have been extended to the whole S line by setting

$$k(S)=\begin{cases}0 & \text{for} \quad S<0,\\ 1 & \text{for} \quad S>1.\end{cases}$$

This fact has practically eliminated negative values of the saturations. Indeed, the condition $k(S)=0$ for $S\leqq 0$ stops further decrease of S in case of an eventual occurrance of a negative value S_j. As a matter of fact, the results represented on

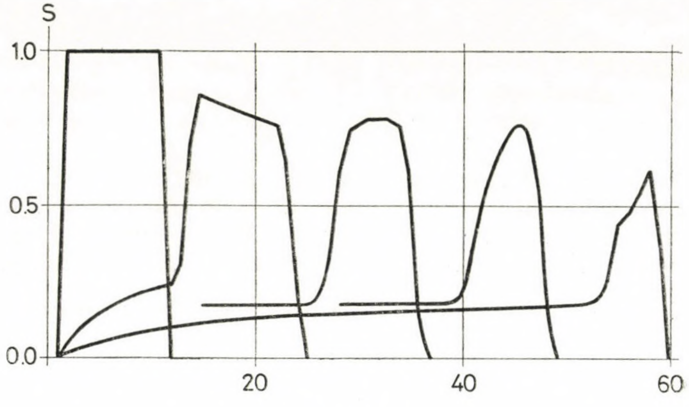


Fig. 2.b

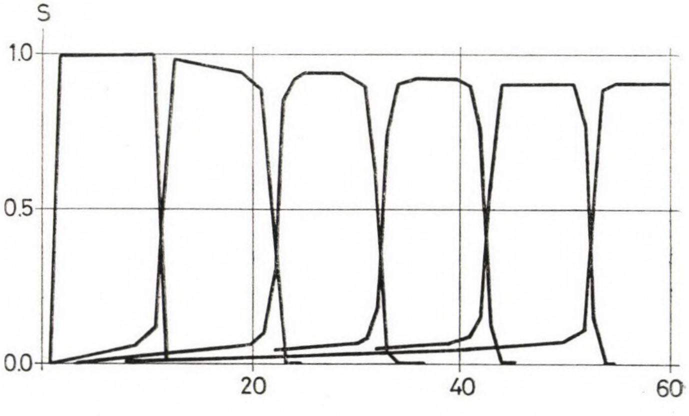


Fig. 3.a

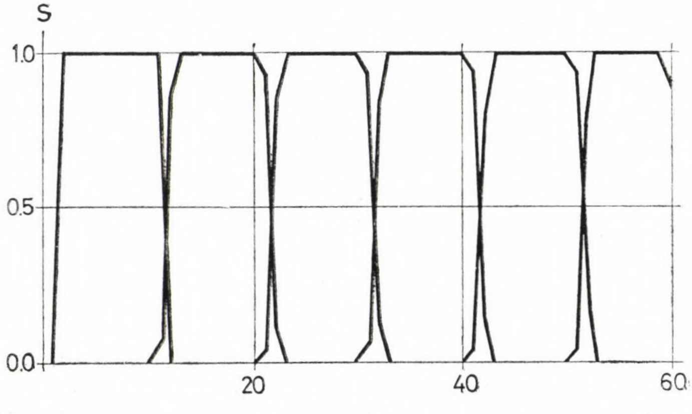


Fig. 3.b

Figures 2.a, 2.b, 3.a and 3.b have been obtained without applying the correction step. (We should observe that in the case of a single-phase flow the correction was essential in obtaining the results of Figure 1.c. Hence we conclude that in those cases where there are several components within a single phase, the correction step cannot be omitted.)

Final conclusions

We note that the inclusion of anti-diffusion in the form discussed above can be performed on already existing multi-phase flow simulators, since it only needs the change of some formulae without introducing essential changes in the computer codes. The inclusion of anti-diffusion practically does not increase the computational time.

It is possible that other finite difference formulae with corresponding anti-diffusions may give still better results than the one presented here. Hence it would be necessary to investigate systematically all possibilities.

REFERENCES

[1] ADLER, G., A Linear Model and a Related Very Stable Numerical Method for Thermal Secondary Oil Recovery, *The Journal of Canadian Petroleum Technology* **14** (1975), 56—65.

[2] BORIS, J. P.—BOOK, D. L., Solution of Continuity Equations by the Method of Flux-Corrected Transport, *Methods in Computational Physics,* Vol. 16, Academic Press, New York, 1976, 85—129.

[3] LANTZ, R. B., Quantitative Evaluation of Numerical Diffusion (Truncation Error), *Soc. Petr. Eng. Journal* (September 1971), pp. 315—320.

Mathematical Institute of the Hungarian Academy of Sciences
H—1053 Budapest, Reáltanoda u. 13—15.

(Received October 26, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 225—228.

FREE MONOIDS AND STRICT RADICALS

by

L. MÁRKI and R. MLITZ

This paper continues the investigations started in L. MÁRKI—R. MLITZ—R. STRECKER [4] and treats some more specific questions in radical theory: we determine the behaviour of free objects in monoid varieties with respect to strict radicals and describe all homomorphically closed strict semisimple classes in terms of (classical) group radical theory.

Throughout the paper we shall make heavy use of the notions, notations, and results presented in [4]. Exactly as was done there, we fix an arbitrary variety $\mathscr{V}$ of monoids which does not consist of groups only. Now we just recall the most important definition from [4]: a *strict radical* in $\mathscr{V}$ is a function ϱ which assigns to each $A\in\mathscr{V}$ a congruence ϱA on A and satisfies the following three conditions for all $A\in\mathscr{V}$:

(I) $\varphi(\varrho A)\subseteqq\varrho(\varphi A)$ for every homomorphism φ from A,

(II) $\varrho(A/\varrho A)=\omega$,

(III) $\varrho A=\omega$ iff $\forall B\leqq A$ $(B\neq\{1\}\Rightarrow\varrho B\neq\iota)$.

According to a result of A. L. ŠMEL'KIN [5], for any radical in the variety of all groups, there exists a cardinal number $\varkappa$ such that G_X is radical if $|X|<\varkappa$ and semisimple if $|X|\geqq\varkappa$, where G_X denotes the free group over the set X; furthermore, for every cardinal number $\varkappa$ there exists a radical for which just $\varkappa$ is the "switching point". Let us see now the situation in varieties of monoids.

THEOREM 1. *Let $\mathscr{V}$ be a variety of monoids which is not a variety of groups, and let ϱ be a strict radical in $\mathscr{V}$. Then three cases are possible for ϱ:*

a) *every free monoid in $\mathscr{V}$ is radical;*

b) *every free monoid in $\mathscr{V}$ is semisimple;*

c) *every free monoid in $\mathscr{V}$ is neither radical nor semisimple.*

Case a) *takes place only for the whole $\mathscr{V}$ being the radical class for ϱ. Case* b) *takes place if and only if* a) *is not the case and either the free cyclic monoid F_1 in $\mathscr{V}$ is infinite or F_1 is finite and the two-element semilattice L is semisimple for ϱ.*

PROOF. If one of the free monoids is radical then so are they all since strict radical classes of monoids are closed under homomorphic images and free products (see [4], Corollary 3.10). Clearly, this takes place only in the trivial case when all monoids in $\mathscr{V}$ are radical. If some free monoid is semisimple then by heredity (see [4], Corollary 3.2) so is F_1. Conversely, suppose that F_1 in $\mathscr{V}$ is semisimple and let F_X be the free monoid over an arbitrary set X. We are going to prove that F_X is also semisimple. Since $\mathscr{V}$ is not a variety of groups, no identity of the form $x^n=1$ can hold in $\mathscr{V}$. This implies that no identity of the form $P(x_1, \ldots, x_k)=1$ can hold either (where P is an arbitrary, say k-ary, term), thus F_X is a semigroup with an "external" identity 1 adjoined. (For the identities which hold in $\mathcal{V}$ this means that they contain the same variables on both sides.) For any $x\in X$, the mapping $x\mapsto x$, $y\mapsto 1$ $(y\in X\setminus\{x\})$ extends to an epimorphism $\varphi_x\colon F_X\to F_1$, the kernel of which will be denoted by Θ_x. Since F_1 is semisimple, for the radical congruence ϱF_X of F_X we have $\varrho F_X\subseteqq \bigwedge_{x\in X}\Theta_x$. Consider now an arbitrary element $w\neq 1$ of F_X and let x be a letter occurring in w. Then $\varphi_x(w)\neq 1$, whence $w\notin [1]_{\varrho F_X}$. Thus we have $[1]_{\varrho F_X}=\{1\}$, i.e., F_X is semisimple.

We still have to determine when case b) takes place. Since we always have $J_1=\{1\}$ in F_1, where J_1 denotes the class of 1 for Green's relation $\mathcal{J}$, by Example 1 in [4] F_1 is certainly semisimple if so is L. If F_1 is finite (i.e., $\mathcal{V}$ has an identity of the form $x^k=x^n$, $1\leqq k<n$) then F_1 must have an idempotent other than 1 and thus it contains L as a submonoid. Therefore F_1 is not semisimple if L is radical. Consider, finally, the case when F_1 is infinite (i.e., in the identities which hold in $\mathcal{V}$, each variable occurs at the same number of times on both sides) and denote by R the largest radical submonoid in F_1 (this exists by [4], Corollary 3.9). Since F_1 is commutative, by Theorem 7.5 in [4] R is a normal submonoid of F_1 (which means that R is the inverse image of 1 under some homomorphism of F_1). Let x be the generating element of F_1 and m be the smallest natural number for which $x^m\in R$. (If $R=\{1\}$ then put $m=0$.) In case $m\neq 0$, let $x^n\in R$, $n=km+p$, $0\leqq p<m$. Since R is a normal submonoid, this implies $x^p\in R$, whence $p=0$, i.e., R is the submonoid generated by x^m. This means, however, that either $R=\{1\}$ and F_1 is semisimple, or $F_1\cong R$ and F_1 is radical.

THEOREM 2. *Let $\mathcal{V}$ be a variety of monoids which is not a variety of groups. $\mathcal{V}$ admits no homomorphically closed strict semisimple classes except the trivial class and $\mathcal{V}$ itself, if and only if F_1 is either infinite or defined by $x^{1+l}=x^l$ for some l. If F_1 is defined by $x^{d+l}=x^l$ where $d>1$ then the homomorphically closed strict semisimple classes $\neq\mathcal{V}$ in $\mathcal{V}$ are exactly the homomorphically closed semisimple classes (in the classical group theoretical sense) in the subvariety defined by $x^d=1$ in $\mathcal{V}$.*

PROOF. Let S be a homomorphically closed strict semisimple class in $\mathcal{V}$. If either F_1 is infinite or F_1 is finite but $L\in$S then all free monoids in $\mathcal{V}$ belong to S by Theorem 1, hence S, being homomorphically closed, equals $\mathcal{V}$. If this is not the case then $L\in\mathcal{R}$S whence the monoids in S have no idempotent elements other than 1. On the other hand, since F_1 is finite, every element of an arbitrary monoid in S has an idempotent power which therefore equals 1. By this we have seen that S consists of groups only.

In what follows we may restrict ourselves to the case that $F_1\in\mathcal{V}$ is finite and is defined by the equation $x^{d+l}=x^l$, and that S is a class (hence a variety) of groups. If $d=1$ then $\mathcal{V}$ contains no non-trivial groups and therefore S is the trivial class. So we may assume that $d>1$. Since S$\subseteqq\mathcal{V}$ is a class of groups, $x^d=1$ holds in S.

Let $\mathcal{G}$ be the subvariety of $\mathcal{V}$ defined by $x^d=1$. Clearly, $\mathcal{G}$ is a variety of groups and S$\subseteqq\mathcal{G}\subseteqq\mathcal{V}$. By Theorem 3.5 in [4], S is a homomorphically closed strict semisimple class in $\mathcal{G}$. By TRẦN VÂN HẠO [6], a class S′ in a variety of groups is a semisimple class (in the classical sense) if and only if S′ is hereditary with respect to normal subgroups and closed under subdirect products and extensions. Since S is a variety, it obviously satisfies the first two properties, and the third holds by Theorem 3.11 in [4].

Conversely, suppose that $\mathbf{S}$ is a homomorphically closed semisimple class in $\mathcal{G}$ (in the classical sense). Firstly we prove that it is also a strict semisimple class in $\mathcal{G}$. Suppose that $A\in\mathcal{G}$ is such that $(\forall B\leqq A)\ \exists\varphi B\neq E, \varphi B\in\mathbf{S}$. Then this holds in particular for all normal subgroups of A, whence $A\in\mathbf{S}$. Conversely, $\mathbf{S}$ is a homomorphically and subdirectly closed class, whence it is a variety by S. R. KOGALOVSKIĬ's well-known theorem [2] (see also [1], Ch. III, § 3, Th. 3). In particular, it is hereditary with respect to subgroups (or equivalently, submonoids). Thus $\mathbf{S}$ is a strict semisimple class in $\mathcal{G}$ by Theorem 3.1 in [4].

Let $\bar{\mathbf{S}}$ be the strict semisimple class generated by $\mathbf{S}$ in $\mathcal{V}$, then $\mathbf{S}=\bar{\mathbf{S}}\cap\mathcal{G}$ by Theorem 3.5 in [4]. Note that $\bar{\mathbf{S}}$ consists of groups only. In fact, if $A\in\bar{\mathbf{S}}$ is not a group then, since A is periodic, there exists an idempotent element $e\neq 1$ in A, but then $L\cong\{1, e\}\leqq A$, $L\in\mathcal{R}\mathbf{S}$, a contradiction to [4], Proposition 3.3. But this means that $\bar{\mathbf{S}}\subseteqq\mathcal{G}$, thus $\mathbf{S}=\bar{\mathbf{S}}$. By this we have proved that every homomorphically closed semisimple class (in the classical sense) in $\mathcal{G}$ is a (homomorphically closed) strict semisimple class in $\mathcal{V}$.

Finally, such a class always exists since $\mathcal{G}$ is clearly a homomorphically closed semisimple class in $\mathcal{G}$ and a proper subclass of $\mathcal{V}$.

By the proof of Theorem 2 we have

COROLLARY. *The homomorphically closed strict semisimple classes in $\mathcal{V}$ are exactly those subvarieties which consist of groups only and are closed under (group) extensions (within $\mathcal{V}$).*

Now we give some indications how to find such subvarieties, though we do not determine all of them explicitly. In doing this, we adopt the notations of Theorem 2.

Let $\mathbf{S}$ be an extension closed group variety in $\mathcal{V}$. Let r be the smallest natural number such that $x^r=1$ is valid in $\mathbf{S}$, then we have $d=rs$ for some natural number s and the free cyclic group in $\mathbf{S}$ is just the cyclic group $C(r)$.

If $r=1$ then $\mathbf{S}$ is the trivial class, so we may henceforth assume that $r>1$.

Suppose next that $(r, s)\neq 1$, then there exist a prime number p and a natural number m such that $p^m|r, p^{m+1}\nmid r, p^{m+1}|d$. Let G_1 be the free cyclic group in $\mathcal{G}$, i.e., $G_1=C(d)$. Since $p^{m+1}|d, C(p^{m+1})\leqq C(d)$, hence $C(p^{m+1})\in\mathcal{G}$. Similarly, we have that $C(p^m)\in\mathbf{S}$ and $C(p)\in\mathbf{S}$. Finally, $p^{m+1}\nmid r$ implies that $C(p^{m+1})\notin\mathbf{S}$, contrary to the extension property of $\mathbf{S}$. Hence for a semisimple class $\mathbf{S}$ $(r, s)\neq 1$ cannot occur.

Suppose now $(r, s)=1, r>1$, and denote by $\mathbf{S}_r$ the subvariety of $\mathcal{G}$ defined by $x^r=1$. We claim that $\mathbf{S}_r$ is a homomorphically closed strict semisimple class in $\mathcal{V}$. In fact, by the Corollary above it suffices to show that $\mathbf{S}_r$ is closed under (group) extensions, but this is clearly true, since if B is a normal subgroup of $A\in\mathcal{G}$ and $B, A/B\in\mathbf{S}_r$ then $A\in\mathbf{S}_r$ must hold in view of $(r, s)=1$.

REMARK. Homomorphically closed semisimple classes of rings are known to be radical classes as well (see R. WIEGANDT [7], Theorem 32.1). This is not the case for strict monoid radicals. In fact, let G be a non-commutative simple group of order $d=p^m s$, p a prime number, $m\geqq 1, (p, s)=1$. Let $\mathcal{V}$ and $\mathbf{S}$ be the monoid varieties defined by $x^{d+2}=x^2$ and $x^{p^m}=1$, respectively. In view of the preceding considerations, $\mathbf{S}$ is a homomorphically closed strict semisimple class in $\mathcal{V}$. We have $G\in\mathcal{V}\setminus\mathbf{S}$, and by the simplicity of G and by $p|d$ there exist at least two

p-Sylow subgroups H_1, H_2 in G. If **S** were a strict radical class, then there should exist a p-subgroup of G containing both H_1 and H_2, but this is impossible.

Recall now that the classical examples for ring radicals (Jacobson, Baer, Koethe, Levitzki, Brown—McCoy) have hereditary radical classes. This is again an instance where strict monoid radicals show up a quite different behaviour. Examples 1 and 3 in [4] show strict radicals with non-hereditary radical classes. Now we shall see that much more can be claimed, at least in sufficiently large varieties $\mathscr{V}$.

THEOREM 3. *Suppose that $\mathscr{V}$ is a variety of monoids which is not a variety of groups, and that every $A\in\mathscr{V}$ can be embedded into a congruence-free monoid in $\mathscr{V}$. Let* **R** *be a non-trivial strict radical class which is different from $\mathscr{V}$, too. Then* **R** *is not hereditary.*

PROOF. Let $A\in\mathscr{V}\setminus\mathbf{R}$ and consider a $B\in\mathbf{R}$, $B\neq E$. Let C be a congruence-free monoid containing the direct product $A\times B$. Now $B\leqq C$ and $B\in\mathbf{R}$, hence C is not semisimple. Then it must be radical in view of its congruence-freeness. Thus $C\in\mathbf{R}$, $A\leqq C$, $A\notin\mathbf{R}$, whence **R** is not hereditary.

REMARK. The condition imposed on $\mathscr{V}$ in Theorem 3 is satisfied if $\mathscr{V}$ is, e.g., the variety of all monoids (see e.g. J. KOLLÁR [3]) but not if $\mathscr{V}$ consists of all commutative monoids. In the latter case the assertion does not hold either:

EXAMPLE. Let $\mathscr{V}$ be the variety of commutative monoids and **R** be the class of abelian torsion groups considered as monoids. We claim that **R** is a hereditary strict radical class in $\mathscr{V}$. In fact, any submonoid of a torsion group is a subgroup, whence itself a torsion group. Thus it remains to show that **R** is a strict radical class in $\mathscr{V}$. We do this by using Corollary 7.6 in [4]. Note that for any $A\in\mathscr{V}$, A has a largest submonoid $\mathbf{R}(A)$ belonging to **R**, this $\mathbf{R}(A)$ is $\{a\in A\mid a^n=1$ for some natural number $n\geqq 1\}$. Now it is clear that **R** is homomorphically closed and that for all $A\in\mathscr{V}$ and $B\leqq A$, $B\in\mathbf{R}$ and $A/\langle B\rangle_A\in\mathbf{R}$ imply $A\in\mathbf{R}$, further $\mathbf{R}(A)$ is a normal submonoid in A, since $g, h\in\mathbf{R}(A)$, $x\in A$ and $xg=h$ imply $x=hg^{-1}\in\mathbf{R}(A)$. Thus Corollary 7.6 in [4] applies, yielding that **R** is a strict radical class.

REFERENCES

[1] GRÄTZER, G., *Universal Algebra*, Van Nostrand, Princeton, 1968.

[2] Когаловский, С. Р., Структурные характеристики универсальных классов, *Сибирск. Мат. Ж.* **4** (1963), 97—119.

[3] KOLLÁR, J., Interpolation in semigroups, *Semigroup Forum* **17** (1979), 337—350.

[4] MÁRKI, L.—MLITZ, R.—STRECKER, R., Strict radicals of monoids, *Semigroup Forum* **21** (1980), 27—66.

[5] Шмелькин, А. Л., Одно свойство полупростых классов групп, *Сибирск. Мат. Ж.* **3** (1962), 950—951.

[6] Чан Ван Хао, О полупростых классах групп, *Сибирск. Мат. Ж.* **3** (1962), 943—949.

[7] WIEGANDT, R., Radical and semisimple classes of rings, *Queen's papers in pure and applied math.*, 37, Queen's University, Kingston, Ontario, 1974.

Mathematical Institute of the Hungarian Academy of Sciences
H—1053 Budapest, Reáltanoda u. 13—15

and

Institut für Angewandte Mathematik, TU Wien
A—1040 Wien, Gusshausstrasse 27—29

(Received November 30, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 229—243.

ON A NEW INDEX FOR CHARACTERISING THE VERTICES OF CERTAIN NON-BIPARTITE GRAPHS

by

R. B. MALLION and D. H. ROUVRAY

Abstract

We present a new, physically inspired, graph-theoretical index which is a characteristic of the vertices of certain non-bipartite graphs. The index is called the 'charge on vertex r' and is given the symbol q_r. Following arguments similar to those of COULSON and RUSHBROOKE [1], we show that, for those graphs to which the index is applicable, $q_r=1$ for all vertices of bipartite graphs, whereas for non-bipartite graphs q_r is, in general, not unity. In both cases, however,

$$\sum_{r=1}^{N} q_r = N,$$

where N is the number of vertices in the graph. Complications which arise when the spectrum of the graph contains repeated eigenvalues are examined and one situation is identified in which the index q_r cannot be defined. Numerical values are given of computed q_r-values for various graphs typical of those commonly encountered.

Introduction

There are very few graph-theoretical parameters which may be used to characterise any vertex, r, of a given graph; an example of such an index is the *degree* of a vertex. We therefore introduce here a quantity, associated with a well-known physical property of molecules, which may have applications within the framework of pure graph theory. The index we define arises from our studies [2] of the 'Pairing theorem' of COULSON and RUSHBROOKE [1] as used by chemists in their study of unsaturated hydrocarbon molecules by simple molecular-orbital| theory [3, 4]. This index is analogous to what the chemist defines as the 'charge on carbon atom r' of a conjugated hydrocarbon species [3, 4]; we therefore propose to call it 'the *charge* on vertex r' of a graph, and to give it the symbol q_r. Our studies have shown that the indices q_r may be associated with graphs which represent molecules well-known in chemistry. It is also found that q_r can be defined for certain graphs which do not, in fact, represent actual molecules [5], although application of the index is not universal, as will be seen.

Preliminaries

In the simple case we consider first, we suppose that a graph, G, has N vertices, and that the $N\times N$ vertex adjacency-matrix, $(\mathbf{A}(G))$, of G has been diagonalised to yield a spectrum of N eigenvalues, denoted by $\{\lambda_i\}$. These eigenvalues are labelled λ_1 to λ_N, *from* the *largest to* the *smallest; subsequent reference to a 'high' or 'low' eigenvalue will refer to its position in this sequence.* There is also a corresponding set of N eigenvectors, $\{\psi_i\}$, the i-th member of which will have the components $c_{i1}, c_{i2}, \ldots, c_{iN}$. These eigenvectors are assumed to be normalised — *i.e.*

$$\sum_{r=1}^{N} c_{ir}^2 = 1. \tag{1}$$

We can now define the index q_r in either of two ways, depending on whether or not the spectrum of G contains repeated eigenvalues. As will be seen later (§ II A), our initial definition (§ I A) is a special case of the more general definition given in § II A. It is, however, pedagogically instructive to treat the case involving distinct eigenvalues first before considering the more general definition of q_r.

Case I. The family of eigenvalues of G, $\{\lambda_i\}$, contains no multiple members (*i.e.* all λ_i are distinct)

A. The Definition of q_r

The vertex adjacency-matrix of G is always a real, symmetric matrix and, because we are assuming here that all the λ_i are distinct, it follows that all the ψ_i will automatically be mutually orthogonal. This requirement, together with the normality condition (1), leads to the general relationship:

$$\sum_{r=1}^{N} c_{ir} c_{jr} = \delta_{ij} \tag{2}$$

for any pair of eigenvectors, ψ_i and ψ_j.

Under these circumstances, the index q_r may be defined by equations (3) and (4):

$$q_r = 2 \sum_{i=1}^{\frac{1}{2}N} c_{ir}^2 \quad \text{(for } N \text{ even)}, \tag{3}$$

$$q_r = 2\left(\sum_{i=1}^{\frac{1}{2}(N-1)} c_{ir}^2 \right) + c_{\frac{1}{2}(N+1),r}^2 \quad \text{(for } N \text{ odd)}. \tag{4}$$

Using arguments similar to those employed by Coulson and Rushbrooke [1], we shall show later (§ I C) that, for each individual vertex, r, of a bipartite graph, G, $q_r = 1$, whereas if G is only three-colourable, q_r is, in general, not unity. First, however, we prove a simple theorem which places a constraint on the *sum* of q_r-values in any given graph — bipartite or non-bipartite — for which the index q_r can be defined.

Theorem 1. *For any graph, G, for which the index q_r can be defined, (see § II C),*

$$\sum_{r=1}^{N} q_r = N. \tag{5}$$

Proof. This follows easily from the normalisation condition (1). For N even we obtain from (1) and (3) the result

$$\sum_{r=1}^{N} q_r = 2 \sum_{r=1}^{N} \sum_{i=1}^{\frac{1}{2}N} c_{ir}^2 = 2\left(\frac{N}{2}\right) = N. \tag{6}$$

For N odd we obtain from (1) and (4)

$$\sum_{r=1}^{N} q_r = \sum_{r=1}^{N} \left\{2\left(\sum_{i=1}^{\frac{1}{2}(N-1)} c_{ir}^2\right) + c^2_{\frac{1}{2}(N+1),r}\right\} = 2\{\tfrac{1}{2}(N-1)\}+1 = N. \tag{7}$$

B. The Pairing Theorem

What we are here calling the 'Pairing theorem' states that any bipartite graph will be possessed of complementary eigenvalues — *i.e.* if λ_i be an eigenvalue of the vertex adjacency-matrix of the graph in question, then $-\lambda_i$ $(=\lambda_{N-i+1}$ in the eigenvalue numbering scheme we devised in *Preliminaries*) will also be an eigenvalue.

This theorem was first put forward (in the *chemical* literature) by Coulson and Rushbrooke in 1940 [1]. Earlier work by Perron [6, 7] and Frobenius [7—10] had established that if the highest eigenvalue of an irreducible matrix having non-negative elements were paired then all the other elements would also be paired; much later this same result was obtained more concisely by Wielandt [11]. Since that time, the theorem has been proved (or, at least, implied) in a number of different contexts [12—16], the various authors concerned apparently being unaware of the existence of the Coulson—Rushbrooke theorem [17]. Recent proofs of the theorem in the chemical literature [18, 19], as well as extensions of it to certain vertex-weighted bipartite graphs [20], have been based on the approach of Sachs [13]. Yet another way of proving this same result was presented by one of the present authors [21]. A corollary to this theorem that we shall have occasion to use in the present discussion concerns the eigenvectors associated with these complementary pairs of eigenvalues of a bipartite graph, G (a corollary which has very recently been generalised to certain vertex-weighted, bipartite graphs [22]). Let the vertices of one set be labelled from 1 to p, and let the vertices of the other set be labelled from $p+1$ to N. The adjacency matrix of G will then be partitioned in the following way (*e.g.* [21]):

$$\begin{array}{c|c:c|} & \leftarrow p \rightarrow & \leftarrow N-p \rightarrow \\ \updownarrow p & \mathbf{0} & \mathbf{B} \\ \hdashline \updownarrow N-p & \mathbf{B}^{\top} & \mathbf{0} \end{array}.$$

Then if an eigenvalue λ_i of the bipartite graph G has the associated eigenvector ψ_i, where

$$\psi_i = \begin{pmatrix} c_{i1} \\ c_{i2} \\ \vdots \\ c_{ip} \\ c_{i,p+1} \\ c_{i,p+2} \\ \vdots \\ c_{iN} \end{pmatrix}, \tag{8}$$

the complementary eigenvalue, $-\lambda_i$, of G will have the associated eigenvector ψ'_i, where

$$\psi'_i = \begin{pmatrix} c_{i1} \\ c_{i2} \\ \vdots \\ c_{ip} \\ -c_{i,p+1} \\ -c_{i,p+2} \\ \vdots \\ -c_{iN} \end{pmatrix} \quad (= \psi_{N-i+1}). \tag{9}$$

(If the eigenvalues are labelled in sequence — in this case, from the highest to the lowest — then the eigenvalue complementary to the one labelled 'λ_i' will be labelled 'λ_{N-i+1}'.) This can be summarised conveniently as follows:

$$c_{N-i+1,j} = c_{ij} \quad (j \leqq p); \quad c_{N-i+1,j} = -c_{ij} \quad (j > p). \tag{10}$$

Both the pairing of the eigenvalues, and the close relation between the corresponding eigenvectors, is a consequence [21, 22] of the way (illustrated above) in which the adjacency matrix of a bipartite graph may be partitioned.

C. The Numerical Value of q_r in a Bipartite Graph

We are now in a position to discuss some properties of the newly defined, graph-theoretical index q_r. For simplicity, we give explicit proof only for the case in which all eigenvalues of G are distinct.

Theorem 2. *In any bipartite graph, G, for which the index q_r can be defined (see § II C), the value of q_r at each vertex r is precisely unity.*

Proof. Let $\mathbf{C}$ be a matrix whose columns are the N eigenvectors of $\mathbf{A}(G)$, the $N\times N$ vertex adjacency-matrix of G. By virtue of condition (2), this matrix will be orthogonal, and hence

$$\mathbf{C}^\top \mathbf{C} = \mathbf{1}_{N\times N}, \tag{11}$$

where $\mathbf{C}^\top$ is the transpose of $\mathbf{C}$. Since, by definition,

$$\mathbf{C}^{-1}\mathbf{C} = \mathbf{1}_{N\times N} = \mathbf{C}\mathbf{C}^{-1}, \tag{12}$$

where $\mathbf{C}^{-1}$ is the inverse of $\mathbf{C}$, it follows that

$$\mathbf{C}^\top = \mathbf{C}^{-1} \tag{13}$$

and hence, from the right-hand side of (12), that

$$\mathbf{C}\mathbf{C}^\top = \mathbf{1}_{N\times N}. \tag{14}$$

By equating diagonal elements on either side of equation (14), we obtain

$$\sum_{i=1}^{N} c_{ir}^2 = 1. \tag{15}$$

We now use this result to prove Theorem 2 for even and odd bipartite graphs of which all the eigenvalues are distinct.

(a) *N even:* If N is even, q_r is given by equation (3). If $1 \leqq r \leqq p$,

$$c_{ir} = c_{N-i+1,r} \tag{16}$$

from equation (10). Equation (3) may thus be partitioned into two terms, as follows:

$$q_r = \sum_{i=1}^{\frac{1}{2}N} c_{ir}^2 + \sum_{i=1}^{\frac{1}{2}N} c_{ir}^2 = \sum_{i=1}^{\frac{1}{2}N} c_{ir}^2 + \sum_{i=\frac{1}{2}N+1}^{N} c_{ir}^2. \tag{17}$$

These two equations can be recombined, and equation (15) called upon, to give

$$q_r = \sum_{i=1}^{N} c_{ir}^2 = 1. \tag{18}$$

If $p+1 \leqq r \leqq N$, equation (10) dictates that (16) is replaced by

$$c_{ir} = -c_{N-i+1,r} \tag{19}$$

and equation (17) then becomes

$$q_r = \sum_{i=1}^{\frac{1}{2}N} c_{ir}^2 + \sum_{i=\frac{1}{2}N+1}^{N} (-c_{ir})^2, \tag{20}$$

which again leads to (18).

(b) *N odd:* If N is odd, q_r is given by equation (4), and it follows from the complementary nature of the eigenvalues of a bipartite graph that at least one λ_i, namely $\lambda_{\frac{1}{2}(N+1)}$, must be zero. The eigenvalue $\lambda_{\frac{1}{2}(N+1)}$ is its own complement in such a case. The genuinely complementary eigenvalues are then λ_1 to $\lambda_{\frac{1}{2}(N-1)}$, and λ_N to $\lambda_{\frac{1}{2}(N+3)}$. By use of arguments entirely analogous to those outlined between equations (15) and (20), equation (4) can be written

$$q_r = \sum_{i=1}^{\frac{1}{2}(N-1)} c_{ir}^2 + \left(\sum_{i=\frac{1}{2}(N+3)}^{N} c_{ir}^2 \right) + c_{\frac{1}{2}(N+1),r}^2 = \sum_{i=1}^{N} c_{ir}^2 = 1 \tag{21}$$

for all r, $1 \leqq r \leqq N$, the latter step again being *via* equation (15). Theorem 2 is, therefore, established.

If G is not bipartite its vertex adjacency-matrix cannot be partitioned by the colouring process as in the bipartite case and hence, in general, no complementary pairs of eigenvalues (and, in particular, no related pairs of eigenvectors as in equations (8) and (9)) will exist. In such cases it is therefore impossible to break down equations (3) and (4) as in equations (17) and (21) since, in general,

$$2 \sum_{i=1}^{\frac{1}{2}N} c_{ir}^2 \neq \sum_{i=1}^{N} c_{ir}^2 \quad \text{(for even } N\text{)} \tag{22}$$

and

$$2 \sum_{i=1}^{\frac{1}{2}(N-1)} c_{ir}^2 \neq \sum_{i=1}^{\frac{1}{2}(N-1)} c_{ir}^2 + \sum_{i=\frac{1}{2}(N+3)}^{N} c_{ir}^2 \quad \text{(for odd } N\text{).} \tag{23}$$

It thus follows that, in general, q_r will not equal unity.

q_r may thus be regarded as an index which is a characteristic of the vertices of *non*-bipartite graphs in that it will vary from one vertex to another within such graphs. The variation in q_r will, however, always be subject to what we term (again following our physical interpretation) the 'conservation of charge'; this constraint is implied in equation (5). The idea of charge is a familiar one in molecular-orbital theory [1, 4, 23, 24] and, from studies of the charges on atoms in molecules, it is known that the definition of q_r given above is not completely general. We therefore now consider a more general definition of q_r.

Case II. The spectrum of G contains degeneracies (*i.e.* repeated eigenvalues)

A. Definition of Eigenvalue Occupation-Numbers and More General Definition of q_r

When multiple eigenvalues (which we shall henceforth refer to as 'degeneracies') occur in the spectrum of a graph, it is possible to invoke the analogy of a principle known in physics and chemistry as the *Aufbau* Principle, which involves the use of Hund's Rules of Maximum Multiplicity [25]. The chemical applications of this Principle have a sound physical interpretation; for our purposes here, however, we treat the Principle merely as a prescription for assigning the appropriate contribution of each eigenvalue, λ_i, (*via* the r-th component, c_{ir}, of its associated eigenvector) to the index q_r.

Each eigenvalue, λ_i, is first assigned an *occupation number*, v_i, which may take only the values 0, 1 or 2. By analogy with the physical terminology, we shall describe the i-th eigenvalue, λ_i, as 'doubly occupied' (or 'filled') if $v_i=2$, 'singly occupied' (or 'half-filled') if $v_i=1$, and 'unoccupied' (or 'empty') if $v_i=0$. Reference to an 'occupied' eigenvalue thus implies one for which $v_i=1$ *or* $v_i=2$. This assignment of values of 0, 1 or 2 to the $\{v_i\}_{i=1,2\dots N}$ is always subject to the condition that

$$\sum_{i=1}^{N} v_i = N. \tag{24}$$

The following rules are adopted when assigning v_i-values to the various λ_i:

1. If N is odd, let a and b be the smallest and largest values of i, respectively, for which $\lambda_i=\lambda_{\frac{1}{2}(N+1)}$.
2. If N is even and $\lambda_{\frac{1}{2}N}=\lambda_{\frac{1}{2}N+1}$, let a and b be the smallest and largest values of i, respectively, for which $\lambda_i=\lambda_{\frac{1}{2}N}$.
3. If N is even and $\lambda_{\frac{1}{2}N}>\lambda_{\frac{1}{2}N+1}$, let $a=\frac{1}{2}N+1, b=\frac{1}{2}N$.
4. If $a+b=N+1$, we define $v_1, v_2, \dots, v_N$ by the rule that $v_i=2$ if $i<a$, $v_i=1$ if $a\leq i\leq b$ and $v_i=0$ if $i>b$.
5. If $a+b\neq N+1$, a difficulty arises, as is discussed further in § II C, and $v_1, v_2, \dots, v_N$ (and, hence, q_r) must be left undefined.

Our most general definition of q_r is then as follows. Let $\mathbf{A}(G)$ be the adjacency matrix of G (with respect to a fixed ordering of the vertices) and let $\psi_1, \psi_2, \dots, \psi_N$

be mutually *orthogonal** eigenvectors of $\mathbf{A}(G)$ corresponding to eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_N$, respectively, these eigenvectors being normalised so that $|\psi_1|=|\psi_2|=\dots=|\psi_N|=1$. Then:

$$q_r = \sum_{i=1}^{N} v_i c_{ir}^2, \tag{25}$$

where c_{ir} denotes the r-th component of ψ_i. It will be immediately evident that when no repeated eigenvalues occur this definition reduces to our previous definitions given by equation (3) (for even N), and by equation (4) (for odd N). It is convenient to draw further on our physical analogy and represent eigenvalue occupation-numbers pictorially by first drawing an ordinate representing, in a qualitative way, the direction of increasing number (*i.e.* decreasing size) of eigenvalue. Repeated (degenerate) eigenvalues are placed at the same level in the pattern. One vertical stroke placed on such an eigenvalue 'level' then signifies a v_i-value of 1, two strokes signify a v_i-value of 2, and unoccupied eigenvalues ($v_i=0$) are left blank. With this convention, the equivalence between definitions (25) and ((3) and (4)), in the non-degenerate case, may be confirmed by reference to Figs. 1 and 2.

This new definition of q_r may be seen to be equivalent to the definitions contained within equations (3) and (4) even when degeneracies are present, *provided that* these degeneracies do not occur amongst the uppermost occupied eigenvalues. Examples are given in Figs. 3 and 4. We necessarily assume again that appropriate

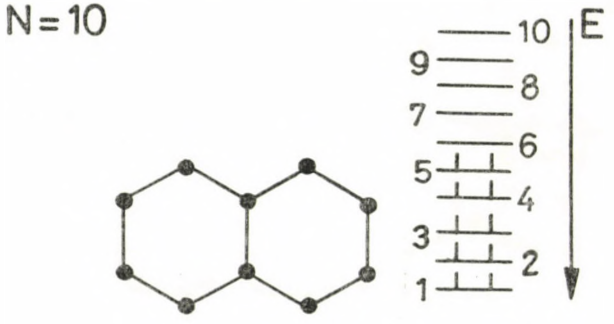


Fig. 1

* From the well-known properties of real-symmetric matrices, the eigenvectors belonging to distinct eigenvalues will automatically be mutually orthogonal. If, however, the i-th eigenvalue (say) is m-fold degenerate (*i.e.*, occurs m times as a root of the characteristic polynomial of $\mathbf{A}(G)$), there will be m linearly independent eigenvectors, $\psi_i^{(1)}, \psi_i^{(2)}, \dots, \psi_i^{(r)}, \dots, \psi_i^{(m)}$, all giving rise to the same eigenvalue, λ_i, but these will not, in general, be mutually orthogonal. Each of these eigenvectors will, of course, be orthogonal to every eigenvector belonging to all the other eigenvalues, λ_r when $r \neq i$; but, in general, the set of eigenvectors $\{\psi_i^{(r)}\}$ will not satisfy equation (2). However, we can *always choose* a new set of eigenvectors, $\psi_i^{o(1)}, \psi_i^{o(2)}, \dots, \psi_i^{o(r)}, \dots, \psi_i^{o(m)}$, by taking different linear combinations of the $\{\psi_i^{(r)}\}$, which *will* be mutually orthogonal. Such a new set will, after normalisation, always satisfy equation (2), even when the spectrum of G contains repeated eigenvalues. Accordingly, in our most general definition of q_r, given in equation (25), it is implicitly assumed that this process of orthogonalisation has already been carried out.

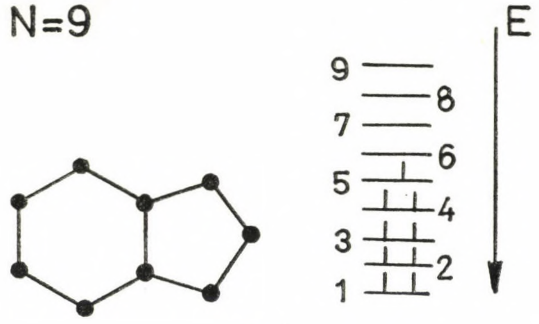


Fig. 2

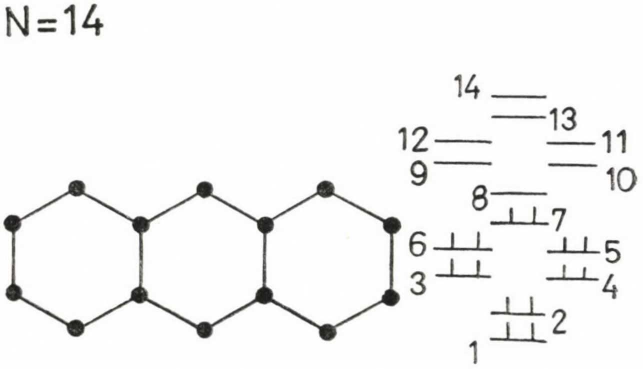


Fig. 3

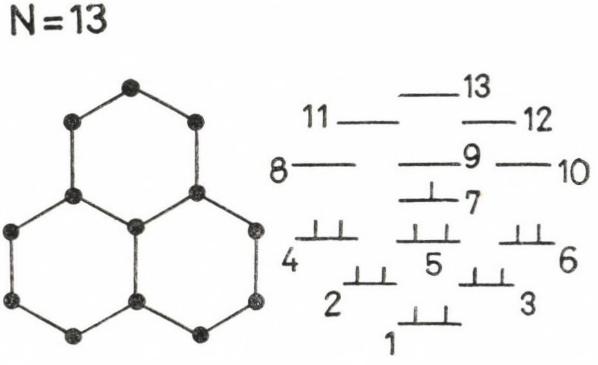


Fig. 4

linear combinations of the linearly independent eigenvectors associated with the degenerate eigenvalues are taken in order that equation (2) is satisfied. Furthermore, when N is even, definition (3) is still applicable in the case where degeneracies occur in the uppermost occupied eigenvalues *provided that* these eigenvalues are all doubly occupied (see Fig. 5).

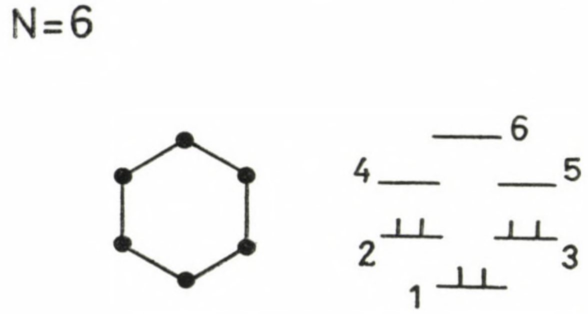


Fig. 5

B. Applicability of Eigenvalue Occupation-Numbers

Where there is degeneracy amongst the highest occupied eigenvalues, and this degeneracy occurs such that each member of the degenerate set is singly occupied ($v_i=1$), we have the situation which the two graphs shown in Figs. 6 and 7 illustrate. In such cases, because equation (25) is no longer reducible to equations (3) or (4), *q_r must be defined by the generalised definition given in equation* (25); otherwise, Theorem 2 will not hold. This is evident by reference, for example, to Figs. 6 and 7.

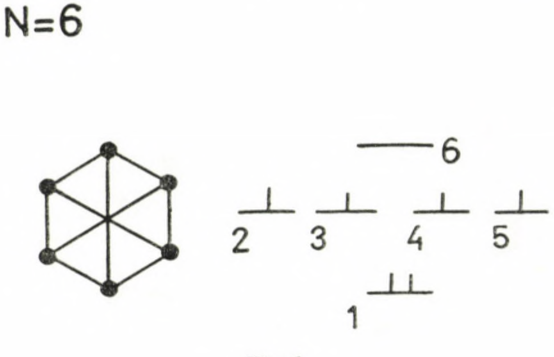


Fig. 6

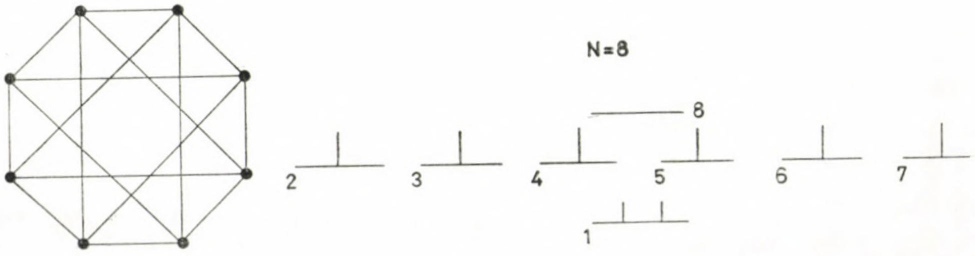


Fig. 7

C. An Ambiguous Case

There is one circumstance in which the index q_r cannot be defined at all; use of the *Aufbau* procedure in this case leads, both physically and mathematically, to an ambiguity. We illustrate this with reference to the graph shown in Fig. 8. * The situation in question arises when, by rules 1—3 of § II A, $a+b \neq N+1$.

The graph in Fig. 8 has six vertices and hence, from equation (24), we should expect that

$$\sum_i v_i = 6. \tag{26}$$

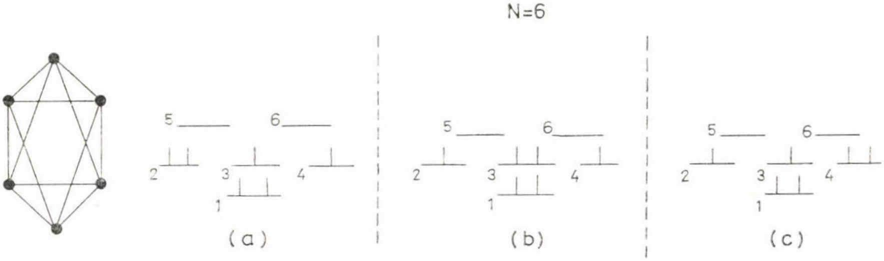


Fig. 8

However, following rules 1 to 3 of § II A leads to the conclusion that $a=2, b=4, N=6$; thus $a+b \neq N+1$ and, by rule 5 of § II A, $v_1, v_2, \ldots, v_N$ are undefined. On the basis of the physical *Aufbau* Principle, which the recipe outlined in § II A is designed to simulate, v_1 is unambiguously defined to be 2; from equation (26), this leaves a total 'occupation-number sum' of four still to dispose of. The *Aufbau* Principle then, however, predicts three equally plausible possible choices, as depicted in Fig. 8 for the cases labelled (a), (b) and (c). Use of the *Aufbau* Principle gives us no decision on which of these three cases our assignment should fall; in this situation, therefore, the $\{v_i\}$, and hence q_r, are left indeterminate.

To summarise, we can say that our definition (25) of q_r becomes ambiguous when the highest occupied eigenvalues are degenerate, though in the cases where the repeated eigenvalues are either *all* doubly or *all* singly occupied, this difficulty does not appear. However, in cases where, on the basis of the *Aufbau* Principle, some members of a family of repeated eigenvalues would have to be singly occupied, while others are doubly occupied, the ambiguity arises. In the present procedure, in which we have translated the physical *Aufbau* Principle into a series of purely abstract, logical, mathematical steps, such cases are diagnosed when a and b, determined by rules 1 to 3 of § II A, are such that

$$a+b \neq N+1. \tag{27}$$

* The graph in Fig. 8 may be visualised in three dimensions as a bipyramid (or octahedron). By use of the symmetry properties of this pyramid, the eigenvalue pattern depicted in Fig. 8 may be written down almost by inspection. We are most grateful to the late Professor C. A. Coulson, F. R. S., for kindly drawing our attention to this aspect of the eigenvalues of graphs.

D. Some Illustrative Examples

To conclude, we give a few numerical examples of q_r-values, calculated for several different graphs. The graphs depicted in Figs. 9 to 14 are all bipartite and, as anticipated from Theorem 2, the index q_r is precisely unity for each vertex of these graphs. The graphs illustrated in Figs. 15 to 20 are all non-bipartite and thus

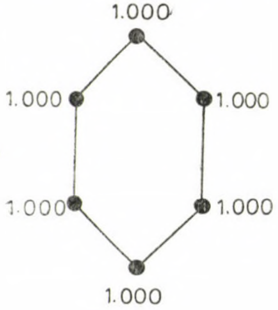


Fig. 9

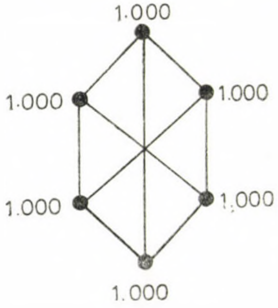


Fig. 10

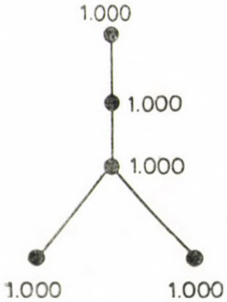


Fig. 11

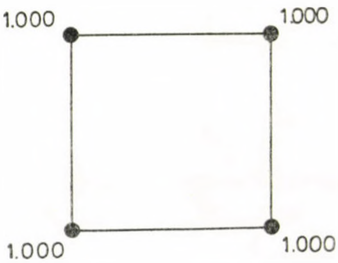


Fig. 12

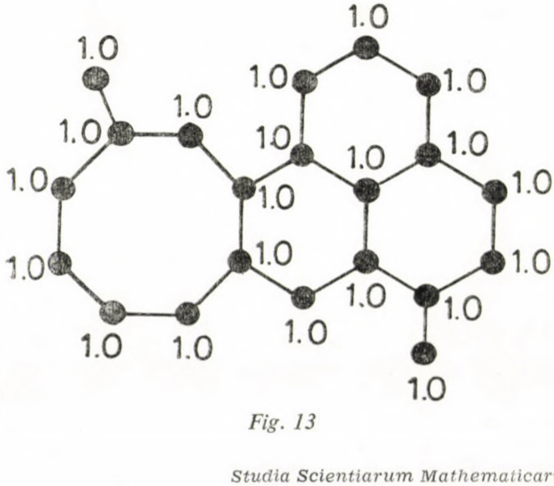


Fig. 13

have q_r-values which, in general, are not equal to unity. For the graphs we illustrate, values of q_r vary from 0.70 to 1.21, and no vertex has its q_r-index exactly equal to one.*

The bipartite graphs in Figs. 9 to 12 and in Fig. 14 have all been drawn so that they possess at least some element of symmetry. The lowest symmetry is that of Fig. 11 which has C_{2v}-symmetry. It should be noted, however, that *neither* symmetry *nor* the possession of an even number (N) of vertices is a prerequisite for having q_r-values equal to unity. Thus, the completely asymmetric graph shown in Fig. 13, having an odd number of vertices ($N=21$), has all its $\{q_r\}$ equal to unity because it is bipartite.

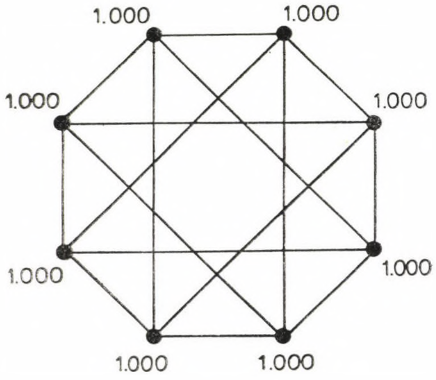


Fig. 14

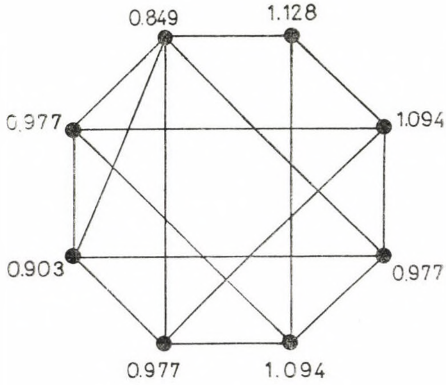


Fig. 15

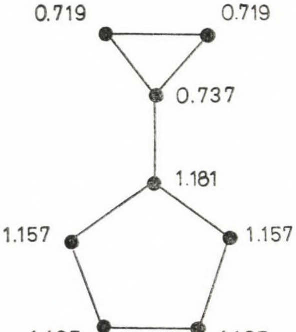
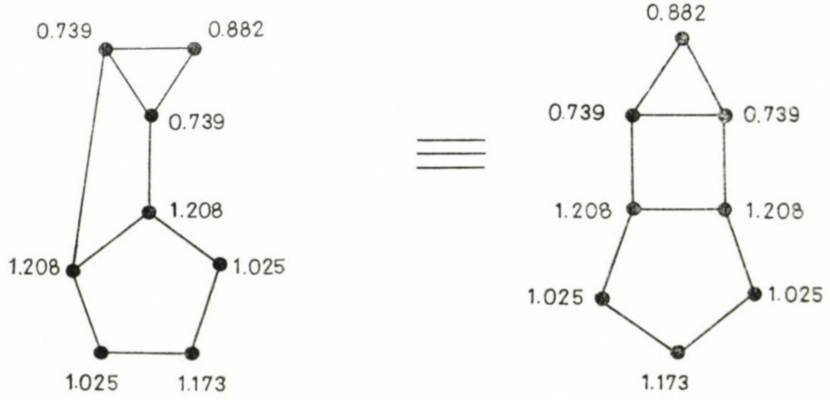


Fig. 16

*Figs. 9—20 illustrate numerical values of computed q_r-values for the vertices of some graphs typical of those commonly encountered. Where graphs represent known molecules, the q_r-values are taken directly from [24]; in all other cases, the q_r-values were computed from equation (25), by use of a standard Jacobian matrix-diagonalisation technique carried out on the Oxford University KDF-9 computer.

Among the non-bipartite graphs, that depicted in Fig. 17*a* has been drawn in such a way as to bring out the relation between it and those shown in Figs. 16 and 18. Both of the latter graphs, as drawn, clearly have C_{2v}-symmetry; but so also may that depicted in Fig. 17*a*, if it is drawn differently. This is evident from inspection of an isomorphic form of the graph, shown in Fig. 17*b*. Furthermore, it may be observed that the q_r-values of Fig. 17*a* reveal this underlying symmetry which the disposition of the vertices and edges in this Figure actually hides. This 'latent' symmetry is a consequence of the invariance of the eigenvalues (and eigenvectors) of an adjacency matrix under the kind of homeomorphism that converts Fig. 17*a* to Fig. 17*b*. Such invariance is, of course, reflected in the q_r-values calculated from the eigenvalues and eigenvectors of the adjacency matrix.



(a)

Fig. 17

(b)

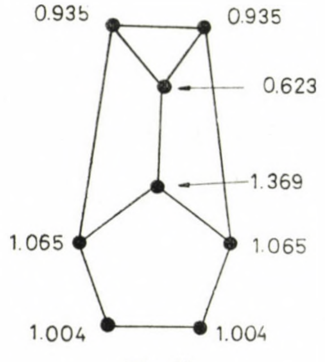


Fig. 18

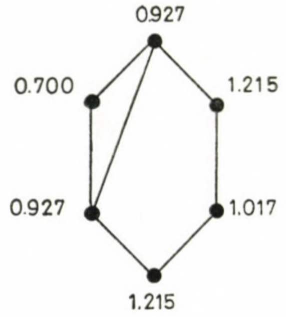


Fig. 19

It is also instructive to compare the q_r-values for the graphs shown in Figs. 14 and 15. These reveal the consequences, as far as the q_r-values are concerned, of introducing even a small 'perturbation' into a large, bipartite graph that makes it non-bipartite. The graph in Fig. 15 is readily obtained from that in Fig. 14 by the conversion of one of the many four-membered circuits it contains into a three-membered

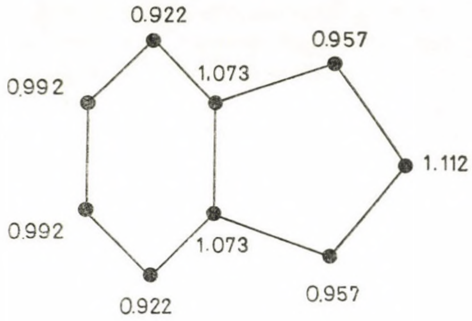


Fig. 20

circuit. Such a conversion yields a new, *non*-bipartite graph, with consequent repercussions on the q_r-values. Finally, we note that, both for the bipartite and non-bipartite graphs we consider, the 'conservation of charge' constraint implied in equation (5) applies. This is universally true; for example, both for the graph in Fig. 14, and that in Fig. 15, $\sum_r q_r = 8$, the number of vertices in each case.

Acknowledgements. We record our indebtedness to the late Professor C. A. Coulson, F.R.S. for his encouragement and for his kindness in making available facilities at the Mathematical Institute, University of Oxford, where most of this work was carried out. We are also grateful to Dr. R. J. Wilson (Open University, U.K.) for helpful discussions. R. B. M. thanks the Science Research Council of Great Britain for a Research Fellowship, and the Governing Body of Christ Church, Oxford, for a *Research Lecturership of the House*, during the tenure of which this manuscript was largely prepared.

REFERENCES

[1] Coulson, C. A. and Rushbrooke, G. S., Note on the method of molecular orbitals, *Proc. Cambridge Phil. Soc.* **36** (1940), 193—199.

[2] Mallion, R. B. and Rouvray, D. H., The Coulson—Rushbrooke pairing theorem: a case study for a multidisciplinary approach to certain aspects of Mathematics and Chemistry, *In preparation.*

[3] Rouvray, D. H., The topological matrix in quantum chemistry, in *Chemical Applications of Graph Theory* (Edited by A. T. Balaban), Academic Press, London, New York and San Francisco, 1976, Chapter 7 (pp. 176—221).

[4] Coulson, C. A., O'Leary, B. and Mallion, R. B., *Hückel Theory for Organic Chemists,* Academic Press, London, 1978.

[5] Mallion, R. B. and Rouvray, D. H., Molecular topology and the *Aufbau* Principle, *Molec. Phys.*, **36** (1978), 125—128.

[6] Perron, O., Über Matrizen, *Math. Ann.* **64** (1907), 248—263.

[7] Gantmacher, F. R., *The Theory of Matrices,* Vol. II, Chelsea Publishing Co., New York, 1960, 53—66.

[8] Frobenius, G., Über Matrizen aus positiven Elementen, *Sitz.-Ber. Deutsch. Akad. Wiss. Berlin, Math. Nat. Kl.* (1909), 471—476.

[9] Frobenius, G., Über Matrizen aus positiven Elementen, *Sitz.-Ber. Deutsch. Akad. Wiss. Berlin, Math. Nat. Kl.* (1909), 514—518.

[10] FROBENIUS, G., Über Matrizen aus nicht-negativen Elementen, *Sitz.-Ber. Deutsch. Akad. Wiss. Berlin, Math. Nat. Kl.* (1912), 456—477.

[11] WIELANDT, H., Unzerlegbare, nicht-negative Matrizen, *Math. Zeitschr.*, **52** (1950), 642—648.

[12] COLLATZ, L. and SINOGOWITZ, U.: Spektren endlicher Graphen, *Abh. Math. Sem. Univ. Hamburg*, **21** (1957), 63—77.

[13] SACHS, H., Über selbstkomplementäre Graphen, *Publ. Math. Debrecen*, **9** (1962), 270—288.

[14] HOFFMAN, A. J., On the polynomial of a graph, *Amer. Math. Monthly*, **70** (1963), 30—36.

[15] MARIMONT, R. B., System connectivity and matrix properties, *Bull. Math. Biophys.*, **31** (1969), 255—274.

[16] CVETKOVIĆ, D., Bihromatičnost i spektar grafa, *Mat. Biblio.*, **41** (1969), 193—194.

[17] MALLION, R. B., Eureka?, *Chem. Britain*, **9** (1973), 242.

[18] GUTMAN, I. and TRINAJSTIĆ, N., Graph theory and molecular orbitals, *Fortsch. Chem. Forsch. (Topics Current Chem.)*, **42** (1973), 49—93.

[19] GRAOVAC, A., GUTMAN, I., TRINAJSTIĆ, N. and ŽIVKOVIĆ, T., Graph theory and molecular orbitals, *Theoret. Chim. Acta*, **26** (1972), 67—78.

[20] MALLION, R. B., SCHWENK, A. J. and TRINAJSTIĆ, N., On the characteristic polynomial of a rooted graph, in *Recent Advances in Graph Theory; Proceedings of the Second Czechoslovak Symposium on Graph Theory* (Edited by M. Fiedler), Academia, Prague, 1975, 345—350.

[21] ROUVRAY, D. H., Les valeurs propres des molécules qui possèdent un graphe bipartit, *Comptes Rend. Acad. Sci. Paris, Sér.* **C**, **274** (1972), 1561—1563.

[22] RIGBY, M. J. and MALLION, R. B., On the eigenvalues and eigenvectors of certain finite, vertex-weighted, bipartite graphs, *J. Combinatorial Theory Ser. B* **27** (1979), 122—129.

[23] SALEM, L., *Molecular Orbital Theory of Conjugated Systems*, W. A. Benjamin, Inc., New York, 1966, 36—43.

[24] COULSON, C. A. and STREITWIESER, A., *Dictionary of Pi-Electron Calculations*, Pergamon Press, Oxford, 1965.

[25] LIBERLES, A., *Introduction to Molecular Orbital Theory*, Holt, Rinehart and Winston, Inc., New York, 1966, 59—66.

Mathematical Institute, University of Oxford, United Kingdom

Present addresses:
R. B. Mallion: The King's School, Canterbury, United Kingdom
D. H. Rouvray: Diebold Europe S. A., 5/6, Argyll Street, London W 1,
United Kingdom

(Received February 2, 1978)

Printed in Hungary

A kiadásért felel az Akadémiai Kiadó igazgatója — Műszaki szerkesztő: Botyánszky Pál
A kézirat a nyomdába érkezett: 1980. III. 6. — Terjedelem: 21,25 (A/5) ív, 47 ábra

80-1202 — Szegedi Nyomda — F. v.: Dobó József igazgató

CONTENTS

31. VII. 1981. **Index: 26 748**

HU ISSN 0081—6906

Die *Studia Scientiarum Mathematicarum Hungarica* ist eine Halbjahrsschrift der Ungarischen Akademie der Wissenschaften. Sie veröffentlicht Originalbeiträge aus dem Bereich der Mathematik in deutscher, englischer, französischer oder russischer Sprache. Es erscheint jährlich ein Band.

Adresse der Redaktion: 1053 Budapest V., Reáltanoda u. 13—15, Ungarn.
Technischer Redaktor: J. Merza

Bestellbar bei Buch- und Zeitungs-Aussenhandelsunternehmen *Kultura* (Budapest 62, P. O. B. 149), oder bei den Vertretungen im Ausland.

Austauschabmachungen können mit der Bibliothek des Mathematischen Instituts (1053 Budapest V., Reáltanoda u. 13—15) getroffen werden.

Die zur Veröffentlichung bestimmten Manuskripte sind in zwei Exemplaren an die Redaktion zu schicken.

Studia Scientiarum Mathematicarum Hungarica est une revue biannuelle de l'Académie Hongroise des Sciences publiant des essais originaux, en francais, anglais, allemand ou russe, du domaine des mathématiques.

Rédaction: 1053 Budapest V., Reáltanoda u. 13—15, Hongrie.
Rédacteur technique: J. Merza

On s'abonne chez *Kultura*, Société pour le Commerce de Livres et Journaux (Budapest 62, P. O. B. 149) ou chez ses représentants à l'étranger.

Pour établir des relations d'échange on est prié de s'adresser à la Bibliothèque de l'Institut de Mathématique (1053 Budapest V., Reáltanoda u. 13—15).

On est prié d'envoyer les articles destinés à la publication en deux exemplaires à l'adresse de la Rédaction.

Studia Scientiarum Mathematicarum Hungarica — выходит два раза в год в Издании Академии Наук Венгрии. Журнал публикует оригинальные исследования в области математики на русском, немецком, английском, и французском языках. Отдельные выпуски составляют ежегодно один том.

Адрес редакции: 1053 Budapest V., Reáltanoda u. 13—15, Венгрия.
Технический редактор: J. Merza

Подписка на журнал принимается Внешнеторговым предприятием „Култура" (Budapest 62, P. O. B. 149) или его представительствами за границей.

По поводу отношения обмена просим обрашаться к Библиотеке Института Математики (1053 Budapest V., Reáltanoda u. 13—15).

Работы, преднасначенные для опубликования в журнале следует направлять по адресу редакции в двух экземплярах.

Printed in Hungary

A kiadásért felel az Akadémiai Kiadó igazgatója — Műszaki szerkesztő: Rózsa Katalin
A kézirat a nyomdába érkezett: 1980. V. 21. — Terjedelem: 20,5 (A/5) ív 59 ábra

80-2235 — Szegedi Nyomda — F. v.: Dobó József igazgató

Studia Scientiarum Mathematicarum Hungarica

A Magyar Tudományos Akadémia matematikai folyóirata

Szerkesztőség: 1053 Budapest V., Reáltanoda u. 13—15.

Technikai szerkesztő: Merza József

Kiadja az Akadémiai Kiadó, 1054 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 (1011 Budapest I., Fő u. 32.).

Cserekapcsolatok felvétele ügyében kérjük az MTA Matematikai Kutató Intézete Könyvtárához (1053 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: 1053 Budapest V., Reáltanoda u. 13—15, Hungary.
Technical Editor: J. Merza

Orders may be placed with *Kultura* 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 (1053 Budapest V., Reáltanoda u. 13—15.)

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

Studia Scientiarum Mathematicarum Hungarica

AUXILIO
CONSILII INSTITUTI MATHEMATICI
ACADEMIAE SCIENTIARUM HUNGARICAE

OMUS XIII.
ASC. 3—4.
978

AKADÉMIAI KIADÓ, BUDAPEST

CONTENTS

31. VII. 1981 **Index: 26 748**

HU ISSN 0081—6906

Die *Studia Scientiarum Mathematicarum Hungarica* ist eine Halbjahrsschrift der Ungarischen Akademie der Wissenschaften. Sie veröffentlicht Originalbeiträge aus dem Bereich der Mathematik in deutscher, englischer, französischer oder russischer Sprache. Es erscheint jährlich ein Band.

Adresse der Redaktion: 1053 Budapest V., Reáltanoda u. 13—15, Ungarn.
Technischer Redaktor: J. Merza

Bestellbar bei Buch- und Zeitungs-Aussenhandelsunternehmen *Kultura* (Budapest 62, P. O. B. 149), oder bei den Vertretungen im Ausland.

Austauschabmachungen können mit der Bibliothek des Mathematischen Instituts (1053 Budapest V., Reáltanoda u. 13—15) getroffen werden.

Die zur Veröffentlichung bestimmten Manuskripte sind in zwei Exemplaren an die Redaktion zu schicken.

Studia Scientiarum Mathematicarum Hungarica est une revue biannuelle de l'Académie Hongroise des Sciences publiant des essais originaux, en francais, anglais, allemand ou russe, du domaine des mathématiques.

Rédaction: 1053 Budapest V., Reáltanoda u. 13—15, Hongrie.
Rédacteur technique: J. Merza

On s'abonne chez *Kultura*, Société pour le Commerce de Livres et Journaux (Budapest 62, P. O. B. 149) ou chez ses représentants à l'étranger.

Pour établir des relations d'échange on est prié de s'adresser à la Bibliothèque de l'Institut de Mathématique (1053 Budapest V., Reáltanoda u. 13—15).

On est prié d'envoyer les articles destinés à la publication en deux exemplaires à l'adresse de la Rédaction.

Studia Scientiarum Mathematicarum Hungarica — выходит два раза в год в Издании Академии Наук Венгрии. Журнал публикует оригинальные исследования в области математики на русском, немецком, английском, и французском языках. Отдельные выпуски составляют ежегодно один том.

Адрес редакции: 1053 Budapest V., Reáltanoda u. 13—15, Венгрия.
Технический редактор: J. Merza

Подписка на журнал принимается Внешнеторговым предприятием „Култура" (Budapest 62, P. O. B. 149) или его представительствами за границей.

По поводу отношения обмена просим обрашаться к Библиотеке Института Математики (1053 Budapest V., Reáltanoda u. 13—15).

Работы, преднасначенные для опубликования в журнале следует направлять по адресу редакции в двух экземплярах.

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 245—252.

SUR LA REPRÉSENTATION BINAIRE DES CATÉGORIES POLYADIQUES

par

V. V. TOPENTCHAROV and Y. N. ARNAOUDOV

Dans la théorie des catégories polyadiques [7], un des cas de la théorie générale des $\{n\}$-catégories [8], on se pose, par analogie avec la théorie des n-quasi-groupes [1] et des n-groupes [2], les problèmes de présentation par catégories binaires. On distingue deux approches. Dans le premier, étant donnée une catégorie polyadique $\mathbf{C}_{[n]}$, on cherche une catégorie $\mathbf{H}$ telle que $\mathbf{C}_{[n]}$ puisse être plongée dans la catégorie polyadique $\mathbf{H}_{[n]}$ dérivée de $\mathbf{H}$; en général $\mathbf{H}_{[n]}$ est à ensemble-support plus grand que le support de $\mathbf{C}_{[n]}$; c'est un analogue du problème de recouvrement des n-groupes de E. L. Post [6]; sa solution pour les catégories polyadiques est donnée dans [7], [9], [11]. Le second approche consiste à construire une catégorie $\mathbf{C}$ sur l'ensemble-support de $\mathbf{C}_{[n]}$ et un endofoncteur F tels que tout composé dans $\mathbf{C}_{[n]}$ soit présentable comme composé dans $\mathbf{C}$, les éléments de la suite composable étant alors les transformés par la puissance $i-1$ de F, i étant le numéro des éléments dans la n-suite composable dans $\mathbf{C}_{[n]}$; ce problème est un analogue du théorème de L. M. Gluskin [3] et de M. Hosszú [4]; pour les catégories polyadiques il est résolu dans [9], [10], [11] par des méthodes déduites de celle de [3].

Nous nous proposons dans cet article de donner une nouvelle démonstration du théorème de représentation binaire d'une catégorie polyadique par une catégorie et un endofoncteur en suivant la méthode de Hosszú [4]. La terminologie et les notations sont pour les catégories celles de S. MacLane [5] et pour les catégories polyadiques — celles de [7].

1. Définitions et généralités sur les catégories polyadiques

Soit $\mathbf{C}$ une catégorie, C son ensemble-support et $\mathbf{C}_0$ l'ensemble de ses objets, identifié à l'ensemble des unités de $\mathbf{C}$; on note $[\mathbf{C}]=(C, \beta, \alpha)$ le graphe (orienté) sous-jacent de $\mathbf{C}$. On appelle $[n]$-graphe le triplet $[C]_{[n]}=(C, \hat{\beta}, \hat{\alpha})$, où pour tout $f\in C$ on a

$$\hat{\beta}\colon f\mapsto (\beta(f))^{n-1}=(\beta(f), \dots, \beta(f)), \quad \hat{\alpha}\colon f\mapsto (\alpha(f))^{n-1}=(\alpha(f), \dots, \alpha(f)).$$

A tout graphe $[\mathbf{C}]$ est canoniquement associé un $[n]$-graphe $[\mathbf{C}]_{[n]}$ et inversement; dans la suite on identifiera $[C]\cong[C]_{[n]}$. Considérons la loi n-aire interne

AMS (MOS) subject classifications (1970). Primary 08A00; Secondary 18A99.

Key words and phrases. Category, polyadic category, binary representation, derivable polyadic category.

$k'_n : C \overset{n}{*} C \to C$ définie sur l'ensemble $C \overset{n}{*} C \subset C^n$ des n-suites composables; le couple (C, k'_n) est appelé *n-classe*. On a [7], [8]:

DÉFINITION 1. Le quadruplet $(C, k'_n, \hat{\beta}, \hat{\alpha})$ est appelé *catégorie polyadique* d'ordre $n \in \mathbf{N}$ (*$[n]$-catégorie*) et noté $\mathbf{C}_{[n]}$ si les conditions (axiomes (K)) suivantes sont vérifiées:

(K.1) (C, k'_n) est une n-classe; $(C, \hat{\beta}, \hat{\alpha})$ est un $[n]$-graphe.

(K.2) Pour toute n-suite composable $(x_i\colon i \leqq n) \in C \overset{n}{*} C$ on a:

$$\hat{\beta} \,.\, k'_n\big((x_i\colon i \leqq n)\big) = \hat{\beta}(x_1) \quad \text{et} \quad \hat{\alpha} \,.\, k'\big((x_i\colon i \leqq n)\big) = \hat{\alpha}(x_n).$$

(K.3) $C \overset{n}{*} C = \{(x_i\colon i \leqq n)\colon \hat{\alpha}(x_p) = \hat{\beta}(x_{p+1}),\ p \leqq n-1\} \subset C^n$.

(K.4) $(\hat{\beta}(x), x) \in C \overset{n}{*} C$ et $[(\hat{\beta}(x), x)] = x$, $(\hat{\alpha}(x), x) \in C \overset{n}{*} C$ et $[(\hat{\alpha}(x), x)] = x$.

(K.5) Soit la suite $(x_s\colon s \leqq 2n-1)$; si $(x_p, \ldots, x_{p+n-1}) \in C \overset{n}{*} C$ pour tout $p \leqq n-1$ et si

$$(x_1, \ldots, x_{p-1}, [(x_p, \ldots, x_{p+n-1})], x_{p+n}, \ldots, x_{2n-1}) \in C \overset{n}{*} C,$$
$$[(x_1, \ldots, x_{n-1}, [(x_n, \ldots, x_{2n-1})])] =$$
$$= [(x_1, \ldots, x_{p-1}, [(x_p, \ldots, x_{p+n-1})], x_{p+n}, \ldots, x_{2n-1})].$$

On en déduit les définitions de n-quasi-groupe et de n-groupe [1], [2]. Plus loin on écrira β et α au lieu de $\hat{\beta}$ et $\hat{\alpha}$ puisque $[\mathbf{C}] \cong [\mathbf{C}]_{[n]}$.

Comme pour les catégories [5], on construit la notion d'homomorphisme entre $[n]$-catégories $F\colon \mathbf{C}_{[n]} \to \mathbf{H}_{[n]}$, appelé *$[n]$-foncteur*.

THÉORÈME 1 [7]. *Toute $[n]$-catégorie admet une catégorie de recouvrement* $\mathbf{H}$, *i. e., une catégorie telle que*

$$(x_i\colon i \leqq n) \in C \overset{n}{*} C, \quad [(x_i\colon i \leqq n)] = x_1 \,.\, x_2 \,.\, \ldots \,.\, x_n \in H.$$

COROLLAIRE (Coset theorem [6]). *Pour tout n-groupe il existe un groupe de recouvrement.*

Soit maintenant $\mathbf{H}$ une catégorie, a un inversible de $\mathbf{H}$ et $F\colon \mathbf{H} \to \mathbf{H}$ un endofoncteur vérifiant les conditions suivantes

$$F(a) = a, \quad a \,.\, x \,.\, a^{-1} = F^{n-1}(x), \quad n \in \mathbf{N} \quad \text{et} \quad x \in H.$$

Au quadruplet $T = T(\mathbf{H}, a, F, n)$ est canoniquement associée la loi n-aire

$$[(x_i\colon i \leqq n)] = x_1 \,.\, F(x_2) \,.\, F^2(x_3) \,.\, \ldots \,.\, F^{n-1}(x_{n-1}) \,.\, a \,.\, x_n.$$

On a le résultat suivant:

THÉORÈME 2 [10]. *Toute $[n]$-catégorie $\mathbf{C}_{[n]}$ est isomorphe à une sous-$[n]$-catégorie de la $[n]$-catégorie canoniquement associée au quadruplet $T = T(\mathbf{L}[\mathbf{C}], e, F, n)$, où $\mathbf{L}[\mathbf{C}]$ est la catégorie libre des chemins et $e \in [\mathbf{C}]_0$.*

COROLLAIRE 1. *Pour toute $[n]$-catégorie $\mathbf{C}_{[n]}$ il existe une catégorie telle que $\mathbf{C}_{[n]}$ soit isomorphe à la $[n]$-catégorie canoniquement associée à $T = T(\mathbf{L}[\mathbf{C}], e, F, n)$, $e \in [\mathbf{C}]_0$, $F\colon \mathbf{L}[\mathbf{C}] \to \mathbf{L}[\mathbf{C}]$.*

COROLLAIRE 2 [3]. *Pour tout n-groupe G_n il existe un groupe (binaire) G tel que G_n soit isomorphe au n-groupe canoniquement associé au quadruplet $T=T(G, e, F, n)$, e étant l'unité du groupe.*

Ce résultat donne la solution du problème de présentation binaire, mais ne fait pas ressortir le fait important que les ensembles-supports de $\mathbf{C}_{[n]}$ et de la $[n]$-catégorie associée à T sont identiques.

2. Catégorie de représentation pour une catégorie polyadique

Soit $\mathbf{C}_{[n]}$ une $[n]$-catégorie. On se propose de construire sur C une loi de composition k' canoniquement associée à $k_n'=[(\dots)]$ de $\mathbf{C}_{[n]}$ telle que si $[\mathbf{C}]=(C, \beta, \alpha)$ est le graphe sous-jacent de $\mathbf{C}_{[n]}$, le quadruplet $C=(C, k', \beta, \alpha)$ soit une catégorie où à l'aide d'un endofoncteur convenablement construit tout composé $[(x_i\colon i\leqq n)]$ soit représentable par des composition dans $\mathbf{C}$.

LEMME 1. *Soit* $\mathbf{C}_{[n]}$ *une* $[n]$*-catégorie et* $[\mathbf{C}]=(C, \beta, \alpha)$ *son graphe sous-jacent. La loi de composition* $.\ \colon C*C\to C$ *telle que:*

$$.\ \colon (x, y)\mapsto \big[(x, \alpha^{n-2}(x), y)\big] \quad \textit{si et seulement si} \quad \alpha(x)=\beta(y),$$

définit sur C la structure de catégorie $\big(\alpha^{n-2}(x)=(\alpha(x), \dots, \alpha(x)), n-2$ fois$\big)$.

DÉMONSTRATION. Évidemment $(C, .)$ est un système multiplicatif et (C, β, α) est un graphe. L'ensemble des couples composables par la loi . est donné comme suit:

$$C*C=\{(x, y)\colon \alpha(x)=\beta(y),\ x, y\in C\}\subset C\times C,$$

ce qui est l'ensemble des couples composable de la catégorie sur $[C]$. Soit $(x, y)\in C*C$, i. e., $x\,.\,y$ est bien définit; on a:

$$\alpha(x\,.\,y)=\alpha\big([(x, \alpha^{n-2}(x), y)]\big)=\alpha(y),$$
$$\beta(x\,.\,y)=\beta\big([(x, \beta^{n-2}(x), y)]\big)=\beta(x),$$

ce qui découle de (K.2); on a donc

$$\alpha(x\,.\,y)=\alpha(y) \quad \text{et} \quad \beta(x\,.\,y)=\beta(x)$$

et l'axiome des unités des projections est vérifié. Soit $x=\alpha(x)=e$; puisque $\alpha(e)=\beta(y)$, on a suivant (K.4):

$$e\,.\,y=\big[(e, \alpha^{n-2}(e), y)\big]=[(e, \dots, e, y)]=y$$

et l'axiome des unités dans la catégorie est vérifié. Soit $x, y, z\in C$ et $\alpha(x)=\beta(y)$, $\alpha(y)=\beta(z)$; alors suivant (K.5) on a:

$$(x\,.\,y)\,.\,z=\big[\big([(x, \alpha^{n-2}(x), y)], \alpha^{n-2}(y), z\big)\big]=\big[\big(x, \alpha^{n-2}(x), [(y, \alpha^{n-2}(y), z)]\big)\big]=x\,.\,(y, z)$$

et l'associativité de . est prouvée. Ceci termine la démonstration.

LEMME 2. *L'application* $F\colon C\to C$ *telle que*

$$F\colon x\mapsto \big[(\beta(x), x, \alpha^{n-2}(x))\big]$$

où $k'_n=[(\ldots)]$ *est la loi n-aire de* $\mathbf{C}_{[n]}$ *et* β, α *les rétractions du graphe sous-jacent* $[\mathbf{C}]_{[n]}$ *définit un endofoncteur de* $\mathbf{C}$.

DÉMONSTRATION. Il faut démontrer que $F(\mathbf{C}_0)\subset\mathbf{C}_0$ et que $F(x\,.\,y)=$ $=F(x)\,.\,F(y)$ pour tout $(x, y)\in C*C$. Soit $x=e\in\mathbf{C}_0$; on a:

$$F(e)=\big[(\beta(e), e, \alpha^{n-2}(e))\big]=[(e, \ldots, e)]=e\in\mathbf{C}_0.$$

Soit maintenant $(x, y)\in C*C$; alors

$$F(x\,.\,y)=\big[(\beta(x\,.\,y), x\,.\,y, \alpha^{n-2}(x\,.\,y))\big];$$

puisque $\mathbf{C}$ est une catégorie, on a suivant le Lemme 1:

$$\beta(x\,.\,y)=\beta(x) \quad \text{et} \quad \alpha(x\,.\,y)=\alpha(y)$$

et remplaçant ci-dessus, on obtient:

$$F(x\,.\,y)=\big[(\beta(x), x\,.\,y, \alpha^{n-2}(y))\big];$$

mais en vertu de la construction de la loi . on a:

$$x\,.\,y=\big[(x, \alpha^{n-2}(x), y)\big],$$

ce qui remplacé dans la dernière expression pour $F(x\,.\,y)$ donne:

$$F(x\,.\,y)=\big[(\beta(x), [(x, \alpha^{n-2}(x), y)], \alpha^{n-2}(y))\big]$$

et suivant l'axiome de l'associativité n-aire (K.5) on obtient

$$F(x\,.\,y)=\big[([(\beta(x), x, \alpha^{n-2}(x))], y, \alpha^{n-2}(y))\big];$$

rappelons que

$$y=\beta(y)\,.\,y=\big[(\beta(y), \alpha^{n-2}(\beta(y)), y)\big]=\big[(\beta^{n-1}(y), y)\big];$$

et remplaçons cette dernière présentation de $y\in C$ dans l'expression pour $F(x\,.\,y)$; utilisant une fois l'associativité n-aire (K.5) on a:

$$F(x\,.\,y)=\big[([(\beta(x), x, \alpha^{n-2}(x))], [(\beta^{n-1}(y), y)], \alpha^{n-2}(y))\big]=$$
$$=\big[([(\beta(x), x, \alpha^{n-2}(x))], \beta^{n-2}(y), [(\beta(y), y, \alpha^{n-2}(y)])\big];$$

mais en vertu de la construction de F et de . on a

$$\big[(\beta(x), x, \alpha^{n-2}(x))\big]=F(x), \big[(\beta(x), y, \alpha^{n-2}(y))\big]=F(y), \quad \alpha(x)=\beta(y);$$

remplaçant on a:

$$F(x\,.\,y)=\big[(F(x), \alpha^{n-2}(x), F(y))\big];$$

mais conformément à (K.2)

$$\alpha\big(F(x)\big)=\alpha\big([(\beta(x), x, \alpha^{n-2}(x))]\big)=\alpha(x),$$

ce qui permet d'écrire

$$F(x\,.\,y)=\big[(F(x), \alpha^{n-2}(F(x)), F(y))\big].$$

En vertu de la construction de la loi . on a définitivement l'expression

$$F(x\,.\,y)=F(x)\,.\,F(y).$$

COROLLAIRE. *L'endofoncteur F vérifie les conditions*

$$F_{|\mathbf{C}_0} = \mathrm{Id}_{\mathbf{C}_0} \quad \text{et} \quad F^n = F.$$

DÉMONSTRATION. Puisque $F(e)=e$ pour tout $e\in\mathbf{C}_0$, on a évidemment $F_{|\mathbf{C}_0}=$ $=\mathrm{Id}_{\mathbf{C}_0}$. Suivant (K.2) l'unité à droite de la composition définissant $F(x)$ est nécessairement $\alpha(x)$ et l'unité à gauche $-\beta(x)$. Alors dans l'expression

$$F^n(x) = \big[(\beta(x), [(\beta(x), \dots, [(\beta(x), x, \alpha^{n-2}(x))], \alpha^{n-2}(x), \dots, \alpha^{n-2}(x))], \alpha^{n-2}(x))\big]$$

on a à gauche exactement n exemplaires de $\beta(x)$ et à droite exactement n exemplaires de $\alpha^{n-2}(x)$; la seconde affirmation du Corollaire s'obtient alors en appliquant $n-1$ fois l'axiome (K.5) de l'associativité n-aire.

THÉORÈME 3. *Pour tout $[n]$-catégorie $\mathbf{C}_{[n]}=(C, k'_n, \hat{\beta}, \hat{\alpha})$ il existe une catégorie sur le même support et de même graphe sous-jacent $[\mathbf{C}]=(C, \beta, \alpha)$ et un endofoncteur $F\colon \mathbf{C}\to\mathbf{C}$, tels que pour toute n-suite composable $(x_i\colon i\leqq n)\in C\overset{n}{*}C$ l'expression suivante est vérifiée:*

$$[(x_i\colon i \leqq n)] = x_1 \,.\, F(x_2) \,.\, F^2(x_3) \,.\, \dots \,.\, F^{n-1}(x_n),$$

où $F^2=F\,.\,F$, $F^k=F\dots F$ (k fois).

DÉMONSTRATION. L'existence de la catégorie $\mathbf{C}$ et de l'endofoncteur F est prouvée dans les Lemmes 1 et 2. Nous démontrerons la formule de représentation. Soit à cet effet la n-suite composable $(x_i\colon i\leqq n)\in C\overset{n}{*}C$ et présentons tout x_p de $(x_i\colon i\leqq n)$ comme suit

$$\big[(x_p, \alpha^{n-1}(x_p))\big] = x_p, \quad p \leqq n;$$

remplaçant ces expressions dans le composé de $(x_i\colon i\leqq n)$, on a:

$$[(x_i\colon i \leqq n)] = \big[([(x_1, \alpha^{n-1}(x_1))], [(x_2, \alpha^{n-1}(x_2))], \dots, [(x_n, \alpha^{n-1}(x_n))])\big];$$

puisque $\alpha(x_{p-1})=\beta(x_p)$ suivant la condition de composabilité dans les $[n]$-catégories, on applique l'associativité n-aire et on obtient:

$$\begin{aligned}
[(x_i\colon i \leqq n)] &= \big[(x_1, \alpha^{n-2}(x_1), [([(\beta(x_2), x_2, \alpha^{n-2}(x_2))], \dots\\
&\dots, [(\beta(x_{n-1}), x_{n-1}, \alpha^{n-2}(x_{n-1}))], [(\beta(x_n), x_n, \alpha^{n-2}(x_n))], \alpha(x_n))])\big] =\\
&= \big[(x_1, \alpha^{n-2}(x_1), [([(\beta(x_2), x_2, \alpha^{n-2}(x_2))], \dots\\
&\dots, [(\beta(x_{n-1}), x_{n-1}, \alpha^{n-2}(x_{n-1}))], F(x_n), \alpha(x_n))])\big] =\\
&= \big[(x_1, \alpha^{n-2}(x_1), [(\beta(x_2), x_2, \alpha^{n-2}(x_2))], \dots\\
&\dots, [(\beta(x_{n-2}), x_{n-2}, \alpha^{n-2}(x_{n-2}))], \beta(x_{n-1}), [(x_{n-1}, \alpha^{n-2}(x_{n-1}), F(x_n))], \alpha(x_n))])\big] =\\
&= \big[(x_1, \alpha^{n-2}(x_1), [([(\beta(x_2), x_2, \alpha^{n-2}(x_2))], \dots\\
&\dots, [(\beta(x_{n-2}), x_{n-2}, \alpha^{n-2}(x_{n-2}))], \beta(x_{n-1}), x_{n-1} \,.\, F(x_n), \alpha(x_n))])\big] =\\
&= \big[(x_1, \alpha^{n-2}(x_1), [([(\beta(x_2), x_2, \alpha^{n-2}(x_2))], \dots, [(x_{n-1} \,.\, F(x_n), \alpha^{n-1}(x_n)], \alpha(x_n))])\big] =\\
&= \big[(x_1, \alpha^{n-2}(x_1), [([(\beta(x_2), x_2, \alpha^{n-2}(x_2))], \dots\\
&\dots, [(\beta(x_{n-1}), x_{n-1} \,.\, F(x_n), \alpha^{n-2}(x_n))], \alpha^2(x_n))])\big] = \dots =\\
&= \big[(\dots, [([(\dots)], \dots, [(\beta(x_{n-1} \,.\, F(x_n)), x_{n-1} \,.\, F(x_n), \alpha^{n-2}(x_{n-1} \,.\, F(x_n))], \alpha^2(x_n))])\big] =\\
&= \big[(\dots, [([(\dots)], \dots, F(x_{n-1}) \,.\, F(x_n))], \alpha^2(x_n))])\big];
\end{aligned}$$

continuant le processus de droite à gauche, on obtient enfin

$$[(x_1 . F(x_2 . F(x_3 . \dots . F(x_n)\dots), \alpha^{n-1}(x_n))] =$$
$$= x_1 . F(x_2) . F^2(x_3) . \dots . F^{n-1}(x_n).$$

Si en particulier $C \overset{n}{*} C = C^n$, la $[n]$-catégorie se réduit à un n-demi-groupe et le Théorème 3 est identique au résultat principal de [12]; si de plus la structure n-aire est un n-groupe [1], notre résultat se réduit au théorème classique de GLUSKIN—HOSSZÚ dans la variante de M. HOSSZÚ [4].

3. Quelques suppléments

La question naturelle suivante se pose: sous quelles conditions $F = F^2 = \dots$ $\dots = F^{n-1} = \mathrm{Id}_C$, i.e., pour tout $(x_i\colon i \leqq n) \in C \overset{n}{*} C$ on a

$$[(x_i\colon i \leqq n)] = x_1 . x_2 . \dots . x_n,$$

où . est la loi de composition dans la catégorie **C** sur le même support que $\mathbf{C}_{[n]}$, construite ci-dessus. Pour en donner la réponse rappelons encore une définition (voir [10]).

On dira que $\mathbf{H}_{[n]}$ est une *$[n]$-catégorie dérivée* pour la catégorie **H**, $[\mathbf{H}] = [\mathbf{H}_{[n]}]$ si la composition n-aire dans $\mathbf{H}_{[n]}$ se réduit à l'application $n-1$ fois de la composition binaire de **H** sur les éléments des n-suites composables respectant l'ordre de composition, i.e., si toute composition n-aire est un n-chemin dans [**C**]. Toute $[n]$-catégorie peut être canoniquement plongée dans la catégorie libre des chemins **L**[**C**], mais si $\mathbf{L}_n[\mathbf{C}]$ est la catégorie des n-chemins, l'isomorphisme $\mathbf{L}_n[\mathbf{C}] \cong \mathbf{C}_{[n]}$, où $\mathbf{L}_n[\mathbf{C}]$ est munie de la loi naturelle de composition n-aire de n-chemins (i.e., on compose n chemins de longueur nk pour des valeurs en général distinctes de k de sorte que le chemin-composé soit de longueur $\left(\sum_{i=1}^{n} k_i\right) . n)$ n'existe pas toujours.

Des $[n]$-catégories primitives existent. En effet soit $[\mathbf{C}] = (C, \beta, \alpha)$ un graphe (orienté), choisissons $m > 2$ et notons **L**[**C**] la catégorie libre des chemins sur [**C**]; notons $L'[C] \subset L[C]$ l'ensemble des chemins de longueur $l(f) = s\,.\,m+1$, $s \in N$; on suppose que entre deux sommets de [**C**] il existe au plus un chemin de $L'[C]$; si [**C**] donne lieu à plusieurs chemins on effectue une factorisation appropriée. On munie $L'[C]$ de la structure de $[m+1]$-catégorie comme suit: les rétractions $\hat{\beta}$ et $\hat{\alpha}$ sont trivialement déduites de β et α du graphe [**C**] et la loi de composition canoniquement induite par celle de **L**[**C**] donne pour toute suite $(f_1, f_2, \dots, f_{m+1})$ de $m+1$ chemins composables (concaténaires) de $L'[C]$:

$$[(f_1, f_2, \dots, f_{m+1})] = f_1 . f_2 . \dots . f_{m+1};$$

puisque $l(f_i) = s_i m + 1$, on a pour le composé

$$l[(f_1, \dots, f_{m+1})] = 1 + m \sum_{i=1}^{m+1} s_i$$

la vérification des axiomes (K) s'effectue sans difficultés, d'où la structure de $[m+1]$-catégorie sur $L'[C]$. Supposons qu'il existe sur $L'[C]$ une loi de composition n-aire k_n qui en fait une $[n]$-catégorie; $m_n > n \geqq 2$; alors la longueur du composé $k_n(f_1, f_2, \dots, f_n)$, $f_i \in L'[C]$ est

$$\sum_{i=1}^{n} l(f_i) = n + m \sum_{i=1}^{n} r_i;$$

mais alors $k_n(f_1, \dots, f_n) \in L'[C]$ puisque dans le cas contraire on aurait

$$n + m \sum_{i=1}^{n} r_i = ms + 1$$

pour une valeur de $s \in \mathbf{N}$; il s'en suit que $n-1$ est multiple de m, ce qui est une contradiction. Donc la $[m+1]$-catégorie $(L'[C], [\dots], \beta, \alpha)$ est primitive.

On a le résultat suivant:

THÉORÈME 4. *Les conditions suivantes sont équivalentes:*

(a) $\mathbf{C}_{[n]}$ *est dérivée de* $\mathbf{C}$.

(b) *L'endomorphisme* F *(du Lemme 2) est l'identité de* $\mathbf{C}$*; i. e.* $F = \mathrm{Id}_{\mathbf{C}}$.

(c) *Pour tout* $x \in C$ *on a* $[(\beta(x), x, \alpha^{n-2}(x))] = x$ *dans* $\mathbf{C}_{[n]}$.

DÉMONSTRATION. Supposons que $\mathbf{C}_{[n]}$ est dérivée de $\mathbf{C}$; prenons un n-chemin $(x_1, \dots, x_n)$ arbitraire dans $[\mathbf{C}]$; on a $(x_i\colon i \leqq n) \in C \overset{n}{*} C$ et aussi

$$[(x_1, x_2, \dots, x_n)] = x_1 . x_2 . \dots . x_n;$$

posant en particulier $x_1 = \beta(x_2) = e'$, $x_2 = x$ et $x_3 = x_4 = \dots = x_n = \alpha(x_2) = \alpha(x) = e$, on obtient:

$$[(e', x, e, \dots, e)] = e' . x . e . \dots . e = x,$$

i.e., (a)⇒(c).

Supposons que dans $\mathbf{C}_{[n]}$ est vérifiée la condition $[(\beta(x), x, \alpha^{n-2}(x))] = x$; alors suivant la construction de F (voir Lemme 2) on a $F\colon x \mapsto x$ pour tout $x \in C$, donc (c)⇒(b).

Soit enfin $F = \mathrm{Id}_C$ conformément à (b); alors

$$[(x_i\colon i \leqq n)] = x_1 . x_2 . \dots . x_n,$$

i. e., toute composition n-aire est exactement un n-chemin dans $[\mathbf{C}]$ et $\mathbf{C}_{[n]}$ est dérivée de la catégorie C, ce qui donne (b)⇒(a).

Si la loi $k'_n = [(\dots)]$ est partout définie, i. e., si $C \overset{n}{*} C = C^n$, on obtient les résultats classiques pour les n-demi-groupes [12] et pour les n-groupes [4].

BIBLIOGRAPHIE

[1] Белоусов, В. Д., *n-арные квазигруппы*, Изд. Штиинца, Кишинев, 1972.

[2] Dörnte, W., Untersuchungen über einen verallgemeinerten Gruppenbegriff, *Math. Z.* **29** (1928), 1—19.

[3] Глускин, Л. М., Позиционные оперативы, *Матем. Сб.* **68 (110)**, No. 3 (1965), 444—472.

[4] Hosszú, M., On the explicite form of n-group operation, *Publ. Math. Debrecen* **10** (1963), 88—92.

[5] MacLane, S., *Categories for the Working Mathematician,* Springer-Verlag, New York—Heidelberg—Berlin, 1971.

[6] Post, E. L., Polyadic Groups, *Trans. Amer. Math. Soc.* **48** (1940), 208—350.

[7] Topentcharov, V., Éléments de la théorie des catégories polyadiques, *Alg. Universalis* **7** (1977), 277—293.

[8] Topentcharov, V., *Theorie der* $\{n\}$*-kategorien,* T. H. Darmstadt, 1980.

[9] Topentcharov, V., Éléments de la théorie des catégories polyadiques, *CRAS de Bulgaire* **27** (1974), 743—746.

[10] Topentcharov, V., Sur les $[n]$-catégories, *Bul. de l'IPI Iaşi,* t. XXIV (XXVIII), f. 3—4 (1978), 13—18.

[11] Топенчаров, В., Елементи от теорията на n-арните категории, *Изд. на ВМЕИ «В. И. Ленин»*, София, 1975, 42 стр.

[12] Zupnik, D., Polyadic semigroups, *Publ. Math. Debrecen,* **14** (1967), 273—280.

Centre de Mathématiques Appliquées
B. P. 384, BG—1000 Sofia

(Reçu le 28 juillet 1977)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 253—258.

WEAKLY REGULAR AND τ-SMOOTH ZERO-ONE MEASURES

by

W. ADAMSKI

1. Introduction

Let $(X, \mathfrak{A})$ be a measurable space and let $\mathfrak{K}$ be a subset of $\mathfrak{A}$ satisfying certain closure properties. Under these assumptions, the author has proved in [1] that X is $\mathfrak{K}$-complete if and only if every 0,1-valued $\mathfrak{K}$-regular measure on $\mathfrak{A}$ is a Dirac measure. It is the purpose of the present paper to characterize in a similar way those measurable spaces on which all 0,1-valued weakly $\mathfrak{K}$-regular [(weakly) τ-smooth] measures are Dirac measures. We shall apply these results to topological situations. Among others, we shall prove by non-measuretheoretic methods that every 0,1-valued τ-smooth Baire [Borel] measure on a [Hausdorff] topological space is a Dirac measure.

2. Definitions and preliminaries

(a) $\mathbf{N}$ denotes the set of positive integers. For an arbitrary set X we denote by $\mathfrak{P}(X)$ the class of all subsets of X.

Let X be a topological space. A zero-set in X is a set of the form $\{f=0\}$ where f is a continuous real-valued function on X. We denote by $\mathfrak{Z}(X)$, $\mathfrak{F}(X)$, $\mathfrak{G}(X)$, $\mathfrak{K}(X)$ the collection of all zero-, closed, open, compact sets in X, respectively. The σ-algebra generated by $\mathfrak{Z}(X)[\mathfrak{F}(X)]$ is denoted by $\mathfrak{B}_0(X)[\mathfrak{B}(X)]$; its members are called the Baire [Borel] subsets of X.

(b) Let X be an arbitrary set and $\mathfrak{A}$ a paving on X (i.e. $\mathfrak{A}$ is a nonvoid subset of $\mathfrak{P}(X)$). A nonempty subset $\mathfrak{N}$ of $\mathfrak{A}$ is called an $\mathfrak{A}$-filter on X provided that

(i) $\emptyset \notin \mathfrak{N}$;

(ii) $N_1, N_2 \in \mathfrak{N}$ implies $N_1 \cap N_2 \in \mathfrak{N}$;

(iii) $\mathfrak{N} \ni N_1 \subseteq N_2 \in \mathfrak{A}$ implies $N_2 \in \mathfrak{N}$.

An $\mathfrak{A}$-filter $\mathfrak{N}$ on X is said

(i) to have the countable intersection property (cip) if $N_1, N_2, \ldots \in \mathfrak{N}$ implies $\bigcap_{k \in \mathbf{N}} N_k \neq \emptyset$;

(ii) to be fixed if $\bigcap \mathfrak{N} \neq \emptyset$;

(iii) to be an $\mathfrak{A}$-ultrafilter if $A \in \mathfrak{A}$ and $A \cap N \neq \emptyset$ for all $N \in \mathfrak{N}$ implies $A \in \mathfrak{N}$.

If each $\mathfrak{A}$-ultrafilter on X with cip is fixed, we say, following [3], that X is $\mathfrak{A}$-complete. In accordance with [3], $\mathfrak{B}(X)$-complete topological spaces are called Borel-complete.

We need the following result the simple proof of which is left to the reader:

LEMMA 2.1. *Let $\mathfrak{N}$ be an $\mathfrak{A}$-ultrafilter on X with cip. If (A_n) is a sequence in $\mathfrak{A}$ such that $\bigcup_{n \in \mathbf{N}} A_n \in \mathfrak{N}$ then $A_{n_0} \in \mathfrak{N}$ for some $n_0 \in \mathbf{N}$.*

(c) Let X be an arbitrary set and $\mathfrak{A}$ a paving on X. Then $R(\mathfrak{A})$, $S(\mathfrak{A})$, $\sigma(\mathfrak{A})$ denotes the ring, σ-ring, σ-algebra in X generated by $\mathfrak{A}$, respectively. Furthermore, $F(\mathfrak{A}):=\{Q\in\mathfrak{P}(X)\colon A\cap Q\in\mathfrak{A}$ for all $A\in\mathfrak{A}\}$. Finally, $\mathfrak{A}_\delta$ denotes the paving of all countable intersections of sets from $\mathfrak{A}$.

(d) A measure is a nonnegative finite countably additive set function defined on some σ-algebra. In particular, a Baire or Borel measure on a topological space X is a measure defined on $\mathfrak{B}_0(X)$ or $\mathfrak{B}(X)$, respectively. Furthermore, δ_x denotes the Dirac measure pertaining to the point $x\in X$.

Let $(X, \mathfrak{A})$ be a measurable space and $\mathfrak{K}$ a subpaving of $\mathfrak{A}$. A measure μ on $\mathfrak{A}$ is called

(i) 0,1-valued if $\mu\neq 0$ and $\mu(A)=0$ or 1 for all $A\in\mathfrak{A}$;
(ii) weakly $\mathfrak{K}$-regular if $\mu(X)=\sup\{\mu(K)\colon K\in\mathfrak{K}\}$;
(iii) $\mathfrak{K}$-regular if $\mu(A)=\sup\{\mu(K)\colon K\in\mathfrak{K},\ K\subseteq A\}$ for all $A\in\mathfrak{A}$;
(iv) weakly τ-smooth with respect to $\mathfrak{K}$ if $\inf\limits_\alpha \mu(K_\alpha)=0$ for each net (K_α) in $\mathfrak{K}$ such that $K_\alpha\downarrow\emptyset$;
(v) τ-smooth w.r.t. $\mathfrak{K}$ if $\inf\limits_\alpha \mu(K_\alpha)=\mu(K)$ for each net (K_α) in $\mathfrak{K}$ such that $K_\alpha\downarrow K\in\mathfrak{K}$.

A Borel measure on a topological space X is called [weakly] τ-smooth if it is [weakly] τ-smooth w.r.t. $\mathfrak{F}(X)$. A Baire measure on a topological space X is called τ-smooth if it is τ-smooth w.r.t. $\mathfrak{Z}(X)$. According to the $\mathfrak{Z}(X)$-regularity of all Baire measures on X (cf. [5], Theorem 18, Part I), a Baire measure on X is τ-smooth if and only if it is weakly τ-smooth w.r.t. $\mathfrak{Z}(X)$.

LEMMA 2.2. *Let μ be a Borel measure on a Hausdorff space X.*
(α) *If μ is $\mathfrak{F}(X)$-regular and weakly $\mathfrak{K}(X)$-regular then μ is $\mathfrak{K}(X)$-regular.*
(β) *If μ is weakly $\mathfrak{K}(X)$-regular then μ is weakly τ-smooth. The converse is true if X is locally compact.*
(γ) *If μ is* 0,1*-valued and $\mathfrak{K}(X)$-regular then μ is a Dirac measure.*

The simple proofs of (α) and (β) are left to the reader, while a proof of (γ) can be found in [1] 2.4.

3. The main results

THEOREM 3.1. *Let $(X, \mathfrak{A})$ be a measurable space, and assume that $\mathfrak{K}$, $\mathfrak{L}$ are two pavings on X satisfying the following conditions:*

(i) $\emptyset\in\mathfrak{K}\subseteq\mathfrak{L}\subseteq\mathfrak{A}=\sigma(\mathfrak{A}\cap F(\mathfrak{L}_\delta))$;
(ii) *$\mathfrak{L}$ is a ring;*
(iii) *every $L\in\mathfrak{L}$ is contained in a countable union of sets from $\mathfrak{K}$.*

Then the following two statements are equivalent:

(1) *X is $\mathfrak{L}$-complete.*
(2) *Every* 0,1*-valued weakly $\mathfrak{K}$-regular measure on $\mathfrak{A}$ is a Dirac measure.*

PROOF. (1)→(2). Let μ be a 0,1-valued weakly $\mathfrak{K}$-regular measure on $\mathfrak{A}$. Then there exists $K_0\in\mathfrak{K}$ such that $\mu(K_0)=1$. We shall show that $\mathfrak{N}:=\{L\in\mathfrak{L}\colon \mu(L)=1\}$

is an $\mathfrak{L}$-ultrafilter on X with cip. Let $L\in\mathfrak{L}$ be such that $L\cap N\neq\emptyset$ for all $N\in\mathfrak{N}$. Assume that $L\notin\mathfrak{N}$, hence $\mu(L)=0$. This implies $\mu(K_0-L)=1$, so $K_0-L\in\mathfrak{N}$ which contradicts $L\cap(K_0-L)=\emptyset$. It is easy to see that $\mathfrak{N}$ satisfies also the other conditions of an $\mathfrak{L}$-ultrafilter with cip. Thus, by (1), there exists $x_0\in\bigcap\mathfrak{N}$. It follows that $\mu(L)==\delta_{x_0}(L)$ for all $L\in\mathfrak{L}_\delta$. Now, let $A\in\mathfrak{A}\cap F(\mathfrak{L}_\delta)$ be given. If $x_0\in A$ then $x_0\in A\cap K_0\in\mathfrak{L}_\delta$, hence $\mu(A\cap K_0)=1$, and so $\mu(A)=1$. If, however, $x_0\notin A$ then $x_0\notin A\cap K_0\in\mathfrak{L}_\delta$ which implies $\mu(A\cap K_0)=0$, hence $\mu(A)=0$, as $\mu(K_0)=1$. Thus we have shown $\mu(A)=\delta_{x_0}(A)$ for all $A\in\mathfrak{A}\cap F(\mathfrak{L}_\delta)$ which implies $\mu=\delta_{x_0}$, since $\sigma(\mathfrak{A}\cap F(\mathfrak{L}_\delta))=\mathfrak{A}$.

(2)→(1). Let $\mathfrak{N}$ be an $\mathfrak{L}$-ultrafilter on X with cip. For any $L\in\mathfrak{L}$, put $\lambda(L)=1$ or $\lambda(L)=0$ according as $L\in\mathfrak{N}$ or $L\notin\mathfrak{N}$. Then λ satisfies the assumptions of [1] 1.2. Therefore λ can be extended to an $\mathfrak{L}_\delta$-regular measure μ defined on $\sigma(F(\mathfrak{L}_\delta))$. As $\mathfrak{N}\neq\emptyset$, the $\mathfrak{L}_\delta$-regularity of μ implies that μ is 0,1-valued. In particular, we have $\mu(L_0)=1$ for some $L_0\in\mathfrak{L}$. Since L_0 is covered by a countable family of sets from $\mathfrak{K}$, we have $\mu(K_0)=1$ for some $K_0\in\mathfrak{K}$. Therefore, by (2), the restriction of μ onto $\mathfrak{A}$ is a Dirac measure, say δ_{x_0}. This implies $x_0\in\bigcap\mathfrak{N}$.

COROLLARY 3.2. *For a Hausdorff space X the following three assertions are equivalent:*

(1) *X is $R(\mathfrak{K}(X))$-complete.*

(2) *X is $S(\mathfrak{K}(X))$-complete.*

(3) *Every 0,1-valued weakly $\mathfrak{K}(X)$-regular Borel measure on X is a Dirac measure.*

PROOF. It suffices to show that $\mathfrak{A}:=\mathfrak{B}(X)$, $\mathfrak{K}:=\mathfrak{K}(X)$, $\mathfrak{L}:=R(\mathfrak{K}(X))$, resp. $\mathfrak{L}:=S(\mathfrak{K}(X))$ satisfy the assumptions of 3.1. Only the verification of the equation $\mathfrak{A}=\sigma(\mathfrak{A}\cap F(\mathfrak{L}_\delta))$ is nontrivial.

(α) If $\mathfrak{L}=R(\mathfrak{K}(X))$ then it suffices to prove $\mathfrak{F}(X)\subseteq F(\mathfrak{L}_\delta)$. For this purpose let $F\in\mathfrak{F}(X)$ and $C\in\mathfrak{L}_\delta$ be given. Then $C=\bigcap_{n\in\mathbf{N}} C_n$ with $C_n\in R(\mathfrak{K}(X))$ for all $n\in\mathbf{N}$. As $C_n=\bigcup_{i=1}^{k_n}(K_i-L_i)$ with $K_i, L_i\in\mathfrak{K}(X), L_i\subseteq K_i$ for $i=1, \dots, k_n$ by [2], Theorem 1, Sect. 58, we obtain $F\cap C_n=\bigcup_{i=1}^{k_n}((F\cap K_i)-(F\cap L_i))$ with $F\cap K_i, F\cap L_i\in\mathfrak{K}(X)$ for $i=1, \dots, k_n$. This means $F\cap C_n\in R(\mathfrak{K}(X))$, $n\in\mathbf{N}$, hence $F\cap C\in\mathfrak{L}_\delta$. This proves $F\in F(\mathfrak{L}_\delta)$.

(β) In the case $\mathfrak{L}=S(\mathfrak{K}(X))$ it suffices to verify $\mathfrak{F}(X)\subseteq F(\mathfrak{L})$. Let $F\in\mathfrak{F}(X)$ be given. We consider the class $\mathfrak{S}_F:=\{C\in\mathfrak{L}\colon F\cap C\in\mathfrak{L}\}$. Then $R(\mathfrak{K}(X))\subseteq\mathfrak{S}_F$, as we have seen in part (α). Since $\mathfrak{S}_F$ is a monotone class, we obtain $\mathfrak{S}_F=\mathfrak{L}$ which implies $F\in F(\mathfrak{L})$.

REMARKS 3.3. The space X of all ordinals less than or equal to the first uncountable ordinal, equipped with the order topology, is a compact Hausdorff space which is not Borel-complete (cf. [1] 2.7b) and thus not $R(\mathfrak{K}(X))$-complete by 3.2. On the other hand, every Hausdorff space X with the property that every 0,1-valued Borel measure on X is $\mathfrak{F}(X)$-regular, is $R(\mathfrak{K}(X))$-complete according to 2.2 (α), (γ) and 3.2. In particular, every perfectly normal Hausdorff space X is $R(\mathfrak{K}(X))$-complete. Consequently, the perfectly normal space L constructed in example 45 of [4] is an example of an $R(\mathfrak{K}(X))$-complete space which is not $\mathfrak{F}(X)$-complete and therefore neither realcompact nor Borel-complete (cf. [1] 2.7).

THEOREM 3.4. *Let $(X, \mathfrak{A})$ be a measurable space, and assume that $\mathfrak{K}$, $\mathfrak{L}$ are two subpavings of $\mathfrak{A}$ satisfying the following conditions:*

(i) $X \in \mathfrak{K} \subseteq \mathfrak{L}$;

(ii) *$\mathfrak{K}$ is closed under finite intersections;*

(iii) *$\mathfrak{L}$ is a ring generating $\mathfrak{A}$.*

Then the following two statements are equivalent:

(1) *Every $\mathfrak{L}$-ultrafilter $\mathfrak{N}$ on X with cip and $\bigcap(\mathfrak{N} \cap \mathfrak{K}) \neq \emptyset$ is fixed.*

(2) *Every* 0,1-*valued measure on $\mathfrak{A}$ being weakly τ-smooth w.r.t. $\mathfrak{K}$ is a Dirac measure.*

PROOF. (1)→(2). Let μ be a 0,1-valued measure on $\mathfrak{A}$ being weakly τ-smooth w.r.t. $\mathfrak{K}$. Setting $\mathfrak{N} := \{L \in \mathfrak{L}\colon \mu(L) = 1\}$ one can easily verify that $\mathfrak{N}$ is an $\mathfrak{L}$-ultrafilter on X with cip. Assume that $\bigcap(\mathfrak{N} \cap \mathfrak{K}) = \emptyset$. Then the weak τ-smoothness of μ implies $\inf\{\mu(N)\colon N \in \mathfrak{N} \cap \mathfrak{K}\} = 0$ which contradicts the definition of $\mathfrak{N}$. Thus, by (1), there exists some $x_0 \in \bigcap \mathfrak{N}$. This implies $\mu(L) = \delta_{x_0}(L)$ for all $L \in \mathfrak{L}$, hence $\mu = \delta_{x_0}$, since $\sigma(\mathfrak{L}) = \mathfrak{A}$.

(2)→(1). Let $\mathfrak{N}$ be an $\mathfrak{L}$-ultrafilter on X with cip and $\bigcap(\mathfrak{N} \cap \mathfrak{K}) \neq \emptyset$. For any $L \in \mathfrak{L}$, put $\lambda(L) = 1$ or $\lambda(L) = 0$ according as $L \in \mathfrak{N}$ or $L \notin \mathfrak{N}$. The set function λ is additive. Furthermore, $\inf \lambda(L_n) = 0$ for each sequence (L_n) in $\mathfrak{L}$ such that $L_n \downarrow \emptyset$. Thus, by Carathéodory's theorem, λ can be extended in a unique way to a 0,1-valued measure μ defined on $\mathfrak{A} = \sigma(\mathfrak{L})$. In order to prove that μ is weakly τ-smooth w.r.t. $\mathfrak{K}$, let (K_α) be a net in $\mathfrak{K}$ such that $K_\alpha \downarrow \emptyset$. Assume that $\inf\limits_{\alpha} \mu(K_\alpha) > 0$, i.e. $\mu(K_\alpha) = 1$ for all α. Then $(K_\alpha) \subseteq \mathfrak{N} \cap \mathfrak{K}$ which contradicts $\bigcap(\mathfrak{N} \cap \mathfrak{K}) \neq \emptyset$. Thus, by (2), $\mu = \delta_{x_0}$ for some $x_0 \in X$. This implies $x_0 \in \bigcap \mathfrak{N}$.

COROLLARY 3.5. *For a topological space X the following three assertions are equivalent:*

(1) *Every $R(\mathfrak{F}(X))$-ultrafilter $\mathfrak{N}$ on X with cip and $\bigcap(\mathfrak{N} \cap \mathfrak{F}(X)) \neq \emptyset$ is fixed.*

(2) *Every $\mathfrak{B}(X)$-ultrafilter $\mathfrak{N}$ on X with cip and $\bigcap(\mathfrak{N} \cap \mathfrak{F}(X)) \neq \emptyset$ is fixed.*

(3) *Every* 0,1-*valued weakly τ-smooth Borel measure on X is a Dirac measure.*

The following corollary is a generalization of [5], Corollary 3)., Theorem 25, Part I to arbitrary topological spaces. However, the analogous result for Borel measures is only valid for Hausdorff spaces (cf. 3.10 and 3.11).

COROLLARY 3.6. *Every* 0,1-*valued τ-smooth Baire measure on a topological space X is a Dirac measure.*

PROOF. In view of 3.4 it suffices to show that every $R(\mathfrak{Z}(X))$-ultrafilter $\mathfrak{N}$ on X with cip is fixed provided that $\bigcap(\mathfrak{N} \cap \mathfrak{Z}(X)) \neq \emptyset$. For this purpose it is sufficient to prove that every $N \in \mathfrak{N}$ contains some element of $\mathfrak{N} \cap \mathfrak{Z}(X)$. Let $N \in \mathfrak{N}$ be given. Then $N = \bigcup\limits_{i=1}^{n} (K_i - L_i)$ with $K_i, L_i \in \mathfrak{Z}(X)$ for $i = 1, \dots, n$ by [2], Theorem 1, Sect. 58. Thus, by 2.1, there exists $i_0 \in \{1, \dots, n\}$ such that $K_{i_0} - L_{i_0} \in \mathfrak{N}$, hence $K_{i_0} \in \mathfrak{N} \cap \mathfrak{Z}(X)$ and $X - L_{i_0} \in \mathfrak{N}$. Furthermore, we have $X - L_{i_0} = \bigcup\limits_{n \in \mathbf{N}} Z_n$ for an appropriate sequence (Z_n) in $\mathfrak{Z}(X)$. Again by 2.1, there exists $n_0 \in \mathbf{N}$ such that $Z_{n_0} \in \mathfrak{N}$. Thus $K_{i_0} \cap Z_{n_0} \in \mathfrak{N} \cap \mathfrak{Z}(X)$ and $K_{i_0} \cap Z_{n_0} \subseteq N$.

REMARKS 3.7. For abbreviation, let us call τ-pseudocomplete any topological space X satisfying the (equivalent) conditions (1)—(3) of 3.5. According to 2.2 (β), 3.2 and 3.5, any τ-pseudocomplete Hausdorff space X is $R(\mathfrak{K}(X))$-complete. Conversely, an $R(\mathfrak{K}(X))$-complete Hausdorff space X is τ-pseudocomplete if every 0,1-valued weakly τ-smooth Borel measure on X is weakly $\mathfrak{K}(X)$-regular. Thus, by 2.2 (β), for locally compact Hausdorff spaces X, τ-pseudocompleteness and $R(\mathfrak{K}(X))$-completeness are equivalent properties.

THEOREM 3.8. *Let $(X, \mathfrak{A})$ be a measurable space, and assume that $\mathfrak{K}$, $\mathfrak{L}$ are two subpavings of $\mathfrak{A}$ satisfying the following conditions:*

(i) $X\in\mathfrak{K}\subseteqq\mathfrak{L}$;

(ii) *$\mathfrak{K}$ is closed under arbitrary intersections;*

(iii) *$\mathfrak{L}$ is a ring generating $\mathfrak{A}$.*

Then the following two statements are equivalent:

(1) *Every $\mathfrak{L}$-ultrafilter $\mathfrak{N}$ on X with cip and $\bigcap(\mathfrak{N}\cap\mathfrak{K})\in\mathfrak{N}$ is fixed.*

(2) *Every* 0,1-*valued measure on $\mathfrak{A}$ being τ-smooth w.r.t. $\mathfrak{K}$ is a Dirac measure.*

PROOF. (1)→(2). If μ is a 0,1-valued measure on $\mathfrak{A}$ being τ-smooth w.r.t. $\mathfrak{K}$ and if $\mathfrak{N}:=\{L\in\mathfrak{L}\colon \mu(L)=1\}$ then $\mathfrak{N}$ is an $\mathfrak{L}$-ultrafilter on X with cip. Assume that $K_0:=\bigcap(\mathfrak{N}\cap\mathfrak{K})\notin\mathfrak{N}$, hence $K_0\in\mathfrak{L}-\mathfrak{N}$ and therefore $\inf\{\mu(N)\colon N\in\mathfrak{N}\cap\mathfrak{K}\}=\mu(K_0)=0$, which contradicts the definition of $\mathfrak{N}$. Now one can proceed as in the proof of 3.4.

(2)→(1). Let $\mathfrak{N}$ be an $\mathfrak{L}$-ultrafilter on X with cip and $\bigcap(\mathfrak{N}\cap\mathfrak{K})\in\mathfrak{N}$. As we have seen in the proof of 3.4, there exists a 0,1-valued measure μ on $\mathfrak{A}$ such that for $L\in\mathfrak{L}$ $\mu(L)=1$ or $\mu(L)=0$ according as $L\in\mathfrak{N}$ or $L\notin\mathfrak{N}$. In order to prove that μ is τ-smooth w.r.t. $\mathfrak{K}$, let (K_α) be a net in $\mathfrak{K}$ such that $K_\alpha\downarrow K\in\mathfrak{K}$. Assume that $\mu(K)<\inf_\alpha\mu(K_\alpha)$, i.e. $\mu(K)=0$ and $\mu(K_\alpha)=1$ for all α, hence $(K_\alpha)\subseteqq\mathfrak{N}\cap\mathfrak{K}$ and therefore $\mathfrak{N}\ni\bigcap(\mathfrak{N}\cap\mathfrak{K})\subseteqq K$ which implies $K\in\mathfrak{N}$, in contrast to $\mu(K)=0$. Thus, by (2), $\mu=\delta_{x_0}$ for some $x_0\in X$. This implies $x_0\in\bigcap\mathfrak{N}$.

COROLLARY 3.9. *For a topological space X the following three assertions are equivalent:*

(1) *Every $R(\mathfrak{F}(X))$-ultrafilter $\mathfrak{N}$ on X with cip and $\bigcap(\mathfrak{N}\cap\mathfrak{F}(X))\in\mathfrak{N}$ is fixed.*

(2) *Every $\mathfrak{B}(X)$-ultrafilter $\mathfrak{N}$ on X with cip and $\bigcap(\mathfrak{N}\cap\mathfrak{F}(X))\in\mathfrak{N}$ is fixed.*

(3) *Every* 0,1-*valued τ-smooth Borel measure on X is a Dirac measure.*

COROLLARY 3.10. *Every* 0,1-*valued τ-smooth Borel measure on a Hausdorff space X is a Dirac measure.*

PROOF. In view of 3.9 it suffices to show that every $R(\mathfrak{F}(X))$-ultrafilter $\mathfrak{N}$ on X with cip is fixed provided that $\bigcap(\mathfrak{N}\cap\mathfrak{F}(X))\in\mathfrak{N}$. We shall prove that the set $F_0:=\bigcap(\mathfrak{N}\cap\mathfrak{F}(X))$, belonging to $\mathfrak{N}$, is a singleton, say $\{x_0\}$, which implies $x_0\in\bigcap\mathfrak{N}$. Assume that x, y are two different points in F_0. As X is Hausdorff, there are disjoint open sets G_1, G_2 such that $x\in G_1$ and $y\in G_2$. At least one of them, say G_1, does not pertain to $\mathfrak{N}$. Thus $X-G_1\in\mathfrak{N}\cap\mathfrak{F}(X)$ which implies $x\in F_0\subseteqq X-G_1$, in contrast to $x\in G_1$.

The following example shows that 3.10 does not remain true for T_1-spaces.

Example 3.11. Let an uncountable set X be equipped with the cofinite topology, i.e. $\mathfrak{G}(X)=\{\emptyset\}\cup\{G\in\mathfrak{P}(X)\colon X-G \text{ finite}\}$. X is a compact T_1- (but not a Hausdorff) space. For any $B\in\mathfrak{B}(X)=\{B\in\mathfrak{P}(X)\colon B \text{ or } X-B \text{ countable}\}$, put $\mu(B)=0$ or $\mu(B)=1$ according as B or $X-B$ is countable. Then μ is a 0,1-valued τ-smooth Borel measure; however, μ is not a Dirac measure. On the other hand, one can easily verify that every 0,1-valued $\mathfrak{F}(X)$-regular Borel measure on X is a Dirac measure; thus, by [1] 2.5, X is $\mathfrak{F}(X)$-complete.

REFERENCES

[1] Adamski, W., Complete spaces and zero-one measures, *Manuscripta mathematica* **18** (1976), 343—352.

[2] Berberian, S. K.. *Measure and integration,* New York, Chelsea, 1970.

[3] Hager, A. W., Reynolds, G. D., Rice, M. D., Borel-complete topological spaces, *Fund. Math.* **75** (1972), 135—143.

[4] Steen, L. A., Seebach, J. A.: *Counterexamples in topology,* New York, Holt, Rinehart and Winston, 1970.

[5] Varadarajan, V. S., Measures on topological spaces, *Amer. Math. Soc. Transl.* (2) **48** (1965), 161—228.

Mathematisches Institut der Universität München, Theresienstraße 39,
D—8000 München 2, West Germany

(Received October 14, 1977)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 259—271.

INDICE D'INERTIE TRILATÈRE ET CLASSES DE CHERN

par
ALBERT CRUMEYROLLE

Résumé

On définit d'abord l'indice d'inertie trilatère qui s'identifie à celui de LERAY [6] seulement quand les lagrangiens sont deux à deux transverses. Cet indice définit un cocycle. On montre que la décomposition de l'indice trilatère à l'aide de l'indice de MASLOV traduit la finesse d'un faisceau et en est une conséquence immédiate.

Dans la partie II on établit que la première classe de CHERN d'une variété presque symplectique est localement représentée par le cocycle d'inertie d'un triplet de lagrangiens deux à deux transverses.

Dans la partie III la recherche d'une généralisation de la propriété établie dans II nous conduit à une construction nouvelle (du moins à notre connaissance) des classes de Chern. On généralise II et finalement on montre que notre définition équivaut à celle de CHERN—WEIL au moyen du théorème de DE RHAM.

I. Indice d'inertie trilatère et indice de Maslov

1° E est un espace vectoriel réel de dimension $2n$, muni d'une forme symplectique F.

LEMME 1. *Soient trois espaces lagrangiens L, L', L'' tels que $L\cap L'=L\cap L''$. Alors il existe σ, en général non unique, appartenant au groupe symplectique* $\mathrm{Sp}(2n, \mathbf{R})$ *qui laisse fixe tous les éléments de L et applique L' sur L''.*

Soit $\{e_\alpha, e_{\beta*}\}$, $\alpha, \beta=1, 2, \dots, n$, une base symplectique de E dont les vecteurs (e_α) sont dans L. Il est immédiat de voir qu'un tel σ s'exprime par

$$\text{(1)}\qquad \begin{cases} \sigma(e_\alpha) = e_\alpha \\ \sigma(e_{\alpha*}) = a^\beta_{\alpha*} e_\beta + e_{\alpha*}, \quad \text{avec} \quad a^\beta_{\alpha*} = a^\alpha_{\beta*}. \end{cases}$$

Il est loisible de supposer que les e_α ont été choisis de manière que $(e_1, e_2, \dots, e_\lambda)$ engendrent $L\cap L'$. Alors $e_{\lambda+1}, \dots, e_n, e_{1*}, \dots, e_{\lambda*}$ engendrent un lagrangien L_1 avec $L_1\cap L'=L_1\cap L''=0$. Soit L_2 le lagrangien de base $(e_1, e_2, \dots, e_\lambda, e_{(\lambda+1)*}, \dots, e_{n*})$, il est immédiat que

$$E = L_1 \oplus L' = L_1 \oplus L'' = L_1 \oplus L_2 \quad \text{et} \quad L\cap L' = L\cap L'' = L\cap L_2.$$

Or il existe une transformation symplectique θ qui conserve L_1 point par point, les vecteurs $e_1, e_2, \dots, e_\lambda$, et envoie L' sur L_2 (envoyer la base ordonnée $(e_{\lambda+1}, \dots, e_n,$

AMS (MOS) subject classifications (1970). Primary 53C15.

Key words and phrases. Global differential geometry, symplectic, Lagrangians, trilateral index, Leray index, Maslov index, Chern classes, symplectic spinor structures, splitting map, fibre bundles, De Rham abstract theorem, cohomology.

$e_{1^*}, \dots, e_{\lambda^*}$) de L_1 sur elle-même et une base ordonnée $(e_1, e_2, \dots, e_\lambda, f_{(\lambda+1)}, \dots, f_n)$ de L' sur la base ordonnée $(e_1, e_2, \dots, e_\lambda, e_{(\lambda+1)^*}, \dots, e_{n^*})$ de L_2). Si θ' envoie de la même manière L'' sur L_2, alors $\sigma=\theta'^{-1}\circ\theta$ répond à la question posée.

Si σ' possède la même propriété $\sigma'^{-1}\circ\sigma$ conserve globalement L' et L point par point. Si $L\cap L'$ est non nul, on voit que σ n'est pas unique. *Par contre si* $L\cap L'=$ $=L\cap L''=0$, (1) *montre l'unicité et établit une correspondance bijective entre l'ensemble des lagrangiens transverses à L et l'ensemble des matrices symétriques $n\times n$.*

2° *Indice d'inertie de trois lagrangiens ordonnés L, L', L'' tels que* $L\cap L'=$ $=L\cap L''=0$.

Avec les hypothèses de lemme 1, prenons dans L' la base $E_i=A_i^{\alpha^*}e_{\alpha^*}+B_i^\alpha e_\alpha$, $i=1, \dots, n$. Si $x\in L'$, formons $Q(x)=F(x, \sigma(x))$. $F(\cdot, \sigma(\cdot))$ est une forme quadratique sur L', car

$$Q_{ij} = F(E_i, \sigma(E_j)) = ({}^tAaA)_{ij}, \quad \text{avec} \quad A = \|A_i^{\alpha^*}\|, \quad a = \|a^\alpha_{\beta^*}\|.$$

Soit $Q_{ij}=\sum\limits_{\alpha,\sigma} A_i^{\alpha^*}A_j^{\sigma^*}a^\alpha_{\sigma^*}$.

σ étant choisi, cette forme quadratique est bien déterminée. Il est toujours possible de choisir (e_α, e_{β^*}) de manière que

$$\text{(2)} \qquad \begin{cases} \sigma(e_\alpha) = e_\alpha \\ \sigma(e_{\alpha^*}) = t_\alpha e_\alpha + e_{\alpha^*}, \quad t = 0,\ 1 \text{ ou } -1, \end{cases}$$

de sorte que $({}^tAaA)_{ij}=\sum\limits_\alpha t_\alpha A_i^{\alpha^*}A_j^\alpha$.

Il faut observer que la signature de Q dépend du choix de σ, si $L\cap L'=$ $=L\cap L''\neq 0$. En effet soit

$$L = (e_1, e_2, \dots, e_\lambda, e_{\lambda+1}, \dots, e_n),$$
$$L' = (e_1, e_2, \dots, e_\lambda, e_{(\lambda+1)^*}, \dots, e_{n^*}),$$
$$\theta(e_\alpha) = e_\alpha, \quad \theta(e_{\alpha^*}) = e_{\alpha^*}, \quad \alpha = 1, 2, \dots, n,$$

d'une part, et d'autre part

$$\theta'(e_\alpha) = e_\alpha, \quad \alpha = 1, 2, \dots, n,$$
$$\theta'(e_{\alpha^*}) = e_{\alpha^*}, \quad \alpha = 1, 2, \dots, \lambda, \quad \theta'(e_{\alpha^*}) = \sum_1^\lambda a^\beta_{\alpha^*}e_\beta + e_{\alpha^*}, \quad \alpha = \lambda+1, \dots, n,$$

θ et θ' conservent L point par point et globalement $L'=L''$, en donnant des signatures distinctes.

Si $L\cap L'=L\cap L''=0$, il existe $\sigma\in\mathrm{Sp}(2n, \mathbf{R})$ unique, qui conserve L point par point et envoie L' sur L'', de sorte que l'on peut donner la définition suivante:

Définition 1. Si $L\cap L'=L\cap L''=0$, et si $\sigma\in\mathrm{Sp}(2n, \mathbf{R})$ conserve L point par point et envoie L' sur L'', la forme quadratique $x\to F(x, \sigma(x))$, $x\in L'$, admet une signature notée: Inert (L', L, L''). La signature est le nombre de signes $(-)$ dans la décomposition de Sylvester. Inert est l'indice d'inertie de Leray [6].

(2) montre que si $r=\dim L'\cap L''$, on peut définir:

$$\text{(3)} \qquad \mathrm{Inert}(L'', L, L') = -\mathrm{Inert}(L', L, L'')+(n-r).$$

L' et L'' étant transverses à L fixé, le choix de (e_α, e_{β^*}) identifie L' et L'' à des matrices symétriques.

Si le triplet (L', L, L'') se déforme continûment, L' et L'' restant transverses à L et $\dim(L' \cap L'')$ demeurant constante, il est immédiat que Inert (L', L, L'') garde une valeur constante.

Observons que si $r=0$, alors

$$e_{\alpha^*} \to e_{\alpha^*}, \quad e_\alpha \to -e_\alpha - t_\alpha e_{\alpha^*}$$

définit une transformation symplectique qui conserve L' point par point et envoie L sur L''; *donc quand* $L' \cap L'' = L \cap L' = L \cap L'' = 0$, *on peut poser:*

$$\text{Inert}(L, L', L'') = -\text{Inert}(L', L, L'') + n$$

et Inert *est invariant par permutation circulaire sur* L, L', L''.

Si $L \cap L' = L \cap L'' = 0$, avec $L' \cap L'' \neq 0$, nous poserons par convention

$$\text{(4)} \quad \begin{aligned} \text{Inert}(L, L', L'') &= -\text{Inert}(L', L, L'') + n \\ \text{Inert}(L', L'', L) &= -\text{Inert}(L', L, L'') + n. \end{aligned}$$

On notera que notre définition d'Inert n'est plus alors identique à celle de Leray, à moins que les trois lagrangiens ne soient deux à deux transverses. En raison de cette modification, l'indice que nous venons de définir sera appelé *indice d'inertie trilatère.*

3° *Indice d'inertie trilatère d'un triplet ordonné quelconque de lagrangiens.*

LEMME 2. *Il existe un lagrangien M transverse simultanément à trois lagrangiens L, L', L'' donnés.*

Selon SOURIAU [7] on peut identifier un lagrangien L à l'image de $a \in U(n)$ par $a \to a\bar{a}^{-1} = l$ et L transverse à M équivaut à $(l-m)$ inversible, si la matrice m représente M. Le lemme est alors immédiat; observons qu'il y a une infinité de choix possibles pour M.[1]

LEMME 3. *Si L_1, L_2, L_3, L_4 sont des lagrangiens tels que L_2 et L_3 soient transverses aux trois autres:*

$$\text{(5)} \quad \text{Inert}(L_1, L_2, L_3) - \text{Inert}(L_1, L_2, L_4) + \text{Inert}(L_1, L_3, L_4) - \text{Inert}(L_2, L_3, L_4) = 0.$$

En effet, selon les remarques faites au (2°), L_2 et L_3 étant transverses à la fois à L_1 et L_4 et

$$\text{Inert}(L_1, L_2, L_3) = \text{Inert}(L_3, L_1, L_2)$$

$$\text{Inert}(L_2, L_3, L_4) = \text{Inert}(L_3, L_4, L_2)$$

les quatre termes de (5) sont égaux deux à deux.

[1] Il serait aussi possible de donner une démonstration indépendante à l'aide des spineurs symplectiques [3a].

Si maintenant L_1, L_2, L_3 sont des lagrangiens quelconques, nous choisissons M transverse à chacun d'eux, alors conformément à (5) nous posons

$$(6)\quad \mathrm{Inert}\,(L_1, L_2, L_3) = \mathrm{Inert}\,(L_1, L_2, M) - \mathrm{Inert}\,(L_1, L_3, M) + \mathrm{Inert}\,(L_2, L_3, M).$$

Comme les trois indices d'inerties du deuxième membre sont constants quand M varie en restant transverse à L_1, L_2, L_3 fixés, la valeur du 1er membre ne dépend pas du choix de M. Il est rappelé que Inert $(L_1, L_2, M) = -$Inert $(L_1, M, L_2)+n$ où les termes du 2ème membre sont calculés selon la définition 1.

En général l'indice trilatère n'est pas invariant par permutation circulaire. Par contre (6) *montre que 4 lagrangiens quelconques L_1, L_2, L_3, L_4 satisfont une relation de cocyclicité sur laquelle on reviendra plus bas.*

4° *L'indice trilatère et la décomposition d'un élément du groupe symplectique en produit de transvections.*

Soit σ l'élément unique qui conserve L point par point et envoie L' sur L'', L' et L'' transverses à L. Introduisons ici l'algèbre de Clifford symplectique [3a]. Au dessus de σ on peut déterminer $\gamma \in G_s$, groupe de Clifford symplectique, modulo un scalaire, tel que

$$p(\gamma) = \sigma \quad \text{par projection:} \quad p(\gamma)(x) = \gamma x \gamma^{-1}, \quad x \in E.$$

Lemme 4. *L'ensemble des éléments $\sigma \in \mathrm{Sp}(2n, \mathbf{R})$ qui conservent L point par point s'identifie à l'ensemble $p\left(\exp\left(\overset{2}{\vee} L\right)\right)$.*

Ce lemme résulte de la formule (1) et des remarques suivantes:

Si $p(\gamma) = \sigma$ conserve L point par point il se factorise en produit de transvections symplectiques.

Si $a \in L$, $\exp\dfrac{ta^2}{2}\, x \exp\left(-\dfrac{ta^2}{2}\right) = x + F(a, x)\, ta$, $t \in \mathbf{R}^*$, et on peut choisir a de manière que $t = \pm 1$. Prenant

$$\gamma = \prod_i \exp\frac{t_i (a_i)^2}{2} = \exp\left(\sum_i \frac{t_i (a_i)^2}{2}\right), \quad a_i \in L, \ t_i = \pm 1,$$

un choix convenable des a_i, t_i permet d'atteindre tout élément σ défini par la formule (1). En particulier on voit que l'on peut choisir la base symplectique (e_α, e_{β^*}) de manière que σ se traduise par la formule (2) on aura alors

$$\gamma = \prod_\alpha \exp\left(\frac{t_\alpha (e_\alpha)^2}{2}\right),$$

ainsi: *Tout $\gamma \in G_s$ qui conserve L point par point peut se factoriser en*

$$(7)\qquad \gamma = \prod_\alpha \exp\frac{t_\alpha (e_\alpha)^2}{2}, \quad t_\alpha = \pm 1$$

la suite des (t_α) est intrinsèquement attachée à γ et si $L' \oplus L = E$, $L'' = \gamma(L')$ la signature de cette suite est l'indice d'inertie de Leray: Inert (L', L, L'').

Remarque. Soient $L_2\oplus L_2'=L_3\oplus L_3'=E, L_2, L_2', L_3, L_3'$ quatre lagrangiens et deux bases symplectiques adaptées à ces décompositions. L'une de ces bases étant fixée, $\mathrm{Sp}(2n, \mathbf{R})$ est en bijection avec l'ensemble des bases symplectiques de E. Construisons L_1 lagrangien transverse à L_2, L_2', L_3, L_3'. Il existe $\sigma_1\in \mathrm{Sp}(2n, \mathbf{R})$ conservant L_1 et envoyant L_2 sur L_3, L_2' sur L_2'' et σ_2 conservant L_3 et envoyant L_2'' sur L_3', donc $\sigma_2\circ\sigma_1$ envoie $L_2\oplus L_2'$ sur $L_3\oplus L_3'$, ainsi:

Toute transformation symplectique est le produit de deux transformations symplectiques conservant chacune un lagrangien point par point, donc le produit de $2n$ transvections symplectiques au plus σ_{a_i,t_i}, avec a_i isotrope $t_i=\pm 1$.

5° *La signification cohomologique de l'indice trilatère et de l'indice de Maslov.*

Considérons une variété réelle V (ou même un espace topologique) paracompacte, de dimension $2n$, à structure presque symplectique et le préfaisceau $\mathscr{P}_{\mathbf{Z}}$ des applications quelconques d'ouverts de V dans $\mathbf{Z}$. Si Pr et Fs désignent respectivement les foncteurs "préfaisceau et "faisceau", construisons $\mathrm{Fs}\,\mathscr{P}_{\mathbf{Z}}=\mathscr{F}_{\mathbf{Z}}$; selon un résultat classique ce faisceau est fin, donc compte-tenu de la paracompacticité de V

$$(8)\qquad H^q(V, \mathscr{F}_{\mathbf{Z}}) \simeq H^q(V, \mathrm{Pr}\,\mathrm{Fs}\,\mathscr{P}_{\mathbf{Z}}) \simeq H^q(V, \mathscr{P}_{\mathbf{Z}}) = 0, \quad \text{pour} \quad q \geqq 1.$$

Si on introduit un recouvrement de V par des ouverts U_α domaines de définition de sections lagrangiennes L_α: $x\in U_\alpha\to L_\alpha(x)$, suffisamment petits pour être munis de sections λ_α: $x\to\lambda_\alpha(x)$ dans le revêtement de la grassmannienne lagrangienne, on peut définir un cocycle par

$$x\in U_\alpha\cap U_\beta\cap U_\gamma \to \mathrm{Inert}\big(L_\alpha(x), L_\beta(x), L_\gamma(x)\big),$$

encore noté $\mathrm{Inert}\big(\lambda_\alpha(x), \lambda_\beta(x), \lambda_\gamma(x)\big)$. Mais en raison de (8) ce cocycle est un cobord, il existe donc m:

$$x\in U_\alpha\cap U_\beta \to m\big(\lambda_\alpha(x), \lambda_\beta(x)\big)\in\mathbf{Z}$$

tel que

$$\mathrm{Inert}(\lambda_\alpha, \lambda_\beta, \lambda_\gamma) = m(\lambda_\alpha, \lambda_\beta)+m(\lambda_\beta, \lambda_\gamma)-m(\lambda_\alpha, \lambda_\gamma).$$

Ainsi: *La décomposition du cocycle d'inertie au moyen de l'indice de Maslov, donnée par Leray* [6], *résulte de la trivialité de la cohomologie $H^q(V, \mathscr{P}_{\mathbf{Z}})$, pour $q\geqq 1$.*

Remarques. 1° Introduisant les divers revêtements du groupe symplectique et de la grassmannienne lagrangienne, il est évident qu'on obtient des résultats analogues avec des indices mod q, $q\in\mathbf{Z}$.

2° L'interprétation cohomologique met en évidence certaines propriétés d'invariance homotopique.

3° On notera que les démonstrations que précèdent ne font intervenir que la structure symplectique, le rôle que joue le groupe unitaire dans certains exposés est donc inessentiel et une simple commodité de preuves.

4° Il est immédiat que l'on pourrait développer une théorie analogue en remplaçant la signature par le rang.

II. Cocycle trilatère et première classe de Chern

1° *Préambule et rappels*. Soit une variété V, réelle, paracompacte, de dimension $n=2r$, à structure presque symplectique. La donnée d'une telle variété équivaut à celle d'une structure presque complexe, car on sait qu'étant donnée une 2-forme extérieure F de rang $2r$, on peut toujours construire, modulo une homotopie, à partir d'une métrique riemannienne arbitraire une autre métrique échangeable avec F et par suite une structure hermitienne admettant F comme 2-forme fondamentale.

De manière indépendante soit un fibré en droites complexes (f. d. c.) au-dessus d'une variété V, les $a_{\alpha\beta}$ étant les fonctions de transitions pour un système d'ouverts de trivialisations (U_α) formant un atlas contractile, posant $f_{\alpha\beta}=\frac{1}{2i\pi}\log a_{\alpha\beta}$, alors

$$c_{\alpha\beta\gamma}=f_{\alpha\beta}+f_{\beta\gamma}-f_{\alpha\gamma}$$

définit un 2-cocycle de ČECH à valeurs dans $\mathbf{Z}$. A l'élément de $H^1(V, \mathbf{C}^*)$ qui définit le f. d. c. on peut associer biunivoquement un élément de $H^2(V, \mathbf{Z})$. En particulier s'il s'agit du f. d. c. associé à un fibré complexe ξ, au moyen des déterminants des matrices unitaires de changements de trivialisations locales, l'élément de $H^2(V, \mathbf{Z})$ obtenu est la première classe de Chern $c_1(\xi)$ de ξ.

2° Sur la variété V du préambule, dont le fibré tangent est muni de la structure complexe J, considérons un recouvrement par des ouverts (U_α), domaines de définition de sections lagrangiennes réelles: $x\to(L_\alpha)_x$. Munissons $(L_\alpha)_x$ d'un repère orthogonal $\{\xi_\alpha\}_x$ alors le système des sections locales $(\xi_\alpha, J\xi_\alpha)$ détermine au-dessus de chaque (U_α) un repère de $T(V)$. Envoyant les $(\xi_\alpha, J\xi_\alpha)$, respectivement, sur les $(\xi_\beta, J\xi_\beta)$, quand $U_\alpha\cap U_\beta\neq\emptyset$, on obtient un cocycle à valeurs dans $U(r, \mathbf{C})$; les $u_{\alpha\beta}(x)$ étant fonctions de transitions en $x\in U_\alpha\cap U_\beta$, posons

$$c_{\alpha\beta}(x)=\det u_{\alpha\beta}(x)$$

et introduisons comme plus haut le cocycle des $c_{\alpha\beta\gamma}$, que nous appellerons aussi *trilatère*.

Les considérations qui suivent sont d'abord purement algébriques. Soient L, L', L'' des lagrangiens deux à deux transverses. Il est possible de munir L et L' de bases $\{(e_\alpha), (e_{\beta^*})\}$, $\alpha, \beta=1, \dots, n$, constituant une base symplectique adaptée à la somme directe $E=L\oplus L'$, et L'' d'une base $\{f_\alpha\}$ telle que

$$f_\alpha=e_\alpha+t_\alpha e_{\alpha^*}, \quad t_\alpha\neq 0,$$

où la signature de la suite des t_α: $\operatorname{sgn}(t_\alpha)$ représente l'indice d'inertie de Leray, Inert (L, L', L'') [6] [3]. L'indice d'inertie est invariant par transformation symplectique, on peut donc supposer que l'on fait une telle transformation de manière que $\{(e_\alpha, e_{\beta^*})\}$ constitue une base orthogonale. Alors $\left\{F_\alpha=\frac{e_\alpha+t_\alpha e_{\alpha^*}}{\sqrt{1+(t_\alpha)^2}}\right\}$ est aussi une base orthogonale de L''.

Nous envoyons L sur L' par J, J considérée comme une application unitaire, $\det J=\exp\left(\frac{i\pi r}{2}\right)$ et nous prenons $\log(\det J)=\frac{i\pi r}{2}$. Nous envoyons ensuite L' sur L'' par $\varphi(e_{\alpha^*})=F_\alpha$ et

$$\varphi(e_\alpha)=-\varphi(Je_{\alpha^*})=-J(F_\alpha)=-\frac{e_\alpha^*+t_\alpha e_\alpha}{\sqrt{1+t_\alpha^2}},$$

obtenant une matrice unitaire avec

$$\text{dét}\,\varphi = \prod_{\alpha} (t_\alpha - i)/\sqrt{1+t_\alpha^2} = \exp\left(i \sum \operatorname{Arg}\left(\frac{t_\alpha - i}{\sqrt{1+t_\alpha^2}}\right)\right).$$

L ira alors sur L'' par ψ avec $\det \psi = \det J \det \varphi$. Pour les calculs de logarithmes nous nous plaçons sur $S^1 - \{-i\}$.

$$\operatorname{Arg}\frac{t_\alpha - i}{1+t_\alpha^2} = \theta + 2k\pi, \quad -\frac{\pi}{2} < \theta < \frac{3\pi}{2}.$$

Si $t_\alpha > 0$, nous prenons pour $\log\left[\exp\left(i \operatorname{Arg}\left(\frac{t_\alpha - 1}{\sqrt{1+t_\alpha^2}}\right)\right)\right]$ la valeur $i\theta$ et pour $\log\left[\exp\left(-i \operatorname{Arg}\left(\frac{t_\alpha - i}{\sqrt{1+t_\alpha^2}}\right)\right)\right]$ la valeur $-i\theta$. Si $t_\alpha < 0$, $\log\left[\exp\left(i \operatorname{Arg}\left(\frac{t_\alpha - i}{\sqrt{1+t_\alpha^2}}\right)\right)\right]$ étant égal à $i\theta$, le logarithme de l'inverse sera $i(2\pi - \theta)$.

Dans le calcul de $\log \det J + \log \det \varphi + \log (\det \psi)^{-1}$, apparaîtra donc $2\pi i \operatorname{sgn}(t_\alpha)$; nous avons donc prouvé, revenant à la variété V la

Proposition 1. *Pour tout $x \in V$, on peut trouver trois voisinages $U_{\alpha'}, U_{\beta'}, U_{\gamma'}$ de x, munis respectivement de sections lagrangiennes $L_{\alpha'}, L_{\beta'}, L_{\gamma'}$ deux à deux transverses sur $U_{\alpha'} \cap U_{\beta'} \cap U_{\gamma'}$, tels qu'il existe des fonctions de transition du fibré presque complexe: $u_{\alpha'\beta'}(x)\colon L_{\alpha'}(x) \to L_{\beta'}(x)$ (et permutation circulaire α', β', γ') avec*

$$\operatorname{Inert}(L_{\alpha'}, L_{\beta'}, L_{\gamma'}) = \frac{1}{2\pi i}\left[\log\left(\det(u_{\alpha'\beta'})\right) + \log\left(\det(u_{\beta'\gamma'})\right) - \log\left(\det(u_{\alpha'\gamma'})\right)\right] = c_{\alpha'\beta'\gamma'}.$$

Ainsi le cocycle trilatère est représenté au-dessus de $U_{\alpha'} \cap U_{\beta'} \cap U_{\gamma'}$ par l'indice d'inertie du triplet de lagrangiens $L_{\alpha'}, L_{\beta'}, L_{\gamma'}$ deux à deux transverses, autrement dit la première classe de Chern de V est représentée localement par cet indice.

Remarques. Notons que l'on pourrait dans la démonstration précédente supposer que $L' \cap L'' = L' \cap L = 0$ et $L \cap L'' \neq 0$, et $t_\alpha = 0$ pour $\alpha = r_1 + 1, \dots, r$, $r_1 = \dim L \cap L''$. On aurait alors à supposer L'_β transverse à $L_{\alpha'}$ et $L_{\gamma'}$ et $\dim (L_{\alpha'} \cap L_{\gamma'})$ constante sur $U_{\alpha'} \cap U_{\beta'} \cap U_{\gamma'}$ pour l'exactitude de l'énoncé. En termes de cohomologie à valeurs dans un faisceau, on sait que l'indice d'inertie trilatère définit un cocycle au-dessus de V qui est également un cobord (I), on voit que localement le cocycle trilatère s'identifie au cocycle d'inertie, exactement comme dans la cohomologie de de Rham tout cocycle s'identifie localement à un cobord.

3° *Conséquences.* Si $\forall x \in V$, il est toujours possible d'envisager $L_{\alpha'}, L_{\beta'}, L_{\gamma'}$ avec un indice d'inertie de parité fixe (que l'on peut supposer pair en modifiant au besoin l'un des $t_\alpha \neq 0$, en $(-t_\alpha)$), alors on peut définir le cocycle trilatère de V par des $c_{\alpha'\beta'\gamma'}(x)$ pairs, $\forall x \in V$. Il existe $\tilde{c}_1 \in H^2(X, \mathbf{Z})$ tel que $c_1 = 2\tilde{c}_1$ si c_1 est la 1ére classe de Chern du fibré tangent, donc selon [5], [8] une structure $Sp_2(r)$-spinorielle symplectique. Réciproquement s'il existe une telle structure spinorielle symplectique il existe $\tilde{c}_1$ tel que $c_1 = 2\tilde{c}_1$ et on peut envisager pour les $c_{\alpha'\beta'\gamma'}$ des valeurs paires; donc:

PROPOSITION 2. *Pour qu'il existe sur V une structure $Sp_2(r)$-spinorielle symplectique il faut et il suffit qu'en tout point l'indice d'inertie du triplet local de lagrangiens soit pair, autrement dit par abus que le cocycle d'inertie, Inert, soit pair.*

On voit, indépendamment de toute considération étrangère à cet article que pour qu'une variété à structure presque symplectique admette une structure spinorielle orthogonale attachée à la métrique hermitienne associée il faut et il suffit que sa 2$^{\text{éme}}$ classe de STIEFEL—WHITNEY ω_2 soit nulle ($\omega_2=0$).

En effet si le cocycle Inert est pair il est évident que $\omega_2=0$, puisque ω_2 est la réduction mod 2 de c_1. Réciproquement, si $\omega_2=0$, il existe un 1-cocycle $b_{\alpha\beta}$ à valeurs dans $\mathbf{Z}$ tel que

$$\bar{c}_{\alpha\beta\gamma}=\bar{b}_{\alpha\beta}+\bar{b}_{\beta\gamma}-\bar{b}_{\alpha\gamma},$$

où les bases désignent les classes modulo 2, alors $a_{\alpha\beta\gamma}$ est cohomologue à $b_{\alpha\beta\gamma}$ avec les $b_{\alpha\beta\gamma}$ pairs, on peut donc représenter localement le cocycle trilatère par des fonctions à valeurs dans $2\mathbf{Z}$.

Plus généralement: Si on peut relever le groupe structural $U(r, \mathbf{C})$ du fibré tangent à $Sp_q(r)$, la première classe de Chern (mod q) de V est nulle et réciproquement. Pour $q=2$ on retrouve directement la propriété de ω_2 car l'anti-image de $U(r, \mathbf{C})$ dans $Sp_2(r)$ est celle d'un groupe de spinorialité élargi.

III. Construction directe élémentaire des classes de Chern, son lien avec le cocycle d'inertie, son équivalence avec la construction de Chern—Weil

Une énorme littérature a été consacrée à la définition et à la signification géométrique des classes de Chern. Dans [1], vol. 81, A. BOREL et F. HIRZEBRUCH en donnent sept définitions. Ces définitions, excepté celle dite de CHERN—WEIL, relèvent essentiellement de la topologie algébrique: elles donnent des classes entières, tandis que la définition de Chern—Weil [2] conduit à des classes réelles. Le passage des classes réelles au classes entières se fait généralement en utilisant un critère d'unicité.

Nous nous proposons de donner une construction des classes de Chern qui n'utilise qu'un seul théorème élaboré, le théorème de décomposition en somme de Whitney de fibrés réels ou complexes par image réciproque (splitting map). Nous établissons directement que les classes de Chern—Weil de tout ordre sont des classes entières en utilisant le théorème généralisé de DE RHAM à partir de cette construction élémentaire. Dans le courant de l'article un lien avec le (II) sera établi.

1° Soit ξ un fibré vectoriel complexe de rang n, au-dessus de la base V paracompacte, $(U_\alpha)_{\alpha\in A}$ est un système d'ouverts trivialisants, définissant un cocyle, $u_{\alpha\beta}(x)\in U(n)$ si $x\in U_\alpha\cap U_\beta$. Le fibré en droites complexes (f. d. c.) associé au cocycle $\det(u_{\alpha\beta})$ est un élément de $H^1(V, \mathbf{C}^*)$ auquel correspond par un isomorphisme classique un élément de $H^2(V, \mathbf{Z})$ que nous appellerons, *par définition, première classe de Chern $c_1(\xi)$ de ξ.*

2° On pourra supposer que le recouvrement $\mathscr{U}$ défini par les $(U_\alpha)_{\alpha\in A}$ est contractile simple.

φ_j, $j=1, 2, \ldots, k$, $k \leqq n$, étant un élément de $U(n)$, définissons

$$(\varphi_1 \square \varphi_2 \square \ldots \square \varphi_k)(e_{\chi_1} \wedge e_{\chi_2} \wedge \ldots \wedge e_{\chi_k}) = \varepsilon^{i_1 i_2 \ldots i_k}_{\chi_1 \chi_2 \ldots \chi_k} \varphi_1(e_{i_1}) \wedge \varphi_2(e_{i_2}) \wedge \ldots \wedge \varphi_k(e_{i_k})$$

les (e_i) constituant une base de $\mathbf{C}^n$ et $1 \leqq \chi_1 < \chi_2 \ldots < \chi_k \leqq n$. Observons que si $\varphi_1 = \varphi_2 = \ldots \varphi_k = \varphi$, alors $\mathop{\square}\limits_{i=1}^{k} \varphi_i = k!(\Lambda^k \varphi)$. La trace de $\mathop{\square}\limits_{i=1}^{k} \varphi_i = (\varphi_1 \square \varphi_2 \square \ldots \square \varphi_k)$, notée $\mathrm{Tr}\,(\square\, \varphi_i)$ est égale à

$$\varepsilon^{i_1 i_2 \ldots i_k}_{l_1 l_2 \ldots l_k} (\varphi_1)^{l_1}_{i_1} (\varphi_2)^{l_2}_{i_2} \ldots (\varphi_k)^{l_k}_{i_k},$$

$\varepsilon^{i_1 \ldots i_k}_{l_1 \ldots l_k}$ étant la signature relative des suites d'indices distincts $(i_1, \ldots, i_k)$ et $(l_1, \ldots, l_k)$, on somme relativement aux indices répétés.

Si $x \in U_\alpha \cap U_\beta$, on peut écrire $u_{\alpha\beta}(x) = \exp\big(2\pi i a_{\alpha\beta}(x)\big)$, $a_{\alpha\beta}(x)$ est une matrice hermitienne, x étant fixé une telle matrice $a_{\alpha\beta}(x)$ existe car l'exponentielle est surjective pour le groupe de Lie connexe compact $U(n)$, mais on ne s'intéresse ni à l'unicité du choix de $a_{\alpha\beta}(x)$, ni à la continuité de $x \to a_{\alpha\beta}(x)$. C'est pour cette raison que nous appellerons *2k-cochaîne formelle,* l'assignation à $x \in U_{\alpha_0} \cap U_{\alpha_1} \cap \ldots \cap U_{\alpha_{2k}} = U_{\alpha_0 \alpha_1 \ldots \alpha_{2k}}$ d'une valeur en x de

$$c_{\alpha_0 \alpha_1 \ldots \alpha_{2k}} = \mathrm{Tr}\,(\sigma_{\alpha_0 \alpha_1 \alpha_2} \square \sigma_{\alpha_2 \alpha_3 \alpha_4} \square \ldots \square \sigma_{\alpha_{2k-2}, \alpha_{2k-1}, \alpha_{2k}}) \tag{9}$$

noté encore: $\mathrm{Tr}\,(\sigma_{\alpha_0 \alpha_1 \ldots \alpha_{2k}})$ avec

$$\sigma_{\alpha_0 \alpha_1 \alpha_2}(x) = a_{\alpha_0 \alpha_1}(x) + a_{\alpha_1 \alpha_2}(x) - a_{\alpha_0 \alpha_2}(x),$$

et notations analogues. Si $k=1$, on a une véritable cochaîne de Čech qui est le cocycle à valeurs dans $\mathbf{Z}$ intervenant au 1°.

3° $\sigma_{\alpha_0 \alpha_1 \ldots \alpha_{2k}}$ *définit un cocycle formel.*

Nous raisonnons par récurrence. Envisageons $U_{\alpha_0 \alpha_1 \ldots \alpha_{2k}}$, écrivons par abus σ_{012} pour $\sigma_{\alpha_0 \alpha_1 \alpha_2}$, $\sigma_{\alpha_0 \alpha_1 \ldots \alpha_{2k}} = \sigma_{01 \ldots 2k}$, … etc … Nous avons à introduire

$$\sigma_{\hat{0}1 \ldots 2k-1} - \sigma_{0\hat{1}2 \ldots 2k+1} + \ldots - \sigma_{012 \ldots \widehat{2k+1}},$$

noté $\delta(c_{012 \ldots 2k})$ et à calculer la trace de

$$(\sigma_{\hat{0}12 \ldots 2k-1} \square \sigma_{2k-1, 2k, 2k+1}) - (\sigma_{0\hat{1}2 \ldots 2k-1} \square \sigma_{2k-1, 2k, 2k+1}) + \ldots$$

$$\ldots + (\sigma_{012 \ldots \widehat{2k-2}, 2k-1} \square \sigma_{2k-1, 2k, 2k+1})$$

$$-(\sigma_{012 \ldots 2k-2, \widehat{2k-1}} \square \sigma_{2k-2, 2k, 2k+1}) + (\sigma_{012 \ldots 2k-2} \square \sigma_{2k-2, 2k-1, 2k+1}) -$$

$$-(\sigma_{012 \ldots 2k-2} \square \sigma_{2k-2, 2k-1, 2k})$$

et les trois derniers termes donnent

$$-\sigma_{012 \ldots 2k-2} \square \sigma_{2k-1, 2k, 2k+1},$$

de sorte que l'on obtient finalement $\delta(\sigma_{01 \ldots 2k-2}) \square \sigma_{2k-1, 2k, 2k+1}$, qui est nul par hypothèse de récurrence puisque $\delta(c_{012}) = 0$.

Il est à noter que si ce cocycle est considéré comme un cocycle formel à valeurs dans $\mathbf{R}$, c'est aussi un cobord formel, c'est le cobord de:

$$\mathrm{Tr}\,(a_{01} \square \sigma_{123} \square \ldots \square \sigma_{2k-3, 2k-2, 2k-1}).$$

4° *Supposons que ξ soit la somme de Whitney de n fibrés complexes ξ_l de rang* 1,

$$\xi = \bigoplus_{l=1}^{n} \xi_l. \tag{10}$$

Alors $\sigma_{120} = \sum_{l=1}^{n} \sigma_{012}^{l}$, et il vient de (10), les $(\sigma_{\alpha\beta\gamma}^{l})_j^i$ étant des matrices scalaires notées $a_{\alpha\beta\gamma}^{l}$:

$$c_{012\ldots 2k} = \sum_{l_1, l_2, \ldots, l_k} a_{012}^{l_1} a_{234}^{l_2} \ldots a_{2k-2, 2k-1, 2k}^{l_k} \tag{11}$$

la somme étant étendue aux arrangements des $1, 2, \ldots, n, k$ à k.

De $\exp(2\pi i a_{\alpha_0\alpha_1}^{l}) \exp(2\pi i a_{\alpha_1\alpha_2}^{l}) \exp(2\pi i a_{\alpha_2\alpha_0}^{l}) = \mathrm{Id}$, on déduit que $a_{\alpha_0\alpha_1}^{l} + a_{\alpha_1\alpha_2}^{l} - a_{\alpha_0\alpha_2}^{l} \in \mathbf{Z}$ et que $c_{012\ldots 2k}$ prend ses valeurs dans $\mathbf{Z}$.

Ainsi, dans le cas envisagé au 4°, les cochaînes formelles sont des cochaînes de Čech à valeurs entières.

5° Revenant au cas général d'un fibré complexe ξ quelconque, on sait (cf. démonstration dans [4]) qu'il existe un fibré complexe λ_1 de base V_1 et une application continue $f\colon V_1 \to V$ tels que $\xi_1 = f^*(\xi)$ soit la somme de Whitney de n fibrés complexes de rang 1 (splitting map), f^* étant une application injective de $H^*(V, \mathbf{Z})$ dans $H^*(V_1, \mathbf{Z})$ (cf. aussi plus bas remarque 1).

Le système des $f^{-1}(U_\alpha)$ constitue un recouvrement trivialisant $f^{-1}(\mathscr{U})$ pour ξ_1 avec les $u_{\alpha\beta}$ comme cocycle, $c_{012\ldots 2k}$ peut être considérée comme une $2k$-cochaîne formelle sur V_1 pour le fibré ξ_1. Il existe un recouvrement trivialisant $\mathscr{U}_1$, plus fin que $\mathscr{U}$, ensemble d'ouverts $(U_{\alpha_1})_{\alpha_1 \in A_1}$ tel que si

$$y \in U_{\alpha_1} \cap U_{\beta_1} \quad \text{et à} \quad f^{-1}(U_\alpha \cap U_\beta)$$

$$u_{\alpha\beta}(f(y)) = \exp\bigl(2\pi i\, a_{\alpha\beta}(f(y))\bigr) = \lambda_{\alpha_1}^{-1}(y) \exp\bigl(2\pi i\, a_{\alpha_1\beta_1}(y)\bigr) \lambda_{\beta_1}(y), \tag{12}$$

$a_{\alpha_1\beta_1}(y)$ étant une matrice diagonale. Mais ici $y \to a_{\alpha_1\beta_1}(y)$ est continue de sorte que l'on peut associer à $\mathscr{U}_1$ une $2k$-cochaîne de Čech: $\underset{(1)}{c}{}_{012\ldots 2k}$, qui d'après le 4° sera à valeurs entières, et d'après le 3° un cocycle. Si deux cochaînes de ce type associées au même fibré $f^*(\xi)$ sont relatives à un même recouvrement ouvert $\mathscr{U}_1$ alors elles diffèrent par un cobord. Passant à la limite inductive pour tous les recouvrements tels que $\mathscr{U}_1$, on obtient un élément de $H^{2k}(V_1, \mathbf{Z})$. Nous allons montrer que cet élément est $k!\,c_k(\xi_1)$ et par $(f^*)^{-1}$ injective, il lui correspondra $k!\,c_k(\xi) \in H^{2k}(V, \mathbf{Z})$.

Rappelons que nous devons résoudre le problème suivant:

Dans la catégorie des fibrés vectoriels complexes à bases paracompactes

(I) A chaque fibré ξ de base V, on sait associer $c_k(\xi) \in H^{2k}(V, \mathbf{Z})$, $c_k(\xi) = 0$, pour $k >$ rang ξ, $c_0(\xi) = 1$.

On pose $c(\xi) = \sum_{0}^{n} c_k(\xi)$.

(II) Si ξ et η sont V-isomorphes, $c(\xi) = c(\eta)$, et si $f\colon B \to V, f^*\bigl(c(\xi)\bigr) = c\bigl(f^*(\xi)\bigr)$.

(III) $c_k\left(\bigoplus_{l=1}^{n} \xi_l\right) = \sum_{i_1, i_2, \ldots, i_k} c_1(\xi_{i_1}) \cup c_1(\xi_{i_2}) \cup \ldots \cup c_1(\xi_{i_k})$, somme étendue à toutes les combinaisons k à k de $1, 2, \ldots, n$, $\cup$ est le cup-produit.

(IV) $c_1(\xi)$ est défini comme au (1°).

(V) $c(\xi)$ ainsi défini est unique.

Nous définissons le cup-produit par:

$$(c \cup c')(U_{i_0, i_1, \ldots, i_{p+q}}) = c(U_{i_0 \ldots i_q}) \cdot c'(U_{i_q \ldots i_{p+q}}).$$

(III) vient du 4° compte tenu de $\lambda \cup \mu = (-1)^{pq} \mu \cup \lambda$ $p = d^0 \lambda$, $q = d^0 \mu$, dans l'algèbre de cohomologie. Les autres propriétés sont immédiates. Donc, en résumé: *la 2k-cochaîne formelle* $\frac{1}{k!}\, c_{\alpha_0 \alpha_1 \ldots \alpha_{2k}}$ *définit la* $k^{\text{ième}}$ *classe de Chern de* ξ, *au moyen de la construction explicitée au* 5°.

Remarque 1. Rappelons la construction de l'espace V_1.

Si $q\colon P(\xi) \to V$ est le fibré projectif associé à $p\colon \xi \to V$, on construit $q^*(\xi)$, si ξ est de rang 1, $V_1 = V$, $f = \mathrm{Id}$. Raisonnons par récurrence. Soit $q^*(\xi) = \lambda_\xi \oplus \sigma_\xi$ σ_ξ de rang $(n-1)$, de base $P(\xi)$. Il existe g et V_1 satisfaisant aux conditions du splitting map pour σ_ξ, $g\colon V_1 \to P(\xi)$, g^* injective en cohomologie $q \circ g\colon V_1 \to V$, $(q \circ g)^*(\xi)$ est complètement réductible, l'injectivité en cohomologie résulte d'une propriété générale des fibrés projectifs.

Remarque 2. On a établi dans (II) que la première classe de Chern est représentée localement par le cocycle d'inertie de trois lagrangiens deux à deux transverses, cette propriété s'étend ici naturellement, (3) correspond à une décomposition locale de la cochaîne qui représente la classe de Chern d'ordre k au moyen d'un polynôme relativement aux cocycles d'inertie de triplets de lagrangiens deux à deux transverses au-dessus d'ouverts U'_{012}, U'_{234}, ... suffisamment petits, contenant un élément $x \in V$.

Ainsi la classe totale de Chern se représente localement par une cochaîne factorisable au moyen de cocycles d'inertie de lagrangiens deux à deux transverses.

6° *L'équivalence avec la définition de Chern—Weil.*

Nous nous proposons d'établir directement que les classes de Chern—Weil sont des classes entières.

V est ici une variété différentiable et les fibrations sont elles-mêmes différentiables. De la suite exacte

$$0 \to \mathbf{R} \xrightarrow{i} \mathscr{A}_0 \xrightarrow{d_0} \mathscr{A}_1 \xrightarrow{d_1} \ldots \mathscr{A}_p \xrightarrow{d_p} \ldots$$

où i est l'inclusion, $\mathscr{A}_p$ le faisceau des germes des formes différentielles de degré p, d_p la différentielle extérieure restreinte à $\mathscr{A}_p$, donnant une résolution fine de $\mathbf{R}$ ($\mathscr{A}_i$ est un faisceau fin), on déduit la suite d'isomorphismes

$$H^i(V, \ker d_{\alpha+1}) \xrightarrow[\delta^i_\alpha]{\sim} H^{i+1}(V, \ker d_\alpha) \xrightarrow[\delta^{i+1}_{\alpha-1}]{\sim} \ldots \xrightarrow[\delta^{i+\alpha}_0]{\sim} H^{i+\alpha+1}(V, \ker d_0) \simeq$$

$$\simeq H^{i+\alpha+1}(V, \mathbf{R})$$

venant de la suite exacte

$$0 \to \ker d_\alpha \subseteqq \mathscr{A}_\alpha \xrightarrow{d_\alpha} \ker d_{\alpha+1} \to 0.$$

Supposons que le fibré complexe soit ξ_1 sur V_1. Selon la définition de l'opérateur cobord (δ^i_α), quand on envisage l'action de $(\delta^i_\alpha)^{-1}$, on considère $\frac{1}{k!} c_{012\dots 2k}$, on lui associe $\operatorname{Tr}(a_{01} \square \sigma_{123} \square \dots \square \sigma_{2k-3,2k-2,2k-1})$ dont on prend la différentielle

$$\frac{1}{k!} \operatorname{Tr}(da_{01} \square \sigma_{123} \square \dots \square \sigma_{2k-3,2k-2,2k-1}) \tag{13}$$

puisque les $\sigma_{\alpha\beta\gamma}$ sont localement constantes. Or

$$d((a_{01})) = \frac{-1}{2\pi i} \operatorname{Tr}(\omega_0 - \omega_1),$$

ω étant une connexion unitaire, ce qui par application de l'opérateur inverse du cobord donne: $\frac{-1}{2\pi i}\ \Omega$, Ω étant la forme de courbure.

On aboutit à

$$\frac{-1}{k!} \frac{1}{2\pi i} \operatorname{Tr}(\Omega \square \sigma_{123} \square \dots \square \sigma_{2k-3,2k-2,2k-1}).$$

En recommençant autant qu'il le faut, on obtient

$$\left(\frac{-1}{2\pi i}\right)^k \operatorname{Tr}(\Omega \wedge \Omega \wedge \dots \wedge \Omega),$$

avec k facteurs Ω. C'est bien un représentant de $c_k(\xi_1)$ dans la méthode de Chern—Weil. En revenant au cas général du fibré ξ sur V, si Δ est l'isomorphisme de $H^k(V_1, \mathbf{R})$ dans $H^k(V_1, \mathbf{Z})$ que l'on vient de mettre en évidence, $(f^*)^{-1} \circ \Delta \circ f^*$ donne un isomorphisme de $H^k(V, \mathbf{R})$ dans $H^k(V, \mathbf{Z})$ qui associe la $k^{\text{ième}}$ classe entière $c_k(\xi)$, à la $k^{\text{ième}}$ classe entière de Chern—Weil (on sait que l'injectivité de f^* s'étend à la cohomologie réelle).

BIBLIOGRAPHIE

[1] Borel, A.—Hirzebruch, F., Characteristic classes and homogeneous spaces I, *Amer. J. Math.* **80** (1958), 458—538.

—, Characteristic classes and homogeneous spaces II, *Amer. J. Math.* **81** (1959), 315—382.

—, Characteristic classes and homogeneous spaces III, *Amer. J. Math.* **82** (1960), 491—504.

[2] Chern, S. S., Characteristic classes of Hermitian manifolds, *Ann. of. Math.* (2) **47** (1946), 85—121.

[3a] Crumeyrolle, A., Algèbre de Clifford symplectique. Revêtements du groupe symplectique. Indices de Maslov et spineurs symplectiques, *J. Math. Pures et Appl.* **56** (1977), 205—230.

[3b] —, Revêtements spinoriels du groupe symplectique et indices de Maslov, *C. R. Acad. Sci. Paris Sér. A* **280** (1975), 1753—1756.

[4] HUSEMOLLER, D., *Fibre bundles,* McGraw—Hill Book Co., New York—London—Sydney, 1966, 235—236.

[5] KOSTANT, B., Symplectic spinors, *Symposia Mathematica,* Vol. XIV (Convegno di Geometria Simplettica e Fisica Matematica, INDAM, Rome, 18—23 Gennaio 1973), 139—152; Academic Press, London, 1974.

[6] LERAY, J., Travaux de Maslov—Arnold. R.C.P. 25 — Vol. 18, Strasbourg, 1973.

[7] SOURIAU, J.-M., Construction explicite de l'indice de Maslov. Applications, *Group theoretical methods in physics* (Fourth Internat. Colloq., Nijmegen, 1975), 117—148; Lecture Notes in Physics, Vol. 50, Springer-Verlag, Berlin, 1976.

[8] TIMBEAU, J., Structure spinorielle sur une variété presque-symplectique, *C. R. Acad. Sci. Paris Sér. A* **279** (1974), 273—276.

Université Paul Sabatier, Faculté des Sciences
118, Route de Narbonne, F—31062 Toulouse Cédex

(Reçu le 26 octobre 1977)

TRIGONOMETRISCHE APPROXIMATION MIT GLEICHVERTEILUNGSMETHODEN

von

P. ZINTERHOF und H. STEGBUCHNER

Abstract

Trigonometric Approximation with Methods of Uniform Distribution. We consider a quantity F_N similar to the Discrepancy D_N, measuring the „quality" of uniform distribution of a sequence $\{x_n\}$ in $\mathbf{R}^s$, which can be computed easily. Further a new method forming „partial sums" of multidimensional Fourier series is introduced.

0. Einleitung

Zum Problemkreis der trigonometrischen Approximation mit Gleichverteilungsmethoden existiert eine umfangreiche Literatur, die sehr vollständig in [2] und [3] angegeben ist. Sind die Funktionswerte von $f(\mathrm{x})$ in den Punkten $\mathrm{x}_1, \ldots, \mathrm{x}_N$ bekannt, so kann der Approximationsfehler durch die Diskrepanz D_N dieser Punkte abgeschätzt werden. Die praktische Anwendung solcher Abschätzungen ist jedoch problematisch, da die Diskrepanz einer Punktfolge im $\mathbf{R}^s$ für große Werte von N und (oder) s wegen des enorm hohen Rechenaufwandes in der Praxis nicht mehr berechnet oder befriedigend abgeschätzt werden kann. In dieser Arbeit wird eine der Diskrepanz ähnliche Größe F_N betrachtet, welche stets mit $O(N^2)$ Rechenschritten, in vielen Fällen jedoch mit wesentlich geringerem Rechenaufwand, ermittelt werden kann. Die Anzahl der Rechenoperationen ist dabei von der Dimension s unabhängig. Für $s=1$ wurde diese Größe in [6] diskutiert.

Im weiteren wird der Begriff der „Partialsumme" einer mehrdimensionalen Fourierreihe dahingehend verallgemeinert, daß nicht, wie allgemein üblich, die Teilsummen über s-dimensionale Quader oder Kugelbereiche sondern über hyperbolische Bereiche gebildet werden. Mit dieser Methode können für viele Funktionenklassen erheblich bessere Konvergenzaussagen hergeleitet werden.

Für die Terminologie sei auf [3] verwiesen.

1. Die Diaphonie F_N

Sei $g(x)=1-\frac{\pi^2}{6}+\frac{\pi^2}{2}(1-2\{x\})^2$ $(x\in\mathbf{R})$ und $H(\mathrm{x})=H(x_1, \ldots, x_s)=\prod_{i=1}^{s} g(x_i)-1$. $H(\mathrm{x})$ besitzt die absolut konvergente Fourierreihe $\sum_{\mathrm{m}\in\mathbf{Z}^s} \frac{\exp(\mathrm{m}\cdot\mathrm{x})}{R(\mathrm{m})^2}$ mit $R(\mathrm{m})= \prod_{i=1}^{s} \max(1, |x_i|)$, $\exp(t)=e^{2\pi i t}$, wobei die Summe über alle $\mathrm{m}\neq(0, \ldots, 0)$ zu erstrecken ist.

Definition 1.1. *Unter der Diaphonie der Punktefolge* $\mathrm{x}_1, \ldots, \mathrm{x}_N$ *wollen wir die Größe*

$$F_N = F_N(\mathrm{x}_1, \ldots, \mathrm{x}_N) = \left(\frac{1}{N^2} \sum_{k=1}^{N} \sum_{l=1}^{N} H(\mathrm{x}_k - \mathrm{x}_l)\right)^{1/2}$$

verstehen.

Es ist wohlbekannt, daß die Diskrepanz einer Punktfolge genau dann gegen Null strebt, wenn die Folge gleichverteilt ist. Dieselbe Tatsache gilt auch für F_N.

Satz 1.1. *Die Folge* $\{\mathrm{x}_n\}$ *ist genau dann gleichverteilt* mod 1, *wenn* $\lim\limits_{N\to\infty} F_N = 0$ *gilt.*

Beweis. Wie man sofort nachrechnet ist $F_N^2 = \sum\limits_{\mathrm{m}}' \dfrac{\left|\frac{1}{N} \sum\limits_{k=1}^{N} \exp(\mathrm{x}_k \cdot \mathrm{m})\right|^2}{R^2(\mathrm{m})}$. Für $\varepsilon > 0$ bel. gilt daher für $N \geqq N(\varepsilon)$ und $h \in \mathbf{N}$ bel.

$$\sum_{\|\mathrm{m}\|_\infty \leqq h}' \frac{\left|\frac{1}{N} \sum\limits_{k=1}^{N} \exp(\mathrm{x}_k \cdot \mathrm{m})\right|^2}{R^2(\mathrm{m})} < \varepsilon$$

falls $\lim\limits_{N\to\infty} F_N = 0$ gilt. Die Ungleichung von Koksma—Erdős—Turán [3] liefert $\lim\limits_{N\to\infty} D_N = 0$.

Die Umkehrung des Satzes folgt unmittelbar aus Satz 1.2.

Lemma 1.1. *Es sei* $\omega = \{\mathrm{a}_1, \ldots, \mathrm{a}_N\}$ *eine Punktfolge aus dem* $\mathbf{R}^s$ *mit Diskrepanz* D_N. *Für ein festes* k $(1 \leqq k \leqq N)$ *betrachten wir die Punktfolge* $\omega^* = \{\mathrm{a}_1 - \mathrm{a}_k, \ldots, \mathrm{a}_N - \mathrm{a}_k\}$ *und bezeichnen ihre Diskrepanz mit* D_N^*. *Dann gilt* $D_N^* \leqq 3^s D_N$.

Beweis. Es seien $0 = \beta_{i,0} < \beta_{i,1} < \ldots < \beta_{i,s_i} = 1$ die verschiedenen Werte, die die i-ten Koordinaten der Punkte aus ω annehmen. Für die Folge ω^* erhalten wir dann analog

$$(1.1)\qquad \begin{aligned} 0 &= \beta_{i,k_i} - a_k^{(i)} < \beta_{i,k_i+1} - a_k^{(i)} < \ldots < \beta_{i,s_i} - a_k^{(i)} = \\ &= 1 + \beta_{i,0} - a_k^{(i)} < \ldots < 1 + \beta_{i,k_i-1} - a_k^{(i)} < 1. \end{aligned}$$

Diese Folge wollen wir mit $0 = \alpha_{i,0} < \alpha_{i,1} < \ldots < \alpha_{i,s_i} = 1$ bezeichnen. Für ein beliebiges s-Tupel $(\alpha_{1,j_1}, \ldots, \alpha_{s,j_s})$ mit $0 \leqq j_i < s_i$ $i = 1, \ldots, s$ definieren wir das Intervall

$$Q^* = \{(x_1, \ldots, x_s)\colon \alpha_{i,j_i} < x_i \leqq \alpha_{i,j_i+1},\ 1 \leqq i \leqq s\}.$$

Σ^* bezeichne die Vereinigung all dieser Intervalle. Ferner sei Σ die Vereinigung aller Intervalle der Form Q, die analog zu Q^* mit den Zahlen $\beta_{i,j}$ definiert werden. Wie in [4] wollen wir mit $\mathrm{y}(Q^*)$ den „oberen" und mit $\mathrm{z}(Q^*)$ den „unteren" Eckpunkt von Q^* bezeichnen. Dann gilt (siehe [4]):

$$(1.2)\quad D_N = \max_{Q^* \in \Sigma^*} \max\left(\left|\frac{A(\mathrm{y}(Q^*);\, N)}{N} - V(\mathrm{y}(Q^*))\right|,\ \left|\frac{A(\mathrm{y}(Q^*);\, N)}{N} - V(\mathrm{z}(Q^*))\right|\right).$$

Betrachten wir nun ein festes $Q^*\in\Sigma^*$ mit $\mathrm{y}(Q^*)=(\alpha_{1,l_1}, \dots, \alpha_{s,l_s})$ und nehmen wir an, daß $\mathrm{a}_n-\mathrm{a}_k$ (bzw. seine periodische Fortsetzung) zum Intervall $[0, \mathrm{y}(Q^*))$ gehört. Nach (1.1) können nun für jedes i $(i=1, 2, \dots, s)$ zwei Fälle eintreten:

$$\begin{aligned}&\text{Fall A: } \alpha_{i,l_i}=\beta_{i,r_i}-a_k^{(i)} \quad \text{d. h.} \quad a_k^{(i)}\leqq a_n^{(i)}<\beta_{i,r_i}\\ (1.3)\quad &\\ &\text{Fall B: } \alpha_{i,l_i}=1+\beta_{i,r_i}-a_k^{(i)} \quad \text{d. h.} \quad 0\leqq a_n^{(i)}<\beta_{i,r_i}\vee a_k^{(i)}\leqq a_n^{(i)}<1.\end{aligned}$$

Es kann somit die Anzahl der Punkte der Folge ω^*, welche in $[0, \mathrm{y}(Q^*))$ liegen, durch die Anzahl der Punkte der ursprünglichen Folge ω, die in bestimmten disjunkten achsenparallelen Quadern $T_1, \dots, T_{j(Q^*)}$ liegen, ausgedrückt werden. Wie unmittelbar aus (1.3) ersichtlich ist, gilt für das Volumen $V(\mathrm{y}(Q^*))$

$$V(\mathrm{y}(Q^*))=\sum_{j=1}^{j(Q^*)} V(T_j).$$

Bezeichnen wir nun mit T einen dieser Quader, so ist $T=\{(x_1, \dots, x_s)\colon c_i\leqq x_i<d_i, i=1, 2, \dots, s\}$ mit $c_i\in\{0, \beta_{i,r}, a_k^{(i)}\}$ und $d_i\in\{1, \beta_{i,r}, a_k^{(i)}\}$, $r=0, 1, \dots, s_i$ und somit gilt

$$\begin{aligned}(1.4)\qquad V(T)&=(d_1-c_1)(d_2-c_2)\dots(d_s-c_s)=\\ &=d_1\dots d_s-c_2d_2\dots d_s\pm\dots=\\ &=e_{1,1}\dots e_{1,s}\pm e_{2,1}\dots e_{2,s}\pm\dots\pm e_{t,1}\dots e_{t,s}.\end{aligned}$$

Nun ist aber jeder Punkt $(e_{i,1}, \dots, e_{i,s})$ ein bestimmtes $\mathrm{y}(Q)$ mit $Q\in\Sigma$; es gilt daher

$$(1.5)\qquad V(T_l)=\pm V(\mathrm{y}(Q_{l_1}))\pm V(\mathrm{y}(Q_{l_2}))\pm\dots\pm V(\mathrm{y}(Q_{l_t})),$$

wobei die Vorzeichen gemäß (1.4) zu wählen sind. Natürlich gilt auch für die Anzahl der Punkte in T_l

$$(1.6)\qquad A(T_l; N)=\pm A(\mathrm{y}(Q_{l_1}); N)\pm\dots\pm A(\mathrm{y}(Q_{l_t}); N),$$

wobei in (1.6) wieder dieselbe Vorzeichenverteilung wie in (1.5) zu wählen ist.

Analog gilt nun auch für den Bereich $\mathrm{z}(Q^*)$, daß seine periodische Fortsetzung in paarweise disjunkte achsenparallele Quader $U_1, \dots, U_{j(Q^*)}$ zerfällt, wobei unter diesen einige entartet sein können. Analog zu (1.5) gilt nun auch wieder

$$V(U_l)=\pm V(\mathrm{z}(\tilde{Q}_{l_1}))\pm\dots\pm V(\mathrm{z}(\tilde{Q}_{l_t})).$$

Wie man sich leicht überlegt, wird aber im allgemeinen $\mathrm{z}(\tilde{Q}_{l_j})$ verschieden von $\mathrm{z}(Q_{l_j})$ sein. Es gelten jedoch die Ungleichungen

$$(1.7)\qquad V(\mathrm{z}(Q_{l_j}))\leqq V(\mathrm{z}(\tilde{Q}_{l_j}))\leqq V(\mathrm{y}(Q_{l_j})).$$

Wir haben nun die Anzahl der Summanden in (1.4) für ein festes Q^* abzuschätzen. Diese hängt natürlich davon ab, wie oft in (1.3) die Fälle A bzw. B auftreten. Nehmen wir an, daß m mal Fall B und $s-m$ mal Fall A eintritt, so rechnet man sofort nach, daß $j(Q^*)=2^m$ gilt. In diesen 2^m Produkten der Gestalt (1.4) treten $\binom{m}{k}$ mal k Werte $c_{i_1}, \dots, c_{i_k}$ auf, die alle gleich Null sind; dies kann ohne viel Mühe mit Induktion bewiesen werden.

Wird nun ein Produkt der Gestalt (1.4) formal ausmultipliziert und sind k Werte der c_{i_j} gleich Null, so besteht die Summe aus 2^{s-k} von Null verschiedenen Summanden. Ihre Gesamtanzahl ergibt sich daher zu

$$\sum_{k=0}^{n}\binom{n}{k}2^{s-k}=\sum_{k=0}^{n}\binom{n}{k}2^{s-n}2^{n-k}=2^{s-n}3^{n}\leqq 3^{s}. \tag{1.8}$$

Wir erhalten somit nach (1.5), (1.6) und (1.8)

$$\left|\frac{A(\mathrm{y}(Q^*);\,N)}{N}-V(\mathrm{y}(Q^*))\right|\leqq\sum_{l=1}^{j(Q^*)}\sum_{j=1}^{t_l}\left|\frac{A(\mathrm{y}(Q_{l_j});\,N)}{N}-V(\mathrm{y}(Q_{l_j}))\right|\leqq 3^s D_N. \tag{1.9}$$

Wegen (1.7) gilt auch

$$\left|\frac{A(\mathrm{y}(Q^*);\,N)}{N}-V(\mathrm{z}(Q^*))\right|\leqq\sum_{l=1}^{j(Q^*)}\sum_{j=1}^{t_l}\left|\frac{A(\mathrm{y}(Q_{l_j});\,N)}{N}-V(\mathrm{z}(\tilde{Q}_{l_j}))\right|\leqq$$

$$\leqq\sum_{l=1}^{j(Q^*)}\sum_{j=1}^{t_l}\max\left(\left|\frac{A(\mathrm{y}(Q_{l_j});\,N)}{N}-V(\mathrm{z}(Q_{l_j}))\right|,\ \left|\frac{A(\mathrm{y}(Q_{l_j});\,N)}{N}-V(\mathrm{y}(Q_{l_j}))\right|\right)\leqq$$

$$\leqq 3^s D_N.$$

Aus (1.2) folgt nun zusammen mit (1.9) und der letzten Ungleichung die Behauptung von Lemma 1.1.

SATZ 1.2. *Die Diaphonie F_N erfüllt die Ungleichung*

$$F_N(\mathrm{x}_1,\ldots,\mathrm{x}_N)<6^s(D_N)^{1/2}.$$

BEWEIS. Nach dem Satz von KOKSMA—HLAWKA [1] und wegen $\int\limits_{E_s}H(x)\,dx=0$ erhalten wir

$$\left|\frac{1}{N}\sum_{l=1}^{N}H(\mathrm{x}_l-\mathrm{x}_k)\right|\leqq D_N^*V^{(s)}(H),$$

wobei D_N^* die Diskrepanz der Punktfolge $\{\mathrm{x}_1-\mathrm{x}_k,\ldots,\mathrm{x}_N-\mathrm{x}_k\}$ bedeutet und $V^{(s)}(H)$ die Variation der Funktion H im Sinne von Hardy und Krause. $V^{(s)}(H)$ ergibt sich im konkreten Fall zu

$$V^{(s)}(H)=\sum_{i=1}^{s}\binom{s}{i}\Big(\int_0^1|g'(x)|\,dx\Big)^i\Big(\int_0^1|g(x)|\,dx\Big)^{s-i}<12^s.$$

Lemma 1.1 liefert nun zusammen mit dieser Abschätzung

$$F_N^2=\frac{1}{N}\sum_{k=1}^{N}\left|\frac{1}{N}\sum_{l=1}^{N}H(\mathrm{x}_l-\mathrm{x}_k)\right|<36^sD_N.$$

Von besonderer numerischer Bedeutung wird F_N für die wichtigen Folgen $\{k\theta\}$, $\theta=(e^{r_1},\ldots,e^{r_s})$, wobei die r_i feste von Null verschiedene rationale Zahlen sind.

SATZ 1.3. *Für die Folgen $\{k\theta\}$ gilt*

$$\text{(i)} \qquad F_N^2 = \frac{H(0)}{N} + \frac{2}{N^2} \sum_{k=1}^{N-1} (N-k)\, H(k\theta).$$

(ii) *Es gelten die Rekursionen*

$$F_1^2 = H(0), \quad B_1 = H(\theta),$$
$$(N+1)^2 F_{N+1}^2 = N^2 F_N^2 = H(0) + 2B_N,$$
$$B_{N+1} = B_N + H\big((N+1)\theta\big).$$

(iii) *Für fast alle θ gilt $F_N = O(N^{\varepsilon-1})$, $\varepsilon > 0$ bel.*

Der Beweis des Satzes kann unter Verwendung der Resultate von [5] sofort vom Beweis der eindimensionalen Version aus [6] übernommen werden.

2. Numerische Integration

Wir geben zwei Integrationsmethoden an, bei denen die Fehlerabschätzung mit Hilfe der Diaphonie F_N durchgeführt wird.

DEFINITION 2.1. *$E_\alpha^s(C)$ sei die Menge aller auf dem $\mathbf{R}^s$ definierten periodischen Funktionen*

$$f(\mathrm{x}) = f(x_1, \dots, x_s) = \sum_{\mathrm{m}=-\infty}^{\infty} C(\mathrm{m}) \exp(\mathrm{m} \cdot \mathrm{x})$$

mit $|C(\mathrm{m})| \leqq C/R^\alpha(\mathrm{m})$ (siehe [2]).

DEFINITION 2.2. *Für $\mathrm{x}_1, \dots, x_N \in \mathbf{R}^s$ und $f(\mathrm{x}) \in E_\alpha^s(C)$ sei*

$$\text{(2.1)} \qquad I_N(f) = I_N(f; \mathrm{x}_1, \dots, \mathrm{x}_N) = \frac{1}{N^2} \sum_{k,l=1}^{N} f(\mathrm{x}_k - \mathrm{x}_l).$$

Für die speziellen Folgen $\{k\theta\}$ wird (2.1) *zu*

$$\text{(2.2)} \qquad I_N(f) = \frac{1}{N^2} \sum_{k=-(N-1)}^{N-1} (N-|k|)\, f(k\theta).$$

DEFINITION 2.3. *Die Gewichte $A_{k,N}^{(t)}$ werden definiert durch die Gleichung*

$$\text{(2.3)} \qquad \left(\sum_{k=0}^{N} z^k \right)^t = \sum_{k=0}^{Nt} A_{k,N}^{(t)} z^k.$$

Siehe dazu [6].

Speziell für die Folgen $\{k\theta\}$ ist nachstehende Definition einer Integrationsregel zugeschnitten.

DEFINITION 2.4. *Für $\theta \in \mathbf{R}^s$ und $f(\mathrm{x}) \in E_\alpha^s(C)$ sei*

$$\text{(2.4)} \qquad J_N^{(t)}(f) = J_N^{(t)}(f; \theta) = (N+1)^{-t} \sum_{k=0}^{Nt} A_{k,N}^{(t)} f(k\theta).$$

Wir werden vorerst einige Rekursionsformeln für die Gewichte $A_{k,N}^{(t)}$ herleiten.

LEMMA 2.1. *Die Gewichte $A_{k,N}^{(t)}$ besitzen für $t, N \in \mathbf{N}$ und $0 \leqq k \leqq Nt$ die Symmetrieeigenschaft*

$$A_{k,N}^{(t)} = A_{Nt-k,N}^{(t)}. \tag{2.5}$$

BEWEIS. Mit Hilfe des polynomischen Lehrsatzes ergibt sich unmittelbar die Gleichung

$$A_{k,N}^{(t)} = \sum_{\substack{\mu_1,\dots,\mu_{N+1}=0\\ \mu_1+\dots+\mu_{N+1}=t\\ \mu_2+2\mu_3+\dots+N\mu_{N+1}=k}}^{t} \frac{t!}{\mu_1!\dots\mu_{N+1}!}. \tag{2.6}$$

Aus (2.6) erhält man dann durch geeignete Umnummerierung der Indizes

$$A_{Nt-k,N}^{(t)} = \sum_{\substack{\mu_1,\dots,\mu_{N+1}=0\\ \mu_1+\dots+\mu_{N+1}=t\\ \mu_N+2\mu_{N-1}+\dots+N\mu_1=Nt-k}}^{t} \frac{t!}{\mu_1!\dots\mu_{N+1}!}. \tag{2.7}$$

Die Behauptung (2.5) folgt somit aus (2.6) und (2.7), wenn wir zeigen können, daß die beiden Aussagen (i) und (ii)

$$\text{(i)} \qquad \sum_{i=1}^{N+1} \mu_i = t \quad \text{mit} \quad 0\mu_1+1\mu_2+\dots+N\mu_{N+1} = k$$

$$\text{(ii)} \qquad \sum_{i=1}^{N+1} \mu_i = t \quad 0\mu_{N+1}+1\mu_N+\dots+N\mu_1 = Nt-k$$

äquivalent sind. Diese Äquivalenz kann jedoch leicht mittels Induktion nach N bewiesen werden.

SATZ 2.1. *Für die in (2.3) definierten Gewichte $A_{k,N}^{(t)}$ gelten die Rekursionen*

$$A_{k,N}^{(t)} \equiv 1, \quad A_{-1,N}^{(t)} \equiv 0 \quad (k, t, N \in \mathbf{N}),$$

$$A_{k,N}^{(t+1)} = A_{k-1,N}^{(t+1)} + A_{k,N}^{(t)} \quad \text{für} \quad 0 \leqq k \leqq N,$$

$$A_{k,N}^{(t+1)} = A_{k-1,N}^{(t+1)} + A_{k,N}^{(t)} - A_{k-1-N,N}^{(t)} \quad \text{für} \quad N < k \leqq \left[\frac{N(t+1)}{2}\right]+1.$$

BEWEIS. Wegen Lemma 2.1 kann man sich auf $k \leqq \left[\frac{N(t+1)}{2}\right]+1$ beschränken. Diese Rekursionsformeln können nun ohne viel Mühe hergeleitet werden, wenn man von der Gleichung

$$(1+z+\dots+z^N)^{t+1} = (1+z+\dots+z^N) \sum_{k=0}^{Nt} A_{k,N}^{(t)} z^k$$

ausgeht und die in der rechten Seite auftretenden Summanden geeignet zusammenfaßt.

Wir geben nun im Anschluß Fehlerabschätzungen für die Integrationsmethoden (2.1) und (2.3) an.

SATZ 2.2. *Sei* $f \in E_\alpha^s(C)$ *mit* $\alpha > 1, \mathrm{x}_1, \dots, \mathrm{x}_N \in \mathbf{R}^s$ *und* $\varepsilon > 0$ *bel. Ist ferner* $R_N = R_N(\mathrm{x}_1, \dots, \mathrm{x}_N) = I_N(f) - \int\limits_{E_s} f(\mathrm{x})\,d\mathrm{x}$, *so gilt:*

(i) *Für* $\alpha \geqq 2$ *ist:* $|R_N| \leqq CF_N^2(\mathrm{x}_1, \dots, \mathrm{x}_N)$,

(ii) *für* $\frac{3}{2} \leqq \alpha \leqq 2$ *ist:* $|R_N| \leqq C_1(\alpha, \varepsilon) F_N^{\beta_1(\alpha,\varepsilon)}$ *mit* $C_1(\alpha, \varepsilon) = C(8/\varepsilon)^{s(2-\alpha)}$ *und* $\beta_1(\alpha, \varepsilon) = 2\alpha - 2 - \varepsilon(2-\alpha)$,

(iii) *für* $1 < \alpha \leqq \frac{3}{2}$ *ist:* $|R_N| \leqq C_2(\alpha, \varepsilon) F_N^{\beta_2(\alpha,\varepsilon)}$ *mit* $C_2(\alpha, \varepsilon) = C\left(\frac{4}{\varepsilon(\alpha-1)}\right)^s$ *und* $\beta_2(\alpha, \varepsilon) = 2\alpha - 2 - \varepsilon(\alpha - 1)$.

BEWEIS.

$$R_N(\mathrm{x}_1, \dots, \mathrm{x}_N) = \frac{1}{N^2} \sum_{k,l=1}^{N} \left(\sum_{\mathrm{m}=-\infty}^{\infty} C(\mathrm{m}) \exp\left(\mathrm{m} \cdot (\mathrm{x}_k - \mathrm{x}_l)\right)\right) - C(0) =$$

$$= \sum_{\mathrm{m}=-\infty}^{\infty}{}' \, C(\mathrm{m}) \frac{1}{N} \sum_{k=1}^{N} \exp(\mathrm{m} \cdot \mathrm{x}_k) \frac{1}{N} \sum_{l=1}^{N} \exp(-\mathrm{m} \cdot \mathrm{x}_l),$$

wobei die Umordnung der Reihe wegen $\alpha > 1$ erlaubt ist. Mit $S(\mathrm{m}, N) = \left| \frac{1}{N} \sum_{k=1}^{N} \exp(\mathrm{m} \cdot \mathrm{x}_k) \right|$ erhalten wir für $\alpha \geqq 2$ die Abschätzung

$$|R_N| \leqq C \sum_{\mathrm{m}=-\infty}^{\infty}{}' \frac{S^2(\mathrm{m}, N)}{R^\alpha(\mathrm{m})} \leqq C \sum_{\mathrm{m}=-\infty}^{\infty}{}' \left(S(\mathrm{m}, N)/R(\mathrm{m})\right)^2 = CF_N^2.$$

Sei nun $\frac{3}{2} \leqq \alpha \leqq 2$. Mit der Hölderschen Ungleichung erhalten wir dann

$$|R_N| \leqq C \sum_{\mathrm{m}=-\infty}^{\infty}{}' \left(S(\mathrm{m}, N)/R(\mathrm{m})\right)^{3\alpha-4} \frac{\left(S(\mathrm{m}, N)\right)^{3(2-\alpha)}}{\left(R(\mathrm{m})\right)^{2(2-\alpha)}} \leqq$$

$$\leqq CF_N^{3\alpha-4} \left(\sum_{\mathrm{m}=-\infty}^{\infty}{}' S^2(\mathrm{m}, N)/R^{4/3}(\mathrm{m}) \right)^{3(2-\alpha)/2} =$$

$$= CF_N^{3\alpha-4} \left(\sum_{\mathrm{m}=-\infty}^{\infty}{}' \left(S(\mathrm{m}, N)/R(\mathrm{m})\right)^{1/3} \frac{\left(S(\mathrm{m}, N)\right)^{5/3}}{R(\mathrm{m})} \right)^{3(2-\alpha)/2} \leqq$$

$$\leqq CF_N^{3\alpha-4} F_N^{(2-\alpha)/2} \left(\sum_{\mathrm{m}=-\infty}^{\infty}{}' S^2(\mathrm{m}, N)/R^{6/5}(\mathrm{m}) \right)^{5(2-\alpha)/4}.$$

Wenden wir nun wiederholt nach diesem Schema die Höldersche Ungleichung an, so erhalten wir nach k Schritten die Abschätzung

$$|R_N| \leqq CF_N^{\delta(\alpha,k)} \left(\sum_{\mathrm{m}=-\infty}^{\infty}{}' S^2(\mathrm{m})/\left(R(\mathrm{m})\right)^{\beta(k)} \right)^{\gamma(k)} \tag{2.8}$$

mit $\beta(k)=(2^{k+1}+2)/(2^{k+1}+1)$, $\gamma(k)=(2^{k+1}+1)/2^{k+1}$ und

$$\delta(\alpha, k) = 3\alpha-4+2(2-\alpha)(2^{-1}+2^{-2}+\ldots+2^{-k}).$$

Wegen $\sum\limits_{\mathrm{m}=-\infty}^{\infty}{}' S^2(\mathrm{m}, N)/(R(\mathrm{m}))^{\beta(k)} \leqq 2^{s(k+2)}$ und $(k+2)\gamma(k) \leqq k+3$ für genügend große k, erhalten wir aus (2.8)

$$|R_N| \leqq C 2^{s(k+3)(2-\alpha)} F_N^{2\alpha-2-(2-\alpha)2^{-k}} \leqq C_1(\alpha, \varepsilon) F_N^{\beta_1(\alpha, \varepsilon)}$$

mit $C_1(\alpha, \varepsilon)=C(8/\varepsilon)^{s(2-\alpha)}$ und $\beta_1(\alpha, \varepsilon)=2\alpha-2-\varepsilon(2-\alpha)$, wobei k so groß zu wählen ist, daß $2^{-k-1} \leqq \varepsilon < 2^{-k}$ gilt.

Ist nun schließlich $1<\alpha\leqq\frac{3}{2}$, so erhalten wir ebenfalls wieder mit Hilfe der Hölderschen Ungleichung die Abschätzung

$$|R_N| \leqq C \sum_{\mathrm{m}=-\infty}^{\infty}{}' (S(\mathrm{m}, N)/R(\mathrm{m}))^{\alpha-1} \frac{(S(\mathrm{m}, N))^{3-\alpha}}{R(\mathrm{m})} \leqq$$

$$\leqq C F_N^{\alpha-1} \left(\sum_{\mathrm{m}=-\infty}^{\infty}{}' (S(\mathrm{m}, N)/R(\mathrm{m}))^{r_1-1} \frac{(S(\mathrm{m}, N))^{3-r_1}}{R(\mathrm{m})} \right)^{s_1}$$

mit $r_1=2/(3-\alpha)$ und $s_1=(3-\alpha)/2$. Wenden wir schließlich die Höldersche Ungleichung auf die verbleibende konvergente Reihe k mal nach diesem Schema an, so erhalten wir

$$|R_N| \leqq C F_N^{(\alpha-1)(1+2^{-1}+\cdots+2^{-k})} \left(\sum_{\mathrm{m}=-\infty}^{\infty}{}' S^2(\mathrm{m}, N)/(R(\mathrm{m}))^{r_{k+1}} \right)^{s_{k+1}} \tag{2.9}$$

mit $r_k=2(2^k-1-(2^{k-1}-1)\alpha)/(2^{k+1}-1-(2^k-1)\alpha)$ und $S_k=2^{-k}(2^{k+1}-1-(2^k-1)\alpha)$. Wird nun die in (2.9) auftretende Reihe in gewohnter Weise abgeschätzt, erhalten wir analog zu obigen Abschätzungen

$$|R_N| \leqq C F_N^{2(\alpha-1)(1-2^{-k-1})} (2^{k+2}/(\alpha-1))^s \leqq F_N^{\beta_2(\alpha, \varepsilon)} C_2(\alpha, \varepsilon)$$

mit $C_2(\alpha, \varepsilon)=C(4/(\varepsilon(\alpha-1)))^s$ und $\beta_2(\alpha, \varepsilon)=2(\alpha-1)-\varepsilon(\alpha-1)$, wobei k wiederum so groß zu wählen ist, daß $2^{-k-1} \leqq \varepsilon < 2^{-k}$ gilt. Damit ist der Satz gezeigt.

Bemerkung. Wie man sich leicht überlegt, gilt (ii) sogar für $\alpha>4/3$ und (iii) für $\alpha>1$. Die Abschätzung (iii) ist aber für $4/3<\alpha\leqq3/2$ besser als (ii); entsprechendes gilt für $\alpha>3/2$ auch für die Abschätzung (ii).

Satz 2.3. *Sei* $t \geqq \alpha \geqq 2$ $(t\in\mathbf{N})$, $f(\mathrm{x})\in E_\alpha^s(C)$ *und* $\theta\in\mathbf{R}^s$, *so gilt*

$$|R_N| = |J_N^{(t)}(f) - \int_{E_s} f(\mathrm{x})\, d\mathrm{x}| \leqq C F_N^{\alpha}. \tag{2.10}$$

Beweis.

$$R_N = (N+1)^{-t} \sum_{\mathrm{m}=-\infty}^{\infty}{}' C(\mathrm{m}) \sum_{k=0}^{Nt} A_{k,N}^{(t)} \exp(\mathrm{m}\cdot k\theta) =$$

$$= (N+1)^{-t} \sum_{\mathrm{m}=-\infty}^{\infty}{}' C(\mathrm{m}) \left(\sum_{k=0}^{N} \exp(\mathrm{m}\cdot k\theta) \right)^t.$$

Damit ergibt sich

$$|R_N| \leqq C\left(\frac{N}{N+1}\right)^t \sum_{\mathrm{m}=-\infty}^{\infty}{}' (S(\mathrm{m}, N)/R(\mathrm{m}))^\alpha \leqq C\left(\sum_{\mathrm{m}=-\infty}^{\infty}{}' S^2(\mathrm{m}, N)/R^2(\mathrm{m})\right)^{\alpha/2} = CF_N^\alpha,$$

wobei letzte Abschätzung aus $\|\{a_n\}\|_p \leqq \|\{a_n\}\|_q$ für $p \geqq q \geqq 1$ folgt.

Sind von der Funktion $f(\mathrm{x}) \in E_\alpha^s(C)$ die Funktionswerte nur an den endlich vielen Stellen $\mathrm{x}_1, \ldots, \mathrm{x}_N$ $(\mathrm{x}_j \in \mathbf{R}^s, 1 \leqq j \leqq N)$ bekannt, so kann bei der Verwendung des arithmetischen Mittels als numerische Integrationsmethode folgende Fehlerabschätzung hergeleitet werden:

SATZ 2.4. *Ist* $f \in E_\alpha^s(C)$ *mit* $\alpha > \frac{3}{2}$ *und* $\mathrm{x}_1, \ldots, \mathrm{x}_N \in \mathbf{R}^s$, *dann gilt für den Integrationsfehler* R_N

$$R_N(\mathrm{x}_1, \ldots, \mathrm{x}_s) = \frac{1}{N} \sum_{k=1}^{N} f(\mathrm{x}_k) - \int_{E_s} f(\mathrm{x})\, d\mathrm{x}$$

die Abschätzung

$$|R_N(\mathrm{x}_1, \ldots, \mathrm{x}_N)| \leqq C \cdot F_N \cdot \left(\frac{2\alpha-2}{2\alpha-3}\right)^{1/2}. \tag{2.11}$$

Der Beweis ergibt sich ohne Mühe aus der eindimensionalen Version des Satzes (siehe [6], Satz 3).

3. Trigonometrische Approximation

Wir wenden nun die in den letzten Paragraphen angeführten Methoden und Ergebnisse auf die Approximation von Funktionen an. Es ist naheliegend, als trigonometrische Approximationspolynome „Partialsummen" der Fourierreihe der Funktion zu verwenden. Die Verallgemeinerung des Begriffs der Partialsumme $s_N(x) = \sum_{m=-N}^{N} c_m \exp(mx)$ vom ein- in den mehrdimensionalen Fall kann aber in verschiedener Weise erfolgen. In der Literatur findet man vor allem zwei solcher Verallgemeinerungen (siehe [7]).

$$s_n(x) = s_{(n_1, \ldots, n_s)}(x) = \sum_{\substack{0 \leqq |k_j| \leqq n_j \\ j=1, \ldots, s}} C(k_1, \ldots, k_s) \exp(\mathrm{k} \cdot \mathrm{x}) \tag{3.1}$$

(Summation über Quaderbereiche).

$$s_R(x) = \sum_{k_1^2 + \ldots + k_s^2 \leqq R} C(k_1, \ldots, k_s) \exp(\mathrm{k} \cdot \mathrm{x}) \tag{3.2}$$

(Summation über Kugelbereiche).

Wir werden nun „Partialsummen" über bestimmte hyperbolische Bereiche definieren und zeigen, daß damit wesentlich bessere Fehlerabschätzungen als mit Summen der Gestalt (3.1) und (3.2) gewonnen werden können. Im folgenden werden zwei Approximationspolynome konstruiert; im ersten Fall berechnen wir uns die Fourierkoeffizienten mit einer Integrationsmethode (2.1), (2.4) oder (2.11).

LEMMA 3.1. *Es sei* $f(\mathrm{x})\in E_\alpha^s(C)$, $\alpha>1$, $C^*(\mathrm{m})=I_N\big(f(\mathrm{x})\exp(-\mathrm{m}\cdot\mathrm{x})\big)$ *und* $\varepsilon>0$ *bel. Dann gilt*

$$(3.3)\qquad |C(\mathrm{m})-C^*(\mathrm{m})|\leqq\begin{cases} CR^\alpha(\mathrm{m})\,F_N^2 & \textit{für}\quad \alpha\geqq 2,\\ C_1(\alpha,\varepsilon)\,R^\alpha(\mathrm{m})\,F_N^{\beta_1(\alpha,\varepsilon)} & \textit{für}\quad \dfrac{3}{2}\leqq\alpha\leqq 2,\\ C_2(\alpha,\varepsilon)\,R^\alpha(\mathrm{m})\,F_N^{\beta_2(\alpha,\varepsilon)} & \textit{für}\quad 1<\alpha\leqq\dfrac{3}{2}.\end{cases}$$

BEWEIS. Wie man sich sofort überzeugt, liegt die Funktion $f(\mathrm{x})\exp(\mathrm{m}\cdot\mathrm{x})$ in der Klasse $E_\alpha^s(CR^\alpha(\mathrm{m}))$, falls $f\in E_\alpha^s(C)$. Der Rest ergibt sich aus Satz 2.2.

SATZ 3.1. *Es sei* $f(\mathrm{x})\in E_\alpha^s(C)$, $\alpha>1$, $\varepsilon>0$ *bel. und* $R\geqq 1$. *Dann gilt für das Approximationspolynom*

$$P_R(\mathrm{x})=\sum_{R(\mathrm{m})\leqq R}C^*(\mathrm{m})\exp(\mathrm{m}\cdot\mathrm{x})$$

die Abschätzung

$$(3.4)\qquad \|f-P_R\|_\infty=O\left(F_N^\delta\left(\log\frac{1}{F_N}\right)^{s-1}\right)$$

mit

$$\delta=\begin{cases}\dfrac{\alpha-1}{\alpha} & \alpha\geqq 2,\\ \dfrac{1+\varepsilon}{\alpha}+\alpha-2-\dfrac{\varepsilon}{2}(3-\alpha) & \dfrac{3}{2}\leqq\alpha\leqq 2,\\ \dfrac{2-\varepsilon}{2}\left(\alpha+\dfrac{1}{\alpha}-2\right) & 1<\alpha\leqq\dfrac{3}{2}.\end{cases}$$

BEWEIS.

$$|f(\mathrm{x})-P_R(\mathrm{x})|=\Big|\sum_{R(\mathrm{m})\leqq R}\big(C(\mathrm{m})-C^*(\mathrm{m})\big)\exp(\mathrm{m}\cdot\mathrm{x})+$$

$$+\sum_{R(\mathrm{m})>R}C(\mathrm{m})\exp(\mathrm{m}\cdot\mathrm{x})\Big|\leqq C^*F_N^\beta\sum_{R(\mathrm{m})\leqq R}R^\alpha(\mathrm{m})+C\sum_{R(\mathrm{m})>R}R^{-\alpha}(\mathrm{m})\leqq$$

$$\leqq C^*F_N^\beta R^\alpha\sum_{R(\mathrm{m})\leqq R}1+C\sum_{R(\mathrm{m})>R}R^{-\alpha}(\mathrm{m})\leqq$$

$$\leqq C^*B_sF_N^\beta R^\alpha R\log^{s-1}R+C\alpha\left(\frac{\alpha}{\alpha-1}\right)^s\frac{(1+\log R)^{s-1}}{R^{\alpha-1}},$$

wobei die letzten Abschätzungen aus (3.3), [2] p. 172 und [2] p. 125 folgen. Dabei ist

$$C^*=\begin{cases}C & \alpha\geqq 2\\ C_1(\alpha,\varepsilon) & \dfrac{3}{2}\leqq\alpha\leqq 2\\ C_2(\alpha,\varepsilon) & 1<\alpha\leqq\dfrac{3}{2}\end{cases}\qquad \beta=\begin{cases}2 & \alpha\geqq 2\\ 2\alpha-2-\varepsilon(2-\alpha) & \dfrac{3}{2}\leqq\alpha\leqq 2\\ 2\alpha-2-\varepsilon(\alpha-1) & 1<\alpha\leqq\dfrac{3}{2}.\end{cases}$$

Wir wählen nun $R=R(F_N)$ so, daß dieser Ausdruck minimal wird und erhalten $R=F_N^\gamma$ mit

$$\gamma=\begin{cases} \dfrac{1}{\alpha} & \alpha \geqq 2 \\ 1+\dfrac{\varepsilon}{2}-\dfrac{1+\varepsilon}{\alpha} & \dfrac{3}{2} \leqq \alpha \leqq 2 \\ \dfrac{2-\varepsilon}{2}\left(1-\dfrac{1}{\alpha}\right) & 1<\alpha \leqq \dfrac{3}{2} \end{cases}$$

woraus schließlich die Behauptung folgt.

SATZ 3.2. *Es sei* $t \geqq \alpha \geqq 2$, $f \in E_\alpha^s(C)$, $R \geqq 1$ *und* $C^{**}(\mathrm{m})=J_N^{(t)}(f(\mathrm{x})\exp(-m\cdot \mathrm{x}))$. *Dann gilt für das Approximationspolynom*

$$P_R^{(t)}(\mathrm{x}) = \sum_{R(m)\leqq R} C^{**}(\mathrm{m})\exp(\mathrm{m}\cdot \mathrm{x})$$

die Abschätzung

$$\|f-P_R^{(t)}\|_\infty = O\left(F_N^{(\alpha-1)/2}\log^{s-1}\frac{1}{F_N}\right). \tag{3.5}$$

Der Beweis verläuft ganz analog wie oben unter Beachtung von (2.10).

Analoge Abschätzungen erhält man bei Verwendung der Integrationsmethode (2.11).

BEMERKUNG. Wird in (3.1) $n_1=\ldots=n_s=R$ gewählt, so hat man bei den Methoden (3.1) und (3.2) ein $O(R^s)$ Fourierkoeffizienten zu berechnen, während die Anzahl der zu schätzenden Koeffizienten mit den Methoden (3.4) und (3.5) nach [2], p. 172 nur ein $O(R\log^{s-1}R)$ beträgt.

Approximationsmethoden, wie sie in Satz 3.1 und 3.2 angeführt wurden, wird man nur dann verwenden, wenn die Koeffizienten $C(\mathrm{m})$ bzw. deren Schätzungen ebenfalls benötigt werden (etwa für gewisse Stopbedingungen oder in der Bildverarbeitung zum Ausfiltern von Störfunktionen). Ist man jedoch nur an den zu approximierenden Funktionswerten interessiert, so braucht man nicht die einzelnen Fourierkoeffizienten getrennt zu berechnen, sondern kann sich gewisser Integraldarstellungen für die Partialsummen bedienen.

DEFINITION 3.1. *Sei* $\mathrm{n}=(n_1, \ldots, n_s)$ *und* $R \geqq 1$, *ganz. Unter dem s-dimensionalen Dirichlet-Kern der Ordnung* n *verstehen wir die Exponentialsumme*

$$D_{\mathrm{n}}^{(s)}(\mathrm{x}) = \sum_{\substack{-n_j \leqq m_j \leqq n_j \\ j=1,\ldots,s}} \exp(\mathrm{m}\cdot \mathrm{x}).$$

Die Exponentialsumme

$$H_R^{(s)}(\mathrm{x}) = \sum_{R(\mathrm{m})\leqq R} \exp(\mathrm{m}\cdot \mathrm{x}) \tag{3.6}$$

wollen wir als s-dimensionalen Hyperbelkern der Ordnung R bezeichnen.

Es ist bekannt, daß $D_n^{(s)}(\mathrm{x})$ die explizite Darstellung

$$D_{\mathrm{n}}^{(s)}(\mathrm{x}) = D_{n_1}(x_1)\dots D_{n_s}(x_s)$$

mit

$$D_k(t) = \frac{\sin(2k+1)\pi t}{\sin \pi t} \tag{3.7}$$

besitzt. Wie man sofort nachrechnet, gelten die Gleichungen

$$s_{\mathrm{n}}(\mathrm{x}) = f * D_{\mathrm{n}}^{(s)}(\mathrm{x}) = \int\limits_{E_s} f(\mathrm{y}) D_{\mathrm{n}}^{(s)}(\mathrm{x}-\mathrm{y})\,dy, \tag{3.8}$$

$$\sum_{R(\mathrm{m})\leqq R} C(\mathrm{m}) \exp(\mathrm{m}\cdot\mathrm{x}) = f * H_R^{(s)}(\mathrm{x}) = \int\limits_{E_s} f(\mathrm{y}) H_R^{(s)}(\mathrm{x}-\mathrm{y})\,dy. \tag{3.9}$$

Auf Grund der Darstellungen (3.8) und (3.9) kann man also ein trigonometrisches Approximationspolynom erhalten, wenn man auf die Funktionen $f(\mathrm{y})D_n^{(s)}(\mathrm{x}-\mathrm{y})$ bzw. $f(\mathrm{y})H_R^{(s)}(\mathrm{x}-\mathrm{y})$ eine numerische Integrationsmethode anwendet. Für (3.9) ist dies natürlich in der Praxis nur dann sinnvoll, wenn wir für (3.6) ebenfalls eine explizite Darstellung ähnlich (3.7) angeben können. Eine solche soll nun im Anschluß hergeleitet werden. Wir werden dazu den Bereich der Gitterpunkte m mit $R(\mathrm{m})\leqq R$ in einer geeigneten Weise charakterisieren. Für $R\in\mathbf{N}$ bel. definieren wir uns dazu folgende Zahlen:

(i) Ist $\sqrt{R+1}=k+r$ mit $0<r\leqq 0.5$, so sei

$$r_j = \begin{cases} j & 0\leqq j\leqq k=[\sqrt{R}], \\ \left[\dfrac{R}{2k-j}\right] & k<j\leqq 2k-1, \\ \infty & j=2k. \end{cases} \tag{3.10}$$

(ii) Ist andernfalls $\sqrt{R+1}=k+r$ mit $0.5<r\leqq 1$, so sei

$$r_j = \begin{cases} j & 0\leqq j\leqq k, \\ \left[\dfrac{R}{2k+1-j}\right] & k+1\leqq j\leqq 2k, \\ \infty & j=2k+1. \end{cases} \tag{3.11}$$

DEFINITION 3.2. *Unter dem Bereich $B_R^{(s)}$ verstehen wir die Menge aller Gitterpunkte* m *für die gilt: Ist $r_{\alpha_j-1}<\bar{n}_j\leqq r_{\alpha_j}$ für $1\leqq\alpha_j\leqq 2k$ (bzw. $1\leqq\alpha_j\leqq 2k+1$) $(j=1,\dots,s)$, so gehöre $\mathrm{n}=(n_1,\dots,n_s)$ genau dann zu $B_R^{(s)}$, wenn $\prod_{j=1}^{s} r_{\alpha_j}\leqq R$ gilt.*

Die Eindeutigkeit dieser Definition zeigen wir in

LEMMA 3.2. *Die in* (3.10) *bzw.* (3.11) *definierten Zahlen sind alle voneinander verschieden.*

BEWEIS. Für $1\leqq j\leqq k$ ist nichts zu zeigen. Aus der Gültigkeit von $\dfrac{R}{l}<\dfrac{R}{l-1}-1$ für $1\leqq l\leqq k$ und $\left[\dfrac{R}{l}\right]\leqq\dfrac{R}{l}<\dfrac{R}{l-1}-1<\left[\dfrac{R}{l-1}\right]$ folgt auch der Rest der Behauptung im Falle (3.10), da ja hier $k=\left[\frac{R}{k}\right]$ gilt. Im Fall (3.11) müssen wir nur noch zeigen, daß $\left[\frac{R}{k}\right]\geqq k+1$ gilt. Wegen $r>\frac{1}{2}$, d. h. $2kr+r^2>k+\frac{1}{4}$ und $R+1=k^2+$ $+2kr+r^2$, d. h. $2kr+r^2\in\mathbf{N}$ erhalten wir die Ungleichung $2kr+r^2\geqq k+1$, woraus schließlich

$$r\geqq\sqrt{k^2+k+1}-k \tag{3.12}$$

folgt. Mit Hilfe von (3.12) läßt sich nun aber sofort zeigen, daß dann auch $k(k+1)\geqq R$ gilt. Aus dieser Ungleichung kann dann aber sofort die Behauptung hergeleitet werden.

Das nachfolgende Lemma wird für die explizite Darstellung von $H_R^{(s)}(\mathrm{x})$ eine große Rolle spielen.

LEMMA 3.3. *Ein Gitterpunkt* $\mathrm{n}=(n_1,\ldots,n_s)$ *gehört genau dann zu* $B_R^{(s)}$, *wenn* $R(\mathrm{n})\leqq R$ *gilt.*

BEWEIS. Die eine Richtung ist trivial, d. h. $\mathrm{n}\in B_R^{(s)}\Rightarrow R(\mathrm{n})\leqq R$. Nehmen wir also an, daß $R(\mathrm{n})\leqq R$ gilt und sei $k=[\sqrt{R}]$ wie in (3.10) bzw. (3.11) definiert. Gilt für ein $\bar{n}_j$: $\bar{n}_j>k$, so folgt daraus $\bar{n}_l\leqq k$ für $l\neq k$ ($j,l=1,\ldots,s$). O. B. d. A. können wir annehmen $\bar{n}_l\leqq k$ für $1\leqq l\leqq s-1$ und nur eventuell $\bar{n}_s>k$. Es gilt somit auf Grund der Definition 3.2: $r_{\alpha_l}=\bar{n}_l$ für $1\leqq l\leqq s-1$. Wegen $R(\mathrm{n})\leqq R$ folgt daher

$$\bar{n}_s\leqq\left[\frac{R}{\bar{n}_1\ldots\bar{n}_{s-1}}\right].$$

Nun ist aber $\left[\frac{R}{\bar{n}_1\ldots\bar{n}_{s-1}}\right]$ stets eine in der Folge (3.10) bzw. (3.11) vorkommende Zahl r_m. Ist daher $\bar{n}_s\leqq r_m=r_{\alpha_s}$, so gilt

$$\prod_{j=1}^{s}r_{\alpha_j}=\bar{n}_1\ldots\bar{n}_{s-1}\left[\frac{R}{\bar{n}_1\ldots\bar{n}_{s-1}}\right]\leqq R,$$

was aber $\mathrm{n}\in B_R^{(s)}$ impliziert. Damit ist das Lemma bewiesen.

Dieses Lemma erlaubt nun eine rekursive Darstellung des in (3.6) definierten Hyperbelkernes $H_R^{(s)}(\mathrm{x})$. Im folgenden wollen wir für $\mathrm{x}=(x_1,\ldots,x_{s-1},x_s)\in\mathbf{R}^s$ mit $\mathrm{x}'=(x_1,\ldots,x_{s-1})$ seine Projektion auf $\mathbf{R}^{s-1}$ bezeichnen.

SATZ 3.3. *Für den in* (3.6) *definierten Kern* $H_{\mathrm{R}}^{(s)}(\mathrm{x})$ $(s\geqq2)$ *gelten die Rekursionsformeln*

$$H_R^{(s)}(\mathrm{x})=\sum_{j=1}^{2k-1}H_{r_{2k-j}}^{(s-1)}(x_1,\ldots,x_{s-1})\big(D_{r_j}(x_s)-D_{r_{j-1}}(x_s)\big) \tag{i}$$

für $\sqrt{R+1}=k+r,\ 0<r\leqq0.5,$

$$H_R^{(s)}(\mathrm{x})=\sum_{j=1}^{2k}H_{r_{2k+1-j}}^{(s-1)}(x_1,\ldots,x_{s-1})\big(D_{r_j}(x_s)-D_{r_{j-1}}(x_s)\big) \tag{ii}$$

für $\sqrt{R+1}=k+r$, $0.5<r\leqq 1$.

$D_{r_0}(t)=D_0(t)$ *ist dabei identisch Null zu setzen und* $H^{(1)}_{r_j}(t)=D_{r_j}(t)$.

Beweis. Wir beweisen etwa (ii) und vermerken, daß $r_{2k}=R$ gilt.

$$H_R^{(s)}(\mathrm{x}) = \sum_{R(\mathrm{m})\leqq R} \exp(\mathrm{m}\cdot\mathrm{x}) = \sum_{\mathrm{m}\in B_R^{(s)}} \exp(\mathrm{m}\cdot\mathrm{x}) =$$

$$= \sum_{r_0\leqq \bar m_s\leqq r_1} \exp(m_s x_s) \sum_{m'\in B_{r_{2k}}^{(s-1)}} \exp(\mathrm{m}'\cdot\mathrm{x}')+\ldots+$$

$$+ \sum_{r_j<\bar m_s\leqq r_{j+1}} \exp(m_s x_s) \sum_{\mathrm{m}'\in B_{r_{2k-j}}^{(s-1)}} \exp(\mathrm{m}'\cdot\mathrm{x}')+\ldots+$$

$$+ \sum_{r_{2k-1}<\bar m_s\leqq r_{2k}} \exp(m_s x_s) \sum_{\mathrm{m}'\in B_{r_1}^{(s-1)}} \exp(\mathrm{m}'\cdot\mathrm{x}') =$$

$$= D_{r_1}(x_s)\, H_{r_{2k}}^{(s-1)}(\mathrm{x}')+\ldots+\big(D_{r_{j+1}}(x_s)-D_{r_j}(x_s)\big)\, H_{r_{2k-j}}^{(s-1)}(\mathrm{x}')+\ldots+$$

$$+\ldots+\big(D_{r_{2k}}(x_s)-D_{r_{2k-1}}(x_s)\big)\, H_{r_1}^{(s-1)}(\mathrm{x}').$$

Dies entspricht jedoch schon der Formel (ii) unter Berücksichtigung von $D_0(t)=0$.

Mit Hilfe dieser Rekursionsformeln kann somit der s-dimensionale Hyperbelkern aus $(s-1)$ dimensionalen Hyperbelkernen und gewöhnlichen Dirichlet Kernen aufgebaut werden. Ein gewisser Nachteil liegt allerdings darin, daß die Berechnung von $H_R^{(s)}(\mathrm{x})$ für große R sehr aufwendig wird. Wir werden daher einen Kern $G_R^{(s)}(\mathrm{x})$ definieren, dessen Berechnung wesentlich einfacher ist und der für genügend große R die gleichen Fehlerabschätzungen wie (3.6) gestattet. Dazu geben wir folgende Definition, die der Einfachheit halber nur für die speziellen Werte $R=2^k$ $(k\in\mathbf{N})$ angegeben wird (hat R nicht diese Gestalt, so sind wieder mehrere Fallunterscheidungen vorzunehmen).

Definition 3.3. *Es sei* $R=2^k$ $(k\in\mathbf{N})$ *und* $q_j=2^j$ *für* $0\leqq j\leqq k$ *sowie* $q_{-1}=0$ *und* $q_{k+1}=\infty$. *Ist* $\mathrm{m}\in\mathbf{Z}^s$ *mit*

$$q_{\alpha_j-1}<\bar m_j\leqq q_{\alpha_j}\quad (0\leqq\alpha_j\leqq k+1)$$

so gehöre m *genau dann zur Menge* $A_R^{(s)}$, *wenn* $\prod_{i=1}^{s} q_{\alpha_i}\leqq R$ *gilt.*

Aus der Definition folgt sofort, daß die Inklusion $A_R^{(s)}\subseteqq B_R^{(s)}$ richtig ist. Einfache Gegenbeispiele zeigen, daß im Allgemeinen nicht das Gleichheitszeichen gilt.

Definition 3.4. *Der Kern* $G_R^{(s)}(x)$ *werde definiert durch die Gleichung*

$$G_R^{(s)}(\mathrm{x}) = \sum_{\mathrm{m}\in A_R^{(s)}} \exp(\mathrm{m}\cdot\mathrm{x}). \tag{3.13}$$

Ganz analog wie in Satz 3.3 zeigt man folgende Rekursionsformel für den Kern $G_R^{(s)}(\mathrm{x})$:

$$G_R^{(s)}(\mathrm{x}) = \sum_{j=0}^{k} G_{q_{k-j}}^{(s-1)}(\mathrm{x}')\big(D_{q_j}(x_s)-D_{q_{j-1}}(x_s)\big), \tag{3.14}$$

wobei wiederum dieselben Nebenbedingungen wie in Satz 3.3 zu gelten haben.

Da alle Fourierkoeffizienten $C(\mathrm{m})$ von $H_R^{(s)}$ bzw. $G_R^{(s)}$ für $R(\mathrm{m})>R$ verschwinden und für $R(\mathrm{m})\leqq R$ höchstens gleich 1 sind, erhalten wir die Abschätzung

$$|C(\mathrm{m})| \leqq \frac{R^\alpha}{R^\alpha(\mathrm{m})} \qquad \forall\, \mathrm{m}\in \mathbf{Z}^s, \tag{3.15}$$

was aber heißt, daß $H_R^{(s)}$ und $G_R^{(s)}$ zur Klass $E_\alpha^s(R^\alpha)$ ($\alpha>1$ bel.) gehören.

Mit Hilfe dieser Kerne $H_R^{(s)}$ und $G_R^{(s)}$ können wir nun für Funktionen aus $E_\alpha^s(C)$ folgende Approximationspolynome definieren:

Definition 3.5. *Es sei* $f\in E_\alpha^s(C)$, $\alpha>1$ *und* $t\geqq\alpha$.

$$HI_R(\mathrm{x}) = I_N\big(f(\mathrm{y})\,H_R^{(s)}(\mathrm{x}-\mathrm{y})\big), \tag{3.16}$$

$$GI_R(\mathrm{x}) = I_N\big(f(\mathrm{y})\,G_R^{(s)}(\mathrm{x}-\mathrm{y})\big), \tag{3.17}$$

$$HJ_R^{(t)}(\mathrm{x}) = J_N^{(t)}\big(f(\mathrm{y})\,H_R^{(s)}(\mathrm{x}-\mathrm{y})\big), \tag{3.18}$$

$$GJ_R^{(t)}(\mathrm{x}) = J_N^{(t)}\big(f(\mathrm{y})\,G_R^{(s)}(\mathrm{x}-\mathrm{y})\big). \tag{3.19}$$

Für diese Approximationspolynome gelten folgende Abschätzungen:

Satz 3.4. *Sei* $f(\mathrm{x})\in E_\alpha^s(C)$ *mit* $\alpha>1$ *und* $\varepsilon>0$ *bel. Dann gelten für die in* (3.16) *und* (3.17) *definierten Approximationspolynome folgende Abschätzungen:*

$$\left.\begin{matrix}\|f-HI_R\|_\infty\\ \|f-GI_R\|_\infty\end{matrix}\right\} = O\big(F_N^{\gamma(\alpha,\varepsilon)}(-\log F_N)^{\delta(\alpha,s)}\big)$$

mit $\delta(\alpha,s)=\dfrac{\alpha(s-1)}{2\alpha-1}$, $\gamma(\alpha,\varepsilon)=\dfrac{2(\alpha-1)}{2\alpha-1}$ *für* $\alpha\geqq 2$, $\gamma(\alpha,\varepsilon)=\dfrac{\alpha-1}{2\alpha-1}\beta_1(\alpha,\varepsilon)$ *für* $\dfrac{3}{2}\leqq\alpha\leqq 2$ *und* $\gamma(\alpha,\varepsilon)=\dfrac{\alpha-1}{2\alpha-1}\beta_2(\alpha,\varepsilon)$ *für* $1<\alpha\leqq\dfrac{3}{2}$. R *muß dabei wie folgt in Abhängigkeit von* F_N *gewählt werden:* $R=\left[\dfrac{(-\log F_N)^{a(\alpha,s)}}{F_N^{b(\alpha)}}\right]$ *mit* $a(\alpha,s)=\dfrac{s-1}{2\alpha-1}$, $b(\alpha)=\dfrac{2}{2\alpha-1}$ *für* $\alpha\geqq 2$, $b(\alpha)=\dfrac{\beta_1(\alpha,\varepsilon)}{2\alpha-1}$ *für* $\dfrac{3}{2}\leqq\alpha\leqq 2$ *und* $b(\alpha)=\dfrac{\beta_2(\alpha,\varepsilon)}{2\alpha-1}$ *für* $1<\alpha\leqq\dfrac{3}{2}$.

Beweis. $|f(\mathrm{x})-HI_R(\mathrm{x})|\leqq|f(\mathrm{x})-f*H_R^{(s)}(\mathrm{x})|+|f*H_R^{(s)}(\mathrm{x})-HI_R(\mathrm{x})|$. Für den ersten Summanden erhalten wir wegen (3.9) und Lemma 28 von [2] die Abschätzung

$$|f(\mathrm{x})-f*H_R^{(s)}(\mathrm{x})| \leqq \Big|\sum_{R(\mathrm{m})>R} C(\mathrm{m})\exp(\mathrm{m}\cdot\mathrm{x})\Big| \leqq \tag{3.20}$$

$$\leqq C\sum_{R(\mathrm{m})>R} R^{-\alpha}(\mathrm{m}) \leqq C\alpha\left(\frac{\alpha}{\alpha-1}\right)^s \frac{(1+\log R)^{s-1}}{R^{\alpha-1}}.$$

Wegen (3.15) und Lemma 9 von [2] liegt $f(\mathrm{x})H_R^{(s)}(\mathrm{x})$ in der Klasse $F_\alpha^s(ACR^\alpha)$ mit einer nur von α und s abhängigen Konstanten A. Satz 2.2 liefert somit die Abschätzung

$$\|f*H_R^{(s)}-HI_R\|_\infty \leqq \begin{cases} ACR^\alpha F_N^2 & \text{für} \quad \alpha \geqq 2, \\ AR^\alpha C_1(\alpha, \varepsilon)(F_N)^{\beta_1(\alpha,\varepsilon)} & \text{für} \quad \dfrac{3}{2} \leqq \alpha \leqq 2, \\ AR^\alpha C_2(\alpha, \varepsilon)(F_N)^{\beta_2(\alpha,\varepsilon)} & \text{für} \quad 1 < \alpha \leqq \dfrac{3}{2} \end{cases}$$

(die Bezeichnungen entsprechen jenen von Satz 2.2). Wird nun R wie oben angegeben gewählt, so wird die Abschätzung der beiden Summanden optimal und man erhält nach kurzer Rechnung die gewünschte Abschätzung für das Approximationspolynom HI_R. Um die Abschätzung für GI_R zu erhalten, müssen wir eine Abschätzung von

$$\Big|\sum_{\mathrm{m}\notin A_R} C(\mathrm{m}) \exp(\mathrm{m}\cdot\mathrm{x})\Big|$$

angeben, welche dieselbe Konvergenzrate wie (3.20) besitzen muß. Dies folgt aber unmittelbar aus Lemma 28 von [2] und nachfolgendem Lemma:

LEMMA 3.5. *Für den in Definition* 3.3 *eingeführten Bereich* $A_R^{(s)}$ *gilt die Inklusion:* $B_{R'}^{(s)} \subseteqq A_R^{(s)}$ *für* $R'=R/2^s$.

BEWEIS. Sei $\mathrm{m}\in B_{R'}^{(s)}$, d. h. $R(\mathrm{m})\leqq R/2^s$ und etwa $2^{\alpha_j-1}<\bar{m}_j\leqq 2^{\alpha_j}$, dann gilt $R/2^s\geqq R(\mathrm{m})\geqq 2^{\Sigma\alpha_j}/2^s$ bzw. $2^{\Sigma\alpha_j}\leqq R$. Laut Definition liegt dann aber bereits m in $A_R^{(s)}$ und das Lemma ist bewiesen.

Für die Approximationspolynome $HJ_R^{(t)}$ und $GJ_R^{(t)}$ gelten folgende Abschätzungen:

SATZ 3.5. *Sei* $f\in E_\alpha^s(C)$ *mit* $\alpha\geqq 2$ *und* $t\geqq\alpha$ *(t ganz). Dann gelten für die in* (3.18) *und* (3.19) *definierten Approximationspolynome nachstehende Abschätzungen:*

$$\left.\begin{matrix}\|f-HJ_R^{(t)}\|_\infty \\ \|f-GJ_R^{(t)}\|_\infty\end{matrix}\right\} = O\big((F_N)^{\eta(\alpha)}(-\log F_N)^{\delta(\alpha,s)}\big)$$

mit $\eta(\alpha)=\alpha(\alpha-1)/(2\alpha-1)$ *und* $\delta(\alpha,s)=\alpha(s-1)/(2\alpha-1)$. *$R$ muß dabei wieder in Abhängigkeit von F_N gewählt werden:*

$$R=\left[\frac{(-\log F_N)^{a(\alpha,s)}}{F_N^{c(\alpha)}}\right]$$

mit $a(\alpha,s)=(s-1)/(2\alpha-1)$ *und* $c(\alpha)=\alpha/(2\alpha-1)$.

Der Beweis kann ganz analog zu obigen Satz geführt werden, wenn die Ergebnisse von Satz 2.3 berücksichtigt werden.

Zum Abschluß wollen wir noch die in Satz 3.4 und 3.5 hergeleiteten Ergebnisse mit jenen vergleichen, die man erhält, wenn man der Hyperbelkern durch den Dirichletkern ersetzt. Wie man sich sofort überlegt, gehört $D_{\mathrm{n}}^{(s)}(\mathrm{x})$ für bel. $\alpha>1$ zur Klasse $E_\alpha^s(R^\alpha(\mathrm{n}))$ und daher $f(\mathrm{y})D_{\mathrm{n}}^{(s)}(\mathrm{x}-\mathrm{y})$ zur Klasse $E^s(CAR^\alpha(\mathrm{n}))$ mit

$A=A(s, \alpha)$, falls $f\in E_\alpha^s(C)$. Analog zu (3.18) betrachten wir das Approximationspolynom $DJ_{\mathrm{n}}^{(t)}(\mathrm{x})=J_N^{(t)}(f(\mathrm{y})D_{\mathrm{n}}^{(s)}(\mathrm{x}-\mathrm{y}))$ und erhalten

$$|f(\mathrm{x})-DJ_{\mathrm{n}}^{(t)}| \geqq |f(\mathrm{x})-f*D_{\mathrm{n}}^{(s)}(\mathrm{x})|-|f*D_{\mathrm{n}}^{(s)}(\mathrm{x})-DJ_{\mathrm{n}}^{(t)}(\mathrm{x})| \geqq$$
$$\geqq \Big|\sum_{\mathrm{m}\in Q'(\mathrm{n})} C(\mathrm{m})\exp(\mathrm{m}\cdot\mathrm{x})\Big|-ACR^\alpha(\mathrm{n})\cdot F_N^\alpha$$

mit $Q'(\mathrm{n})=\mathbf{Z}^s\diagdown Q(\mathrm{n})$ und $Q(\mathrm{n})=\{\mathrm{m}\in\mathbf{Z}^s\colon\ |m_i|\leqq n_i,\ i=1, \dots, s\}$. Betrachten wir speziell die Faltung der Funktion $f(x)$ mit Fourierkoeffizienten $C(\mathrm{m})=C/R^\alpha(\mathrm{m})$ mit dem Dirichletkern im Punkte $\mathrm{x}=0$, so erhalten wir für $\mathrm{n}=(n, \dots, n)$

$$\sup_{\mathrm{x}} |f(\mathrm{x})-DJ_{\mathrm{n}}^{(t)}(\mathrm{x})| \geqq \sum_{\mathrm{m}\in Q'(\mathrm{n})} C/R^\alpha(\mathrm{m})-ACn^{\alpha s}F_N^\alpha \geqq$$

$$\geqq 4C/2^\alpha\left(\frac{\alpha+1}{\alpha-1}\right)^s\left(1-\left(\frac{\alpha}{\alpha+1}\right)^s\right)n^{1-\alpha}-ACn^{\alpha s}F_N^\alpha = \tag{3.21}$$

$$= CF_N^{\delta(\alpha,\varepsilon)}\left(4/2^\alpha\left(\frac{\alpha+1}{\alpha-1}\right)^s\left(1-\left(\frac{\alpha}{\alpha+1}\right)^s\right)F_N^{-\eta(\alpha,\varepsilon)}-A\right) \geqq C(\alpha, s)F_N^{\delta(\alpha,\varepsilon)}$$

mit $\delta(\alpha, \varepsilon)=\dfrac{\alpha(\alpha-1+\varepsilon)}{\alpha-1+\varepsilon+\alpha s}$ und $\eta(\alpha, \varepsilon)=\dfrac{\alpha\varepsilon}{\alpha-1+\varepsilon+\alpha s}$. Für hinreichend große N ist dabei die Konstante $C(\alpha, s)>0$. Es existiert somit in jeder Klasse $E_\alpha^s(C)$ $(\alpha>1)$ eine Funktion $f(\mathrm{x})$, für die das mit Hilfe des Dirichletkernes definierte Approximationspolynom keine bessere Abschätzung als obige liefert. Ein Vergleich mit Satz 3.4 zeigt aber, daß die damit erreichte Approximationsgüte nur etwa die s-te Wurzel von jener aus Satz 3.4 beträgt. Analoges gilt natürlich auch für die Approximationspolynome (3.16), (3.17) und (3.19).

Zu den oben angeführten Integrations- und Approximationsmethoden wurde eine Reihe von numerischen Experimenten durchgeführt, über die an einer anderen Stelle berichtet werden wird.

LITERATUR

[1] Hlawka, E., Funktionen von beschränkter Variation in der Theorie der Gleichverteilung, *Annali di Mat. (Ser. IV)* **54** (1961), 325—334.

[2] Korobow, N. M., *Zahlentheoretische Methoden in der Näherungsanalysis,* Moskau: Fismatgis 1963 (russisch).

[3] Kuipers, L. and Niederreiter, H.: *Uniform Distribution of Sequences,* New York: John Wiley & Sons, Inc. 1974.

[4] Niederreiter, H., Discrepancy and Convex Programming, *Annali di Mat. (Ser. IV)* **93** (1972), 89—97.

[5] Zinterhof, P., Einige zahlentheoretische Methoden zur numerischen Quadratur und Interpolation, *S. B. Akad. Wiss., math.-naturw. Klasse, Abt. II,* **117** (1969), 51—77.

[6] Zinterhof, P., Über einige Abschätzungen bei der Approximation von Funktionen mit Gleichverteilungsmethoden, *S. B. Akad. Wiss., math.-naturw. Klasse, Abt. II,* **185** (1976), 121—132.

[7] Zygmund, A., *Trigonometric Series, Vol. 1 and 2,* Cambridge: University Press, 1968.

Mathematisches Institut der Universität,
Petersbrunnstraße 19, A—5020 Salzburg

(Eingegangen am 28. November 1977)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 291—317.

PERTURBATION OF LINEAR DIFFERENTIAL EQUATIONS BY A HALF-LINEAR TERM DEPENDING ON A SMALL PARAMETER

by

I. BIHARI

I. The case of a first order autonomous system of differential equations

1. The system in question is as follows

$$
\begin{aligned} u' &= v+\varepsilon f(u,v) \\ v' &= -u+\varepsilon g(u,v) \end{aligned} \qquad (|\varepsilon| \ll 1) \tag{1}
$$

where we assume

1° $f(\lambda u, \lambda v)=\lambda f(u,v)$, $g(\lambda u, \lambda v)=\lambda g(u,v)$, $\forall \lambda, u, v$;
2° f and g are analytic *on* the circle $u^2+v^2=1$;

We wish to determine the first, second, ... approximations of the general solution of (1) which are accurate in order $\varepsilon, \varepsilon^2, \dots$ for an arbitrary value of εt. This is not the case with expansions in power series of the form $u=\sum_{n=0}^{\infty} \varepsilon^n u_n(t)$, $v=\sum_{n=0}^{\infty} \varepsilon^n v_n(t)$, where secular terms appear (see [1], p. 35).

2. First of all let it be remarked that system (1) can be integrated in closed form. Namely, in the case of uniqueness both of u an v cannot vanish simultaneously. Say, u does not vanish and is monotonic in the neighbourhood of a point, then there we have

$$\frac{dv}{du} = H\left(\frac{v}{u}\right), \quad H(z) = \frac{-1+\varepsilon g(1,z)}{z+\varepsilon f(1,z)}$$

which by the substitution $\frac{v}{u}=z$ can be integrated as follows:

$$\int \frac{du}{u} = \int \frac{dz}{H(z)-z} \qquad (H(z) \neq z). \tag{*}$$

On the other hand for the exponential solutions — if any — of the form $u=Ae^{\lambda t}$, $v=Be^{\lambda t}$ we have

$$
\begin{aligned} \lambda A &= B+\varepsilon f(A,B), \\ \lambda B &= -A+\varepsilon g(A,B) \end{aligned}
$$

whence $\frac{B}{A}=H\left(\frac{B}{A}\right)$. Thus if the zeros of the equation $H(z)=z$ — if they exist

AMS (MOS) subject classifications (1970). Primary 34C15; Secondary 34C25, 34C35, 34D10, 34E05, 34E10.

Key words and phrases. Differential equations, perturbation, small parameter, asymptotic methods, nonlinear oscillations, periodic solution, dynamical systems.

at all — are $z_1, z_2, \ldots$, then the exponential solutions correspond to the points z_i and the solutions of the form (*) to the intervals $z_i<z<z_{i+1}$, $i=1, 2, \ldots$.[1] — Another approach for solving (1) is to apply the substitution $u=r\cos\alpha$, $v=r\sin\alpha$ giving the system

$$\frac{r'}{r}=\varepsilon\,[\cos\alpha f(\cos\alpha,\,\sin\alpha)+\sin\alpha g(\cos\alpha,\,\sin\alpha)],$$

$$\alpha'=-1+\varepsilon[-\sin\alpha f(\cos\alpha,\;\sin\alpha)+\cos\alpha g(\cos\alpha,\;\sin\alpha)],$$

where the second equation is independent from the first one and can be integrated immediately and we obtain easily the following statements concerning the orbits on the phase plane (u, v).

1) If ε is small enough, then $\alpha'<0$, $\alpha\searrow$, $\lim\limits_{t=\infty}\alpha=-\infty$.

2) If $uf(u,v)\geqq 0$, $vg(u,v)\geqq 0$, then $r'>0$, $r\nearrow$, $\lim\limits_{t=\infty} r=\infty$.

3) If the conditions 1) and 2) hold simultaneously then the origin O is an unstable spiral point with respect to (1) (Fig. 1).

4) If the assumptions of 2) hold but that of 1) do not, then it can happen that $\alpha'=0$, but $r\nearrow$ by all means.

EXAMPLE. Letting $f=\dfrac{c}{4}\dfrac{u^3}{u^2+v^2}$, $g=\dfrac{c}{4}\dfrac{v^3}{u^2+v^2}$, $\varepsilon=1$, then

$$\alpha'=-1-\frac{c}{4}\sin 4\alpha$$

and $\alpha'=0$ provided, $\sin 4\alpha=-\dfrac{4}{c}$.

Now if $C=4$, then $\alpha'=0$ at $\alpha_k=-\dfrac{\pi}{8}+\dfrac{k\pi}{2}$, $(k=0, 1, 2, 3)$, but does not change its sign at these points and $\alpha\searrow$, $\lim\limits_{t=\infty}\alpha=-\infty$. The orbits corresponding to these α_k's are straight lines (Fig. 2). — If $C>4$, then α' changes sign at its zeros and α is increasing and decreasing alternately. (Fig. 3 corresponds to $C=8$; then $\alpha'=0$ when $\alpha_k=-\dfrac{\pi}{24}+\dfrac{k\pi}{2}$ or $\alpha_l=\dfrac{\pi}{4}+\dfrac{\pi}{24}+\dfrac{l\pi}{2}$ $(k, l=0, 1, 2, 3)$.)

5) If instead of the rule of sign given in 2) we assume that $vf(u,v)\geqq 0$, $ug(u,v)\geqq 0$, $\varepsilon=1$, then both of r' and α' can vanish. E.g., if

$$f=\frac{v^3}{u^2+v^2},\quad g=\frac{u^3}{u^2+v^2},\quad \text{then}\quad \frac{r'}{r}=\frac{1}{2}\sin 2\alpha.$$

At the points $\alpha_k=\dfrac{k\pi}{2}$ $(k=0, 1, 2, 3)$ r has minimum and

$$\alpha'=2(\cos^4\alpha+\cos^2\alpha-1),\quad \alpha'=0\quad\text{if}\quad \cos\alpha_0=\frac{\sqrt{5}-1}{2}$$

and α' changes sign at $\alpha=\pm\alpha_0$ (Fig. 4).

[1] See below much more later the example where $u^2+v^2=a_0^2e^{\frac{3}{2}\varepsilon t}(1+O(\varepsilon))$ in first approximation.

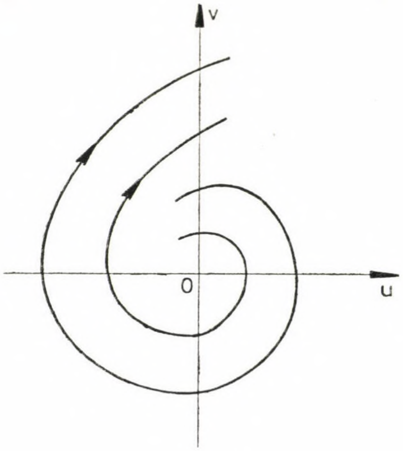


Fig. 1

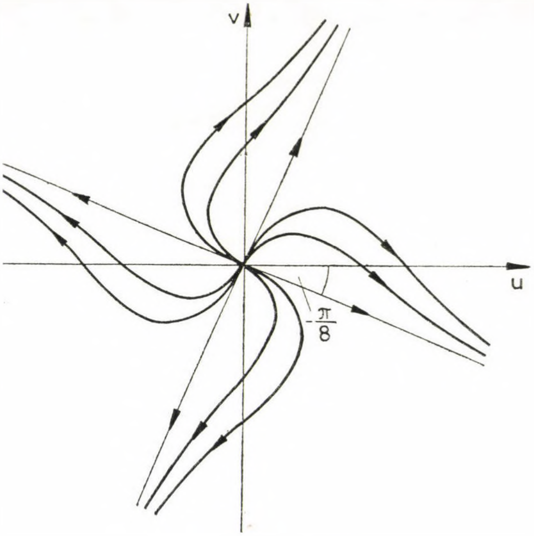


Fig. 2

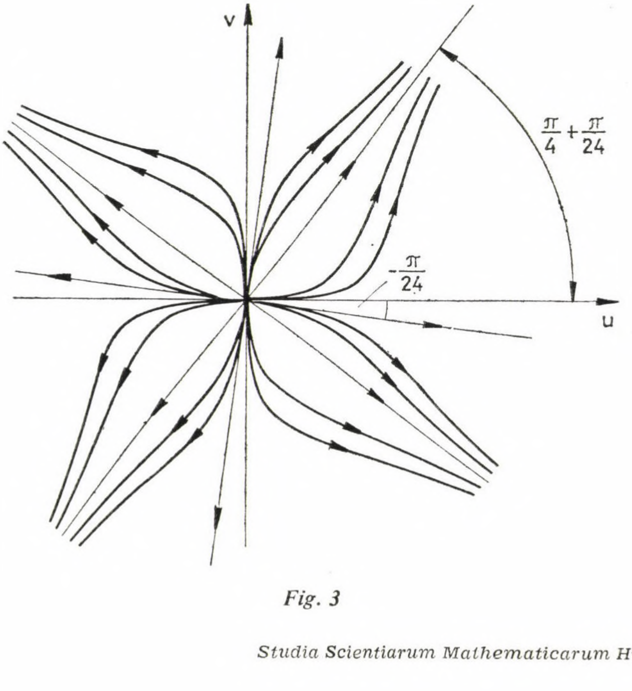


Fig. 3

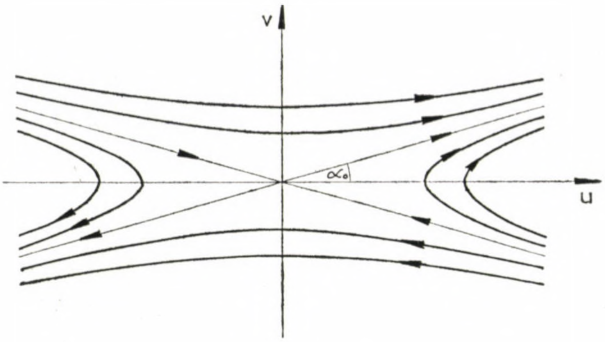


Fig. 4

6) If f and g are such as in 5), but $\varepsilon>0$ is small enough, then r' can vanish, but α' not. Then O is a stable or unstable spiral point or it can happen that the orbits are all closed, the solution is periodic. In fact, the necessary and sufficient condition of the periodicity of *every* solution is as follows

$$\text{(C)}\qquad 0=[\log r]_0^T=\varepsilon\int_0^T[\cos\alpha f_\alpha+\sin\alpha g_\alpha]\,dt=\varepsilon\int_0^{2\pi}\frac{\cos\alpha f_\alpha+\sin\alpha g_\alpha}{-1+\varepsilon[-\sin\alpha f_\alpha+\cos\alpha g_\alpha]}\,d\alpha,$$

$$[f_\alpha=f(\cos\alpha,\sin\alpha),\ g_\alpha=g(\cos\alpha,\sin\alpha)].$$

Here T means the time of one revolution and dt has been replaced by $\frac{d\alpha}{\alpha'}$. E.g. this relation is satisfied provided f is odd in its second argument and g is even in its second one, since then the integrand has opposite signs at $\alpha=\pi\pm\beta$. In the example of 5) $\frac{r'}{r}=\frac{\varepsilon}{2}\sin 2\alpha,\ \alpha'=-1+\varepsilon\cos 2\alpha$ and condition (C) is $\int_0^{2\pi}\frac{\sin 2\alpha}{1-\varepsilon\cos 2\alpha}\,d\alpha=0$ ($|\varepsilon|\ll 1$) which is satisfied and $r'=0$ at $\alpha=\frac{k\pi}{2}$ ($k=0,1,\dots$) where r has maxima and minima alternately (Fig. 5). — If $f=\frac{v^3}{u^2+v^2}$, $g=-\frac{u^3}{u^2+v^2}$ then (C) reads

$$\int_0^{2\pi}\frac{\sin 4\alpha}{1+\varepsilon(1-\frac{1}{2}\sin^2 2\alpha)}\,d\alpha=0,$$

which is also fulfilled and $r'=-\frac{\varepsilon}{4}\sin 4\alpha=0$ if $\alpha=\frac{k\pi}{4}$ ($k=0,1,2,\dots$) where r has successive maxima and minima (Fig. 6).

3. Now let us return to our original task, namely to obtain asymptotic solution of (1). The origin is a critical point with respect to (1) and we are faced with the so-called critical case of a critical point, since the eigenvalues of the matrix on the right-hand side of the corresponding homogeneous system

$$\text{(2)}\qquad x'=y,\quad y'=-x$$

are pure imaginary. Therefore it can have some interest how the method followed by KRYLOV—BOGOLIUBOV—MITROPOLSKI [1] works in this case. — The solution of (2) is $x=a\cos(t+t_0)$, $y=-a\sin(t+t_0)$, $(a, t_0=\text{const})$ and the orbits are circles.

The method for *general* nonlinear system is to assume the solution of (1) in the form

$$\begin{aligned} u &= a\cos\psi+\varepsilon u_1(a,\psi)+\varepsilon^2 u_2(a,\psi)+\dots, \\ &\qquad\qquad (3) \\ v &= -a\sin\psi+\varepsilon v_1(a,\psi)+\varepsilon^2 v_2(a,\psi)+\dots, \end{aligned}$$

where u_i and v_i have the period 2π in ψ and have continuous partial derivatives. Here a and ψ' are not constant anymore, but they have to satisfy the system

$$\begin{aligned} a' &= \varepsilon A_1(a)+\varepsilon^2 A_2(a)+\dots, \\ &\qquad\qquad (4) \\ \psi' &= 1+\varepsilon B_1(a)+\varepsilon^2 B_2(a)+\dots \end{aligned}$$

of differential equations and A_i, B_i, u_i, v_i are to be determined that (3) be solution of (1). Applying this method no secular terms appear. In the present case making intensive use of the half-linearity of f and g our

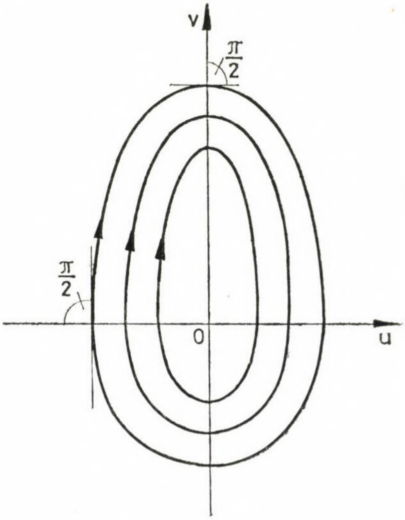


Fig. 5

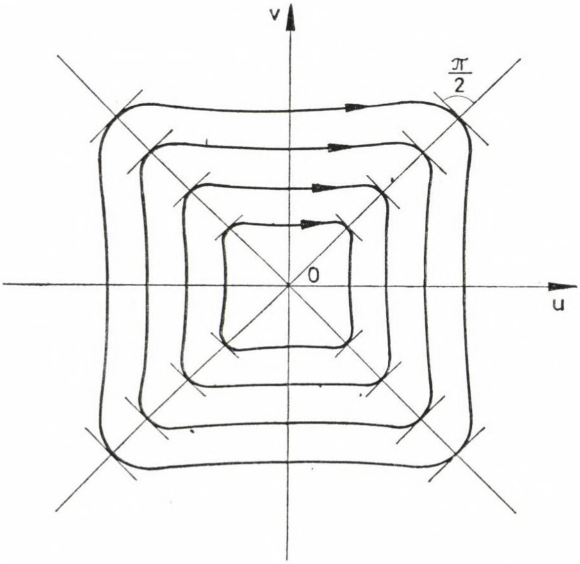


Fig. 6

purpose will be reached much more easily than having any different nonlinear f and g. The simplifications obtained here are:

(i) It is sufficient to assume f and g to be analytic on the unit-circle (instead of the whole plane);[2]

(ii) The functions u_i, v_i, A_i $(i=1, 2, \dots)$ are linear in a, and B_i are constant i.e., instead of (3)—(4) we have now

$$(5)\qquad \begin{aligned} u &= a[\cos\psi+\varepsilon u_1(\psi)+\varepsilon^2 u_2(\psi)+\dots],\\ v &= a[-\sin\psi+\varepsilon v_1(\psi)+\varepsilon^2 v_2(\psi)+\dots],\end{aligned}$$

$$(6)\qquad \begin{aligned} a' &= a(\varepsilon A_1+\varepsilon^2 A_2+\dots)\\ \psi' &= 1+\varepsilon B_1+\varepsilon^2 B_2+\dots\end{aligned}$$

where A_i, B_i $(i=1, 2, \dots)$ are constant;

(iii) As we shall see later, the Fourier expansions of such functions as $f(\cos\psi, -\sin\psi)$ etc. will involve terms only of the forms $\cos(2n+1)\psi$, $\sin(2n+1)\psi$ which simplifies strongly the computations and formulae.

Let us see the details. Now with the notations $\frac{du_i}{d\psi}=\dot u_i, \frac{dv_i}{d\psi}=\dot v_i$ corresponding to (5)—(6) we have

$$\begin{aligned} u' &= a'(\cos\psi+\varepsilon u_1+\varepsilon^2 u_2+\dots)+a\psi'(-\sin\psi+\varepsilon\dot u_1+\varepsilon^2\dot u_2+\dots) =\\ &= a[(\varepsilon A_1+\varepsilon^2 A_2+\dots)(\cos\psi+\varepsilon u_1+\varepsilon^2 u_2+\dots)+\\ &\quad +(1+\varepsilon B_1+\varepsilon^2 B_2+\dots)(-\sin\psi+\varepsilon\dot u_1+\varepsilon^2\dot u_2+\dots)] =\\ &= a[-\sin\psi+\varepsilon(A_1\cos\psi-B_1\sin\psi+\dot u_1)+\\ &\quad +\varepsilon^2(A_2\cos\psi-B_2\sin\psi+A_1u_1+B_1\dot u_1+\dot u_2)+\varepsilon^3\dots],\\ v' &= a'(-\sin\psi+\varepsilon v_1+\varepsilon^2 v_2+\dots)+a\psi'(-\cos\psi+\varepsilon\dot v_1+\varepsilon^2\dot v_2+\dots) =\\ &= a[(\varepsilon A_1+\varepsilon^2 A_2+\dots)(-\sin\psi+\varepsilon v_1+\varepsilon^2 v_2+\dots)+\\ &\quad +(1+\varepsilon B_1+\varepsilon^2 B_2+\dots)(-\cos\psi+\varepsilon\dot v_1+\varepsilon^2\dot v_2+\dots)] =\\ &= a[-\cos\psi+\varepsilon(-A_1\sin\psi-B_1\cos\psi+\dot v_1)+\\ &\quad +\varepsilon^2(-A_2\sin\psi-B_2\cos\psi+A_1v_1+B_1\dot v_1+\dot v_2)+\varepsilon^3\dots].\end{aligned}$$

Furthermore $f(u, v)=af(\cos\psi+\varepsilon u_1+\varepsilon^2 u_2+\dots, -\sin\psi+\varepsilon v_1+\varepsilon^2 v_2+\dots)$. Let us expand the function $f(\cos\psi+\varepsilon\dots, -\sin\psi+\varepsilon\dots)$ in Taylor series about the point $\Omega=(\cos\psi, -\sin\psi)$ situated on the unit circle

$$\begin{aligned} \varepsilon f(u, v) &= a[\varepsilon f(\Omega)+\varepsilon f_u(\Omega)(\varepsilon u_1+\varepsilon^2 u_2+\dots)+\varepsilon f_v(\Omega)(\varepsilon v_1+\varepsilon^2 v_2+\dots)+\dots] =\\ &= a\{\varepsilon f(\Omega)+\varepsilon^2[u_1 f_u(\Omega)+v_1 f_v(\Omega)]+\varepsilon^3\dots\}.\end{aligned}$$

In the same way

$$\varepsilon g(u, v) = a\{\varepsilon g(\Omega)+\varepsilon^2[u_1 g_u(\Omega)+v_1 g_v(\Omega)]+\varepsilon^3+\dots\}.$$

[2] However, this involves the analyticity on the whole plane except the origin.

Putting all these in (1) and comparing the coefficients of ε, ε^2, on both sides of the equation we obtain for the determination of u_i, v_i, A_i, B_i $(i=1, 2)$ the following two systems of equations:

$$\text{(7)}\qquad \begin{aligned} A_1 \cos\psi - B_1 \sin\psi + \dot{u}_1 &= v_1 + f(\Omega),\\ -A_1 \sin\psi - B_1 \cos\psi + \dot{v}_1 &= -u_1 + g(\Omega), \end{aligned}$$

$$\text{(8)}\qquad \begin{aligned} A_2 \cos\psi - B_2 \sin\psi + \dot{u}_2 - v_2 &= -A_1 u_1 - B_1 \dot{u}_1 + u_1 f_u(\Omega) + v_1 f_v(\Omega),\\ -A_2 \sin\psi - B_2 \cos\psi + \dot{v}_2 + u_2 &= -A_1 v_1 - B_1 \dot{v}_1 + u_1 g_u(\Omega) + v_1 g_v(\Omega). \end{aligned}$$

These equations serve to determine the first and second approximations, respectively. Now by the half-linearity $f(\cos\psi, -\sin\psi) = \cos\psi\, f(1, -\operatorname{tg}\psi)$. The function $f(1, -\operatorname{tg}\psi)$ has the period π and if it is smooth enough, it can be expanded in a Fourier series of the form

$$f(1, -\operatorname{tg}\psi) = \sum_{n=0}^{\infty} (f_n \cos 2n\psi + F_n \sin 2n\psi) \quad (F_0 = 0),$$

involving

$$\cos\psi f(1, -\operatorname{tg}\psi) = \sum_{n=0}^{\infty} (\bar{f}_n \cos(2n+1)\psi + \bar{F}_n \sin(2n+1)\psi),$$

where $\bar{f}_0 = f_0 + \frac{1}{2} f_1$, $\bar{F}_0 = \frac{1}{2} F_1$, $\bar{f}_n = \frac{1}{2}(f_n + f_{n+1})$, $\bar{F}_n = \frac{1}{2}(F_n + F_{n+1})$, $n = 1, 2, \ldots$. In the same way

$$\cos\psi g(1, -\operatorname{tg}\psi) = \sum_{n=0}^{\infty} (\bar{g}_n \cos(2n+1)\psi + \bar{G}_n \sin(2n+1)\psi),$$

where $\bar{g}_0 = g_0 + \frac{1}{2} g_1$, $\bar{G}_0 = \frac{1}{2} G_1$, $\bar{g}_n = \frac{1}{2}(g_n + g_{n+1})$, $\bar{G}_n = \frac{1}{2}(G_n + G_{n+1})$ $(n = 1, 2, \ldots)$. Here g_n, G_n are the Fourier coefficients of $g(1, -\operatorname{tg}\psi)$. Let the Fourier series of u_1 and v_1 be

$$u_1 = \sum_{n=0}^{\infty} (\alpha_n \cos n\psi + \beta_n \sin n\psi), \quad v_1 = \sum_{n=0}^{\infty} (\gamma_n \cos n\psi + \delta_n \sin n\psi), \quad (\beta_0 = \delta_0 = 0),$$

respectively. Then

$$\dot{u}_1 = \sum_{n=0}^{\infty} n(-\alpha_n \sin n\psi + \beta_n \cos n\psi), \quad \dot{v}_1 = \sum_{n=0}^{\infty} n(-\gamma_n \sin n\psi + \delta_n \cos n\psi).$$

Putting all these in (7) and comparing on both sides the coefficients of $\cos n\psi$ and $\sin n\psi$ we have for $n=0$ $\alpha_0 = \gamma_0 = 0$. For $n=1$ we get the system

$$\begin{aligned} A_1 + \beta_1 &= \gamma_1 + \bar{f}_0, & B_1 + \alpha_1 &= -\delta_1 - \bar{F}_0,\\ -B_1 + \delta_1 &= -\alpha_1 + \bar{g}_0, & A_1 + \gamma_1 &= \beta_1 - \bar{G}_0 \end{aligned}$$

of equations, whence

$$A_1 = \frac{1}{2}(\bar f_0 - \bar G_0), \quad \beta_1 - \gamma_1 = \frac{1}{2}(\bar f_0 + \bar G_0), \tag{9}$$
$$B_1 = -\frac{1}{2}(\bar F_0 + \bar g_0), \quad \delta_1 + \alpha_1 = \frac{1}{2}(\bar g_0 - \bar F_0).$$

Thus A_1, B_1 are determined while among $\alpha_1, \beta_1, \gamma_1, \delta_1$ we received two relations, i.e., two of them remains undetermined for the moment and can be determined only by some supplementary conditions as we shall see.

For $n=2k$, $(k=1, 2, \ldots)$

$$2k\beta_{2k} = \gamma_{2k}, \qquad 2k\alpha_{2k} = -\delta_{2k},$$
$$2k\delta_{2k} = -\alpha_{2k}, \quad 2k\gamma_{2k} = \beta_{2k},$$

whence $\alpha_{2k}=\beta_{2k}=\gamma_{2k}=\delta_{2k}=0$, $(k=1, 2, \ldots)$.

This great simplification is the consequence of (iii). For $n=2k+1$ $(k=1, 2, \ldots)$

$$(2k+1)\beta_{2k+1} = \gamma_{2k+1} + \bar f_k, \quad -(2k+1)\alpha_{2k+1} = \delta_{2k+1} + \bar F_k,$$
$$(2k+1)\delta_{2k+1} = -\alpha_{2k+1} + \bar g_k, \quad (2k+1)\gamma_{2k+1} = \beta_{2k+1} + \bar G_k,$$

whence

$$4k(k+1)\alpha_{2k+1} = -[(2k+1)\bar F - \bar g_k], \quad 4k(k+1)\gamma_{2k+1} = \bar f_k - (2k+1)\bar G_k, \tag{9'}$$
$$4k(k+1)\beta_{2k+1} = (2k+1)\bar f_k - \bar G_k, \quad 4k(k+1)\delta_{2k+1} = (2k+1)\bar g_k + \bar F_k.$$[3]

Therefore in first approximation

$$a' = a\varepsilon\lambda, \quad \lambda = \frac{1}{2}(\bar f_0 - \bar G_0),$$

$$\psi' = 1 + \varepsilon\mu, \quad \mu = -\frac{1}{2}(\bar F_0 + \bar g_0)$$

and so

$$a = a_0 e^{\varepsilon\lambda t}, \quad \psi = (\psi_0 + t) + \mu\varepsilon t$$

where a_0 and ψ_0 are arbitrary constants. Furthermore

$$u_1 = \sum_{n=0}^{\infty}\left(\alpha_{2n+1}\cos(2n+1)\psi + \beta_{2n+1}\sin(2n+1)\psi\right),$$

$$v_1 = \sum_{n=0}^{\infty}\left(\gamma_{2n+1}\cos(2n+1)\psi + \delta_{2n+1}\sin(2n+1)\psi\right).$$

Here are two relations among $\alpha_1, \beta_1, \gamma_1, \delta_1$, the rest of the coefficients are determined.

[3] Compared to $\bar f_k, \bar F_k, \bar g_k, \bar G_k$ the order of magnitudes of $\alpha_{2k+1}, \ldots, \delta_{2k+1}$ are $\frac{1}{k}$. Therefore the series of u_1 and v_1 are convergent.

EXAMPLE. Letting $f(u, v)=\frac{u^3}{u^2+v^2}$, $g(u, v)=\frac{v^3}{u^2+v^2}$ we have

$$f(\cos\psi, -\sin\psi) = \cos^3\psi = \frac{3}{4}\cos\psi+\frac{1}{4}\cos 3\psi,$$

$$g(\cos\psi, -\sin\psi) = -\sin^3\psi = -\frac{3}{4}\sin\psi+\frac{1}{4}\sin 3\psi,$$

consequently $\bar{f}_0=\frac{3}{4}$, $\bar{F}_0=0$, $\bar{f}_1=\frac{1}{4}$, $\bar{F}_1=0$, $\bar{G}_0=-\frac{3}{4}$, $\bar{g}_1=0$, $\bar{g}_0=0$, $\bar{G}_1=\frac{1}{4}$ and the remaining of the coefficients are zero. Therefore

$$\lambda=\frac{3}{4},\quad \mu=0,\quad \beta_1=\gamma_1,\quad \delta_1=-\alpha_1,$$

$$\alpha_3=\delta_3=0,\quad \beta_3=-\gamma_3=\frac{1}{16}$$

and the rest of α_i and β_i are zero. So we have

$$u_1=\alpha_1\cos\psi+\beta_1\sin\psi+\frac{1}{16}\sin 3\psi,$$

$$v_1=\beta_1\cos\psi-\alpha_1\sin\psi-\frac{1}{16}\cos 3\psi$$

and the first approximation is

$$u=a(\cos\psi+\varepsilon u_1)=a_0e^{\frac{3}{4}\varepsilon t}\left[\cos\psi+\varepsilon\left(\alpha_1\cos\psi+\beta_1\sin\psi+\frac{1}{16}\sin 3\psi\right)\right],$$

$$v=a(-\sin\psi+\varepsilon v_1)=a_0e^{\frac{3}{4}\varepsilon t}\left[-\sin\psi+\varepsilon\left(\beta_1\cos\psi-\alpha_1\sin\psi-\frac{1}{16}\cos 3\psi\right)\right].$$

The $a=a_0e^{\frac{3}{4}\varepsilon t}$ will be the complete amplitude of the basic harmonic $\cos\psi$ and $-\sin\psi$ if and only if $\alpha_1=0$. If in this case $\beta_1\neq 0$, then the basic harmonic will be subjected to a phase-shift, namely

$$\begin{aligned}\cos\psi+\varepsilon\beta_1\sin\psi &= C\cos(\psi-\bar{\psi}),\\ -\sin\psi+\varepsilon\beta_1\cos\psi &= C\sin(\psi-\bar{\psi}),\end{aligned}\qquad C^2=1+\varepsilon^2\beta_1^2,\ \mathrm{tg}\,\bar{\psi}=\varepsilon\beta_1.$$

If $\beta_1=0$, then there is no phase-shift. Then

$$u=a_0e^{\frac{3}{4}\varepsilon t}\left(\cos\psi+\frac{\varepsilon}{16}\sin 3\psi\right),$$

$$v=a_0e^{\frac{3}{4}\varepsilon t}\left(-\sin\psi-\frac{\varepsilon}{16}\cos 3\psi\right)$$

and $u^2+v^2=a_0^2e^{-\frac{3}{2}\varepsilon t}\left(1+\frac{\varepsilon}{8}\sin 4\psi+\frac{\varepsilon^2}{256}\right)\approx a_0^2e^{-\frac{3}{2}\varepsilon t}$. This is in very good agreement with the exact value. Therefore the orbits are for small εt near to a circle. Summarizing: in first approximation there is always a weak third upper harmonic, too, and the basic harmonic bears a phase-shift in general.

Let it be *remarked* that we have from the relation $f(u,v)=uf\left(1,\frac{v}{u}\right)$

$$f_u(u,v)=f\left(1,\frac{v}{u}\right)+uf_v\left(1,\frac{v}{u}\right)\left(-\frac{v}{u^2}\right)=f\left(1,\frac{v}{u}\right)-\frac{v}{u}f_v\left(1,\frac{v}{u}\right),$$

$$f_v(u,v)=uf_v\left(1,\frac{v}{u}\right)\frac{1}{u}=f_v\left(1,\frac{v}{u}\right),$$

which shows that $f_u(\cos\psi,-\sin\psi)$ and $f_v(\cos\psi,-\sin\psi)$ are the functions of $\operatorname{tg}\psi$ only and they are periodic with period π. Since u_1, v_1 involve only terms $\cos(2k+1)\psi$, $\sin(2k+1)\psi$ and $f_u(\cos\psi,-\sin\psi)$, $f_v(\dots)$, $g_u(\dots)$, $g_v(\dots)$ only terms $\cos 2l\psi$, $\sin 2l\psi$, the functions

$$\mathcal{F}=u_1f_u(\dots)+v_1f_v(\dots)\quad\text{and}\quad\mathcal{G}=u_1g_u(\dots)+v_1g_v(\dots)$$

will contain only terms of the form $\cos(2k+1)\psi$, $\sin(2k+1)\psi$. Let the Fourier series of $\mathcal{F}$, $\mathcal{G}$

$$\mathcal{F}=\sum_{n=0}^{\infty}(p_n\cos(2n+1)\psi+P_n\sin(2n+1)\psi),$$

$$\mathcal{G}=\sum_{n=0}^{\infty}(r_n\cos(2n+1)\psi+R_n\sin(2n+1)\psi)$$

and let

$$u_2=\sum_{n=0}^{\infty}(h_n\cos n\psi+H_n\sin n\psi),\quad v_2=\sum_{n=0}^{\infty}(k_n\cos n\psi+K_n\sin n\psi),\quad H_0=K_0=0.$$

By putting all these in (8) and comparing the coefficients of $\cos n\psi$, $\sin n\psi$ we have for $n=0$, $h_0=k_0=0$, for $n=1$

$$A_2+H_1+A_1\alpha_1+B_1\beta_1=k_1+p_0,\quad -B_2-h_1+A_1\beta_1-B_1\alpha_1=K_1+P_0,$$
$$-B_2+K_1+A_1\gamma_1+B_1\delta_1=-h_1+r_0,\quad -A_2-k_1+A_1\delta_1-B_1\gamma_1=-H_1+R_0,$$

whence

$$\begin{aligned}
A_2&=\frac{1}{2}[p_0-R_0-A_1(\alpha_1-\delta_1)-B_1(\beta_1+\gamma_1)],\\
H_1-k_1&=\frac{1}{2}[p_0+R_0-A_1(\alpha_1+\delta_1)-B_1(\beta_1-\gamma_1)],\\
B_2&=-\frac{1}{2}[P_0+r_0-A_1(\beta_1+\gamma_1)-B_1(\delta_1-\alpha_1)],\\
h_1+K_1&=\frac{1}{2}[r_0-P_0-A_1-(\gamma_1-\beta_1)-B_1(\delta_1+\alpha_1)],
\end{aligned}\tag{10}$$

i.e., h_1, H_1, k_1, K_1 are not known, but they are connected by two relations.

For $n=2l$ $(l=1, 2, \dots)$

$$2lH_{2l} = k_{2l}, \quad 2lK_{2l} = -h_{2l}, \quad 2lh_{2l} = -K_{2l}, \quad 2lk_{2l} = -H_{2l},$$

whence $h_{2l}=H_{2l}=k_{2l}=K_{2l}=0$.

For $n=2l+1$ $(l=1, 2, \dots)$

$$\begin{aligned} nH_n+A_1\alpha_n+B_1n\beta_n &= k_n+p_l, \\ nK_n+A_1\gamma_n+B_1n\delta_n &= -h_n+r_l, \\ -nh_n+A_1\beta_n-B_1n\alpha_n &= K_n+P_l, \\ -nk_n+A_1\delta_n-B_1n\gamma_n &= -H_n+R_l, \end{aligned}$$

whence

$$\begin{aligned} 4l(l+1)h_n &= A_1(\gamma_n+n\beta_n)+nB_1(\delta_n-n\alpha_n)-r_l-nP_l, \\ 4l(l+1)H_n &= -A_1(n\alpha_n-\delta_n)-nB_1(n\beta_n+\gamma_n)+np_l-R_l, \\ 4l(l+1)k_n &= -A_1(\alpha_n-n\delta_n)-nB_1(\beta_n+n\gamma_n)+p_l-nR_l, \\ 4l(l+1)K_n &= -A_1(n\gamma_n+\beta_n)-nB_1(n\delta_n-\alpha_n)+nr_l+P_l.^4 \end{aligned} \tag{10'}$$

In the above example by an elementary computation

$$A_2 = 0, \quad B_2 = -\frac{1}{32}, \quad k_1 = H_1, \quad K_1 = -h_1,$$

$$h_3 = \frac{1}{512}(45-96\beta_1), \quad H_3 = -k_3 = \frac{\alpha_1}{16}, \quad K_3 = -\frac{1}{512}(135+96\beta_1),$$

$$h_5 = H_5 = k_5 = K_5 = 0,$$

$$h_7 = K_7 = \frac{1}{512}, \quad H_7 = k_7 = 0,$$

while the rest of the coefficients of u_2 and v_2 are zero. Thus in *second approximation* a remains invariant, while from $\psi'=1+\varepsilon B_1+\varepsilon^2B_2=1-\frac{\varepsilon^2}{32}$ we have $\psi=(\psi_0+t)-\frac{\varepsilon^2}{32}t$ and

$$u_2 = h_1\cos\psi+H_1\sin\psi+\frac{1}{512}(45-96\beta_1)\cos 3\psi+\frac{\alpha_1}{16}\sin 3\psi+\frac{1}{512}\cos 7\psi,$$

$$v_2 = H_1\cos\psi-h_1\sin\psi-\frac{\alpha_1}{16}\cos 3\psi-\frac{1}{512}(135+96\beta_1)\sin 3\psi+\frac{1}{512}\sin 7\psi.$$

[4] A simple consideration like in footnote [3] gives the convergence of the series of u_2 and v_2.

Again: the a remains the complete amplitude of the basic harmonic only in the case where $\alpha_1=h_1=0$ and arises no phase shift, too, provided $\beta_1=H_1=0$. Then

$$u_2=\frac{1}{512}(45\cos 3\psi+\cos 7\psi),\quad v_2=\frac{1}{512}(-135\sin 3\psi+\sin 7\psi).$$

Summarizing: in second approximation some third and seventh upper harmonic appear with amplitudes step by step decreasing; namely

$$u=a\left[\cos\psi+\frac{\varepsilon}{16}\sin 3\psi+\frac{\varepsilon^2}{512}(45\cos 3\psi+\cos 7\psi)\right],$$

$$v=a\left[-\sin\psi-\frac{\varepsilon}{16}\cos 3\psi+\frac{\varepsilon^2}{512}(-135\sin 3\psi+\sin 7\psi)\right].$$

REMARK. Our formulae will be simplified in great measure when we assume in (10) $\alpha_1-\delta_1=\beta_1+\gamma_1=0$. Then by (9) $\alpha_1, \beta_1, \gamma_1, \delta_1$ can be uniquely determined and from (10) then $A_2=\frac{1}{2}(p_0-R_0)$, $B_2=-\frac{1}{2}(p_0-R_0)$. Furthermore H_1-k_1 and h_1+K_1 will have known values. — This is an alternative method of the above one which helped us to make the values of $\alpha_1, \beta_1, \gamma_1, \delta_1$ definite.

II. The case of the second order non-autonomous differential equation

A. *The case of non-resonance*

Let us consider the equation

$$\ddot{x}+\omega^2 x=\varepsilon f(\nu t, x, \dot{x})\quad (|\varepsilon|\ll 1) \tag{1}$$

where $f(\theta, u, v)$ is half-linear in u, v, i.e.,

$$f(\theta, \lambda u, \lambda v)=\lambda f(\theta, u, v),\quad \forall\,\theta, u, v, \lambda$$

and analytic *on* the ellipse $u^2+\frac{v^2}{\omega^2}=1$, finally it is periodic in θ with period 2π[5]. We look for the general solution of (1) for arbitrary εt with an accuracy $\varepsilon, \varepsilon^2, \ldots$. In this Case A we assume that $\frac{\nu}{\omega}$ is an irrational constant, moreover it cannot be "well" approximated with rational numbers. This problem and that of the accuracy and convergence will be treated in the last section. Now we wish to deal with the formal determination of the first, second, ... approximations, only.

[5] If f does not depend explicitely on t and $\omega=1$, then — as it is easily seen — every solution is periodic provided (see Section 1),

$$\int_0^{2\pi}\frac{\sin\varphi f(\cos\varphi, \sin\varphi)}{-1+\varepsilon\cos\varphi f(\cos\varphi, \sin\varphi)}\,d\varphi=0,$$

which is satisfied, e.g., when f is even in its second argument.

We assume $f(\nu t, x, \dot{x})$ in the form

$$f(\nu t, x, \dot{x}) = \sum_{n=-\infty}^{\infty} f_n(x, \dot{x}) e^{in\nu t} \qquad f_n(x, \dot{x}) = \frac{1}{2\pi} \int_0^{2\pi} f(\theta, x, \dot{x}) e^{-in\theta}\, d\theta \tag{2}$$

or, for a while, in the form

$$f(\nu t, x, \dot{x}) = \sum_{n=-N}^{N} f_n(x, \dot{x}) e^{in\nu t} \qquad (N \in \mathbf{Z}). \tag{3}$$

The solutions of (1) for $\varepsilon=0$ is $x=a\cos(\omega t+t_0)$, $a=\text{const}$, $t_0=\text{const}$. Therefore we assume the solution in the form

$$x = a\,[\cos\psi + \varepsilon u_1(\nu t, \psi) + \varepsilon^2 u_2(\nu t, \psi) + \dots] \tag{4}$$

where $u_i(\theta, \psi)$ is periodic of period 2π in θ and ψ as well. The $\dot{a}$ and $\dot{\psi}$ will be assumed as

$$\dot{a} = a(\varepsilon A_1 + \varepsilon^2 A_2 + \dots), \quad \dot{\psi} = \omega + \varepsilon B_1 + \varepsilon^2 B_2 + \dots (= \text{const!}) \tag{5}$$

where A_i, B_i are constant. The form of (4)—(5) is much more simpler than that of the *general* assumption

$$x = a\cos\psi + \varepsilon u_1(a, \nu t, \psi) + \varepsilon^2 u_2(a, \nu t, \psi) + \dots$$

$$\dot{a} = \varepsilon A_1(a) + \varepsilon^2 A_2(a) + \dots, \quad \dot{\psi} = \omega + \varepsilon B_1(a) + \varepsilon^2 B_2(a) + \dots$$

applicable for *any* nonlinear perturbations. — Now by (4)—(5)

$$\dot{x} = \dot{a}(\cos\psi + \varepsilon u_1 + \dots) + a\dot{\psi}\left(-\sin\psi + \varepsilon\frac{\partial u_1}{\partial\psi} + \dots\right) + a\nu\left(\varepsilon\frac{\partial u_1}{\partial\theta} + \varepsilon^2\frac{\partial u_2}{\partial\theta} + \dots\right) =$$

$$= a\left[-\omega\sin\psi + \varepsilon\left(A_1\cos\psi - B_1\sin\psi + \omega\frac{\partial u_1}{\partial\psi} + \nu\frac{\partial u_1}{\partial\theta}\right) + \right.$$

$$\left. + \varepsilon^2\left(-B_2\sin\psi + A_2\cos\psi + \omega\frac{\partial u_2}{\partial\psi} + a\nu\frac{\partial u_2}{\partial\theta} + A_1 u_1 + B_1\frac{\partial u_1}{\partial\psi}\right) + \varepsilon^3\dots\right].$$

Furthermore

$$\ddot{x} = \ddot{a}(\cos\psi + \varepsilon u_1 + \dots) + 2\dot{a}\dot{\psi}\left(-\sin\psi + \varepsilon\frac{\partial u_1}{\partial\psi} + \dots\right) + a\dot{\psi}^2\left(-\cos\psi + \varepsilon\frac{\partial^2 u_1}{\partial\psi^2} + \dots\right) +$$

$$+ 2\dot{a}\nu\left(\varepsilon\frac{\partial u_1}{\partial\theta} + \varepsilon^2\frac{\partial u_2}{\partial\theta} + \dots\right) + a\nu^2\left(\varepsilon\frac{\partial^2 u_1}{\partial\theta^2} + \varepsilon^2\frac{\partial^2 u_2}{\partial\theta^2} + \dots\right) +$$

$$+ 2a\dot{\psi}\nu\left(\varepsilon\frac{\partial^2 u_1}{\partial\theta\,\partial\psi} + \varepsilon^2\frac{\partial^2 u_2}{\partial\theta\, d\psi} + \dots\right).$$

Here we have

$$\ddot{a} = \dot{a}(\varepsilon A_1 + \dots) = a(\varepsilon A_1 + \dots)^2 = a(\varepsilon^2 A_1^2 + 2A_1 A_2 \varepsilon^3 + \dots),$$

$$\dot{a}\dot{\psi} = a\,[\varepsilon\omega A_1 + \varepsilon^2(A_2\omega + A_1 B_1) + \varepsilon^3\dots],$$

$$\dot{\psi}^2 = \omega^2 + \varepsilon\cdot 2\omega B_1 + \varepsilon^2(B_1^2 + 2\omega B_2) + \varepsilon^3\dots\,.$$

Thus

$$\ddot{x} = a\{-\omega^2 \cos\psi + \varepsilon[-2\omega(A_1 \sin\psi + B_1 \cos\psi) + S_1(u_1)] +$$
$$+\varepsilon^2[-2\omega(A_2 \sin\psi + B_2 \cos\psi) + S_1(u_2) + (A_1^2 - B_1^2)\cos\psi -$$
$$-2A_1B_1 \sin\psi + 2\omega A_1 \frac{\partial u_1}{\partial\psi} + 2\omega B_1 \frac{\partial^2 u_1}{\partial\psi^2} + 2A_1 \nu \frac{\partial u_1}{\partial\theta} + 2B_1 \nu \frac{\partial^2 u_1}{\partial\theta\,\partial\psi}\Big] + \varepsilon^3 \ldots\Big\}.$$

$$S_1(u_i) = \omega^2 \frac{\partial^2 u_i}{\partial\psi^2} + 2\omega\nu \frac{\partial^2 u_i}{\partial\theta\, d\psi} + \nu^2 \frac{\partial^2 u_i}{\partial\theta^2} \quad (i = 1, 2, \ldots).$$

Putting the above values of x and $\dot{x}$ in $f(\nu t, x, \dot{x})$ and expanding the function thus obtained in Taylor series about the point $\Omega(\nu t, \cos\psi, -\omega\sin\psi)$ lying on the cylinder $u^2 + \frac{\nu^2}{\omega^2} = 1$ we have

$$\varepsilon f(\nu t, x, \dot{x}) = \varepsilon a f\big(\nu t, \cos\psi + \varepsilon u_1 + \varepsilon^2 u_2 + \ldots;\ -\omega\sin\psi + \varepsilon(\ldots) + \varepsilon^2(\ldots) + \ldots\big) =$$
$$= \varepsilon a f(\Omega) + \varepsilon a f_x(\Omega)(\varepsilon u_1 + \varepsilon^2 u_2 + \ldots) +$$
$$+\varepsilon a f_{\dot{x}}(\Omega)\left[\varepsilon\left(A_1 \cos\psi - B_1 \sin\psi + \omega \frac{\partial u_1}{\partial\psi} + \gamma \frac{\partial u_1}{\partial\theta}\right) + \varepsilon^2 \ldots\right] =$$
$$= \varepsilon a f(\Omega) + \varepsilon^2 a\left[f_x(\Omega) u_1 + f_{\dot{x}}(\Omega)\left(A_1 \cos\psi - B_1 \sin\psi + \omega \frac{\partial u_1}{\partial\psi} + \nu \frac{\partial u_1}{\partial\theta}\right)\right] + \varepsilon^3 \ldots.$$

Replace all these expressions in (1) and compare the coefficients of $\varepsilon, \varepsilon^2, \ldots$. So we obtain the equations of the consecutive approximations, serving for the determination of A_i, B_i, u_i $(i=1, 2, \ldots)$. The equation of the first approximation is

$$(6)\qquad \begin{gathered} -2\omega(A_1 \sin\psi + B_1 \cos\psi) + S_2(u_1) = f(\Omega), \\ S_2(u_i) = \omega^2 \frac{\partial^2 u_i}{\partial\psi^2} + 2\nu\omega \frac{\partial^2 u_i}{\partial\theta\,\partial\psi} + \nu^2 \frac{\partial^2 u_i}{\partial\theta^2} + \omega^2 u_i \quad (i = 1, 2, \ldots). \end{gathered}$$

Again by the formula $f(\nu t, \cos\psi, -\omega\sin\psi) = \cos\psi\, f(\nu t, 1, -\omega\,\mathrm{tg}\,\psi)$, having the second factor the period π, the product by $\cos\psi$ involves terms of the form $\cos(2m+1)\psi$, $\sin(2m+1)\psi$ only. Therefore

$$(7)\qquad \begin{gathered} f(\nu t, \cos\psi, -\omega\sin\psi) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f_{2m+1,n}\, e^{i((2m+1)\psi + n\nu t)} \\ f_{2m+1,n} = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} f(\theta, \cos\xi - \omega\sin\xi)\, e^{-i((2m+1)\xi + n\theta)}\, d\xi\, d\theta. \end{gathered}$$

For the sake of simplicity we write m instead of $2m+1$ keeping in mind that m is odd. The m is also odd in the function

$$u_1(\nu t, \psi) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} u_{1mn} e^{i(m\psi + n\nu t)}, \quad u_{1mn} = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} u_1(\theta, \xi)\, e^{-i(m\xi + n\theta)}\, d\xi\, d\theta$$

since

$$S_2(u_1) = \sum_{m,n=-\infty}^{\infty} (m, n) u_{1mn} e^{i(m\psi+nvt)} \quad \text{where} \quad (m, n) = \omega^2-(m\omega+nv)^2 \tag{9}$$

and thus by (6) $S_2(u_1)$ and u_1 itself can involve in m odd terms only. Making use of (7) and (9), (6) will have the form

$$-2\omega A_1 \sin\psi - 2\omega B_1 \cos\psi + \sum_{m,n} (m, n) u_{1mn} e^{i(m\psi+nvt)} = \sum_{m,n} f_{mn} e^{i(m\psi+nvt)}. \tag{10}$$

The case $(m, n)=0$ occurs only when $n=0, m=\pm 1$. Namely $(m, n)=0$ is equivalent to

$$m\omega+nv = \pm\omega \quad \text{or} \quad (m\pm 1)\omega+nv = 0 \tag{11}$$

and this can hold only in the above mentioned case. Otherwise $\frac{\omega}{v}$ would be rational which is excluded. Therefore with the expression $-2\omega A_1 \sin\psi - 2B_1\omega\cos\psi$ on the left of (10) the terms corresponding to $n=0, m=\pm 1$ of the right side must be equated, i.e.,

$$\begin{aligned} -2\omega A_1 \sin\psi - 2\omega B_1 \cos\psi &= f_{10} e^{i\psi} + f_{-10} e^{-i\psi} = \\ &= (f_{10}+f_{-10}) \cos\psi + i(f_{10}-f_{-10}) \sin\psi. \end{aligned} \tag{12}$$

Here

$$\begin{aligned} f_{10}+f_{-10} &= \frac{1}{4\pi^2} \int_0^{2\pi}\int_0^{2\pi} f(\ldots)(e^{-i\xi}+e^{i\xi})\, d\xi\, d\theta = \\ &= \frac{1}{2\pi^2} \int_0^{2\pi}\int_0^{2\pi} f(\theta, \cos\xi, -\omega\sin\xi) \cos\xi\, d\xi\, d\theta = I_1, \\ i(f_{10}-f_{-10}) &= \frac{1}{4\pi^2} \int_0^{2\pi}\int_0^{2\pi} f(\ldots) i(e^{-i\xi}-e^{i\xi})\, d\xi\, d\theta = \\ &= \frac{1}{2\pi^2} \int_0^{2\pi}\int_0^{2\pi} f(\theta, \cos\xi, -\omega\sin\xi) \sin\xi\, d\xi\, d\theta = I_2, \end{aligned} \tag{13}$$

involving

$$A_1 = -\frac{1}{2\omega} I_2, \quad B_1 = -\frac{1}{2\omega} I_1 \tag{14}$$

and from (10)

$$u_{1mn} = \frac{1}{(m, n)} f_{mn}, \quad (m, n) \neq 0 \text{ or } (m\pm 1)^2+n^2 > 0 \tag{15}$$

or

$$u_1(vt, \psi) = \sum_{\substack{m \quad n \\ m,n)\neq 0}}\sum \frac{1}{(m, n)} f_{mn} e^{i(m\psi+nvt)}. \tag{16}$$

The real form of u_1 is

(17) $$u_1 = \sum_{\substack{m=0\\(m,n)\neq 0}}^{\infty} \sum_{n=0}^{\infty} \frac{1}{(m,n)} [f_{mn} e^{i\Psi} + f_{-m,-n} e^{-i\Psi}] = \sum\sum \frac{1}{(mn)} (\alpha_{mn} \cos \Psi + \beta_{mn} \sin \Psi)$$
$$(\Psi = m\psi + n\nu t)$$

where similarly as above in (13)

(18) $$\alpha_{mn} = f_{mn} + f_{-m,-n} = \frac{1}{2\pi^2} \int_0^{2\pi} \int_0^{2\pi} f(\theta, \cos\xi, -\omega \sin\xi) \cos(m\xi + n\theta)\, d\xi\, d\theta,$$
$$\beta_{mn} = i(f_{mn} - f_{-m,-n}) = \frac{1}{2\pi^2} \int_0^{2\pi} \int_0^{2\pi} f(\theta, \cos\xi, -\omega \sin\xi) \sin(m\xi + n\theta)\, d\xi\, d\theta.$$

(m is odd!)

In first approximation $\dot{a} = a\varepsilon A_1$, $\dot{\psi} = \omega + \varepsilon B_1$ whence

$$a = a_0 e^{\varepsilon A_1 t}, \quad \psi = (\psi_0 + \omega t) + \varepsilon B_1 t.$$

Example. Letting $f(\theta, u, v) = \cos\theta \dfrac{u^3}{u^2 + \dfrac{v^2}{\omega^2}}$ we have

(19) $$f(\nu t, \cos\psi, -\omega \sin\psi) = \cos \nu t \cos^3 \psi = \frac{3}{8} \cos(\psi - \nu t) + \frac{3}{8} \cos(\psi + \nu t) +$$
$$+ \frac{1}{8} \cos(3\psi - \nu t) + \frac{1}{8} \cos(3\psi + \nu t)$$

whence $\alpha_{1,\pm 1} = \frac{3}{8}$, $\alpha_{3,\pm 1} = \frac{1}{8}$, $\beta_{m,n} = 0$ the rests of α_{mn}'s are zero, involving

(20) $$u_1 = \frac{3}{8(1,-1)} \cos(\psi - \nu t) + \frac{3}{8(1,1)} \cos(\psi + \nu t) +$$
$$+ \frac{1}{8(3,-1)} \cos(3\psi - \nu t) + \frac{1}{8(3,1)} \cos(3\psi + \nu t),$$

where

$$(1,-1) = \nu(2\omega - \nu), \quad (1,1) = -\nu(2\omega + \nu)$$

$$(3,-1) = -(2\omega - \nu)(4\omega - \nu), \quad (3,1) = -(2\omega + \nu)(4\omega + \nu)$$

and $A_1 = B_1 = 0$ and so $a = a_0$, $\psi = \psi_0 + \omega t$. The first approximation is

(21) $$x = a_0[\cos\psi + \varepsilon u_1(\nu t, \psi)], \quad \psi = \psi_0 + \omega t$$

i.e., in general beside the vibration with frequency ω also the combination vibrations with frequencies $\omega \pm \nu$, $3\omega \pm \nu$ appear, too.

The equation of the second approximation reads

$$-2\omega A_2 \sin\psi - 2\omega B_2 \cos\psi + S_2(u_2) +$$

$$+(A_1^2 - B_1^2)\cos\psi - 3A_1B_1\sin\psi + 2\omega A_1 \frac{\partial u_1}{\partial\psi} + 2A_1\nu\frac{\partial u_1}{\partial\theta} + \tag{22}$$

$$+2B_1\frac{\partial^2 u_1}{\partial\psi^2} + 2B_1\nu\frac{\partial^2 u_1}{\partial\theta\,\partial\psi} = F(\nu t, \cos\psi, \sin\psi),$$

where

$$F = f_x(\Omega)u_1 + f_{\dot x}(\Omega)\left(A_1\cos\psi - B_1\sin\psi + \omega\frac{\partial u_1}{\partial\psi} + \nu\frac{\partial u_1}{\partial\theta}\right) = \sum_{-\infty}^{\infty}\sum F_{mn}e^{i(m\psi+n\nu t)}. \tag{23}$$

If

$$u_2 = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} u_{2mn}e^{i(m\psi+n\nu t)}, \tag{24}$$

then

$$S_2(u_2) = \sum_{m,n=-\infty}^{\infty}\sum (m,n)\,u_{2mn}\,e^{i(m\psi+n\nu t)}. \tag{25}$$

The terms of $S_2(u_2)$ with $m=\pm 1, n=0$ vanish. u_1 and its derivatives do not involve $\sin\psi$ and $\cos\psi$, therefore A_2, B_2 must be determined from the following equation:

$$-2\omega A_2\sin\psi - 2\omega B_2\cos\psi + (A_1^2 - B_1^2)\cos\psi - 3A_1B_1\sin\psi =$$

$$= F_{10}e^{i\psi} + F_{-1,0}e^{-i\psi} = F_c\cos\psi + F_s\sin\psi, \quad F_c = F_{10} + F_{-10}, \quad F_s = i(F_{10} - F_{-10})$$

whence $-2\omega A_2 - 3A_1B_1 = F_s$, $-2\omega B_2 + A_1^2 - B_1^2 = F_c$ or $A_2 = \dfrac{3A_1B_1 - F_s}{2\omega}$, $B_2 = \dfrac{A_1^2 - B_1^2 - F_c}{2\omega}$. Then

$$\sum_{(m,n)\neq 0}\sum (m,n)\,u_{2mn}e^{i(m\psi+n\nu t)} = G(\nu t, \cos\psi, \sin\psi),$$

where

$$G = \sum_{(m,n)\neq 0}\sum F_{mn}e^{i(m\psi+n\nu t)} - 2\omega A_1\frac{\partial u_1}{\partial\psi} - 2A_1\nu\frac{\partial u_1}{\partial\theta} - 2B_1\frac{\partial^2 u_1}{\partial\psi^2} - 2B_1\nu\frac{\partial^2 \mu_1}{\partial\theta\,\partial\psi},$$

which does not involve $\sin\psi$ and $\cos\psi$. Hence

$$u_{2mn} = \frac{1}{(m,n)}G_{mn}, \quad (m,n)\neq 0.$$

Here F_{mn} and G_{mn} are the Fourier coefficients of F and G, respectively. Finally

$$u_2 = \sum_{\substack{m=0\\(m,n)\neq 0}}^{\infty}\sum_{n=0}^{\infty}\frac{1}{(m,n)}\left(\gamma_{mn}\cos(m\psi+n\nu t) + \delta_{mn}\sin(m\psi+n\nu t)\right), \tag{26}$$

where

$$\gamma_{mn} = G_{mn}+G_{-m,\,-n} = \frac{1}{2\pi^2}\int_0^{2\pi}\int_0^{2\pi} G(\theta, \cos\xi, \sin\xi)\cos(m\xi+n\theta)\,d\xi\,d\theta,$$

$$\delta_n = i(G_{mn}-G_{-m,\,-n}) = \frac{1}{2\pi^2}\int_0^{2\pi}\int_0^{2\pi} G(\theta, \cos\xi, \sin\xi)\sin(m\xi+n\theta)\,d\xi\,d\theta.$$

In the above example $A_1=B_1=0$ and (22) reads

$$(27)\qquad -2\omega A_2\sin\psi-2\omega B_2\cos\psi+S_2(u_2)=f_x(\Omega)u_1+f_{\dot x}(\Omega)\left(\omega\frac{\partial u_1}{\partial\psi}+\nu\frac{\partial u_1}{\partial\theta}\right).$$

Now

$$f_u=\cos\theta\frac{u^2+\dfrac{3v^2}{\omega^2}}{\left(u^2+\dfrac{v^2}{\omega^2}\right)^2},\quad f_v=-\cos\theta\frac{2u^3v}{\left(u^2+\dfrac{v^2}{\omega^2}\right)^2},$$

thus

$$f_x(\Omega)=\frac{1}{8}[6\cos\nu t+2\cos(2\psi-\nu t)+2\cos(\psi+\nu t)-\cos(4\psi-\nu t)-\cos(4\psi+\nu t)],$$

$$f_{\dot x}(\Omega)=\frac{1}{8\omega}[2\sin(2\omega-\nu t)+2\sin(2\psi+\nu t)+\sin(4\psi-\nu t)+\sin(4\psi+\nu t)]$$

and so the right side of (27) can be formed and the values of A_2, B_2, u_{2mn} can be computed. From these we get a and ψ in second approximation. The result of the somewhat lengthy computation is the following

$$A_2=\frac{72\omega^3-6\omega^2\nu-11\omega\nu^2-2\nu^3}{128\omega^2(4\omega^2-\nu^2)(16\omega^2-\nu^2)},\quad B_2=-\frac{744\omega^2-6\omega\nu-51\nu^2}{256\omega(4\omega^2-\nu^2)(16\omega^2-\nu^2)},$$

$$\dot a=a\varepsilon^2A_2,\quad \dot\psi=\omega+\varepsilon^2B_2\quad\text{whence}\quad a=a_0e^{\varepsilon^2A_2t},\quad \psi=(\psi_0+\omega t)+\varepsilon^2B_2t,$$

$$u_2=\sum_{m=0}^{m}\sum_{n=0}^{\infty}(\alpha_{mn}\cos(m\psi+n\nu t)+\beta_{mn}\sin(m\psi+n\nu t))$$

where the coefficients different from zero are as follows

$$\alpha_{1,\,-2}=\frac{48\omega^2+34\omega\nu+7\nu^2}{256\nu^2(4\omega^2-\nu^2)(16\omega^2-\nu^2)},\quad \beta_{1,\,-2}=\frac{-48\omega^3+10\omega^2\nu+3\omega\nu^2-\nu^3}{256\omega\nu(\omega-\nu)(4\omega^2-\nu^2)(16\omega^2-\nu^2)},$$

$$\alpha_{12}=\frac{-384\omega^2+376\omega^2\nu+42\omega\nu^2+25\nu^3}{512\nu^2(\omega+\nu)(4\omega^2-\nu^2)(16\omega^2-\nu^2)},\quad \beta_{12}=-\frac{3\omega+\nu}{256\omega\nu(\omega+\nu)(2\omega+\nu)(4\omega+\nu)},$$

$$\alpha_{30}=3\frac{-32\omega^3-16\omega^2\nu+2\omega\nu^2+5\nu^3}{1024\omega^3\nu(4\omega^2-\nu^2)(16\omega^2-\nu^2)},\quad \beta_{30}=\frac{9}{512\omega^2(4\omega^2-\nu^2)},$$

$$\alpha_{3,-2} = -\frac{3}{128\nu(\omega-\nu)(2\omega-\nu)(4\omega-\nu)}, \quad \beta_{3,-2} = -\frac{3}{256\omega\nu(2\omega-\nu)^2},$$

$$\alpha_{3,2} = \frac{24\omega^2+4\omega\nu-3\nu^2}{512\nu(\omega+\nu)(2\omega+\nu)(4\omega^2-\nu^2)(16\omega^2-\nu^2)}, \quad \beta_{32} = \frac{3}{256\omega\nu(2\omega+\nu)^2},$$

$$\alpha_{50} = \frac{-96\omega^3-48\omega^2\nu+6\omega\nu^2+7\nu^3}{3072\omega^2\nu(4\omega^2-\nu^2)(16\omega^2-\nu^2),}, \quad \beta_{50} = \frac{32\omega^2+\nu^2}{512\omega^2(4\omega^2-\nu^2)(16\omega^2-\nu^2)},$$

$$\alpha_{52} = -\frac{12\omega+\nu}{512\nu(2\omega+\nu)^2(3\omega+\nu)(4\omega+\nu)}, \quad \beta_{52} = \frac{12\omega^2+21\omega\nu+5\nu^2}{512\omega\nu(2\omega+\nu)^2(3\omega+\nu)(4\omega+\nu)},$$

$$\alpha_{5,-2} = \frac{1}{256(2\omega-\nu)^2(3\omega-\nu)(4\omega-\nu)}, \quad \beta_{5,-2} = \frac{-12\omega^2+9\omega\nu+7\nu^2}{512\omega\nu(2\omega-\nu)^2(3\omega-\nu)(4\omega-\nu)},$$

$$(\alpha_{7,-2} = 0), \qquad \beta_{7,-2} = \frac{1}{512\omega(2\omega-\nu)(4\omega-\nu)^2},$$

$$\alpha_{70} = -\frac{1}{6144\omega^2(2\omega-\nu)(4\omega-\nu)}, \quad \beta_{70} = \frac{16\omega^3-4\omega^2\nu-\omega\nu^2}{2044\omega^3(4\omega^2-\nu^2)(16\omega^2-\nu^2)},$$

$$\alpha_{72} = -\frac{1}{512(2\omega+\nu)(3\omega+\nu)(4\omega+\nu)}, \quad \beta_{72} = \frac{1}{512(2\omega+\nu)(4\omega+\nu)^2}.$$

In consequence, a lot of new combination frequencies appear, but the higher ones have step by step smaller amplitudes.

B. *The case of resonance*

By this denomination it will be meant the case where a relation of the form

$$\omega^2 = \left(\frac{p}{q}\nu\right)^2 + \Delta\varepsilon, \quad p, q \in \mathbf{Z}, (p, q) = 1 \tag{28}$$

holds between ω and ν where ν, Δ and the integers p, q are fixed, i.e. ω is function of ε. This is the point in which cases A and B differ. The case $\Delta=0$ is included. In the present case the series used here and expressing the solution will not be convergent. However, its partial sums give good approximation. Otherhand the series is often finite, causing no problem. On account of the nature of its influence on the solution the term $-\varepsilon\Delta x$ will be attached to the forcing term as follows

$$\ddot{x}+\omega'^2 x = \varepsilon[f(\nu t, x, \dot{x})-\Delta x], \qquad \left(\omega' = \frac{p}{q}\nu\right). \tag{29}$$

The function $f(\tau, u, v)$ has the period 2π in τ and is assumed to be analytic on the cylinder $u^2+\frac{v^2}{\omega'^2}=1$. The solution of the homogeneous equation $\ddot{x}+\omega'^2 x=0$

is $x=a\cos(\omega' t+t_0)$, $a=\text{const}$, $t_0=\text{const}$. In view of this the solution of (29) will be assumed in the form

$$x = a\,[\cos\psi+\varepsilon u_1(\nu t,\psi)+\varepsilon^2 u_2(\nu t,\psi)+\dots] \tag{30}$$

where in general the "phase" ψ and "amplitude" a are not constant anymore, but they have to satisfy the equations

$$\begin{aligned}\dot a &= a\,[\varepsilon A_1(\vartheta)+\varepsilon^2 A_2(\vartheta)+\dots],\\ \dot\psi &= \omega'+\varepsilon B_1(\vartheta)+\varepsilon^2 B_2(\vartheta)+\dots,\quad (\vartheta=\psi-\omega' t)\end{aligned} \tag{31}$$

involving

$$\dot\vartheta = \varepsilon B_1(\vartheta)+\varepsilon^2 B_2(\vartheta)+\dots, \tag{31'}$$

i.e., we suppose that $\dot a$ and $\dot\psi$ are functions of the unique variable $\vartheta=\vartheta(t)$ — the "phase-shift" — which is also time dependent. We assume that $u_i(\tau,\psi)$ is periodic in τ and ψ with the period 2π and the derivatives appearing here all exist and are continuous. These assumptions are suggested by arguments taken from the physics but must be justified by the tools of the analysis. — Now we have by (30)—(31)—(31′)

$$\dot x = \dot a\,(\cos\psi+\varepsilon u_1+\dots)+a\dot\psi\left(-\sin\psi+\varepsilon\frac{\partial u_1}{\partial\psi}+\dots\right)+a\nu\left(\varepsilon\frac{\partial u_1}{\partial\tau}+\varepsilon^2\frac{\partial u_2}{\partial\tau}+\dots\right)=$$

$$=a\left\{-\omega'\sin\psi+\varepsilon\left[A_1\cos\psi-B_1\sin\psi+\omega'\frac{\partial u_1}{\partial\psi}+\nu\frac{\partial u_1}{\partial\tau}\right]+\right.$$

$$\left.+\varepsilon^2\left[A_2\cos\psi-B_2\sin\psi+\omega'\frac{\partial u_2}{\partial\psi}+\nu\frac{\partial u_2}{\partial\tau}+A_1u_1+B_1\frac{\partial u_1}{\partial\psi}\right]+\varepsilon^3\dots\right\}$$

and

$$\ddot x = \ddot a\,(\cos\psi+\varepsilon u_1+\dots)+2\dot a\dot\psi\left(-\sin\psi+\varepsilon\frac{\partial u_1}{\partial\psi}+\dots\right)+2\dot a\nu\left(\varepsilon\frac{\partial u_1}{\partial\tau}+\varepsilon^2\frac{\partial u_2}{\partial\tau}+\dots\right)+$$

$$+a\ddot\psi\left(-\sin\psi+\varepsilon\frac{\partial u_1}{\partial\psi}+\dots\right)+a\dot\psi^2\left(-\cos\psi+\varepsilon\frac{\partial^2u_1}{\partial\psi^2}+\dots\right)+$$

$$+2\nu a\dot\psi\left(\varepsilon\frac{\partial^2u_1}{\partial\tau\,\partial\psi}+\varepsilon^2\frac{\partial^2u_2}{\partial\tau\,\partial\psi}+\dots\right)+a\nu^2\left(\varepsilon\frac{\partial^2u_1}{\partial\tau^2}+\varepsilon^2\frac{\partial^2u_2}{\partial\tau^2}+\dots\right).$$

By (31)—(31′)

$$\ddot a = \dot a\,(\varepsilon A_1+\dots)+a\,(\varepsilon A_1'+\varepsilon^2A_2'+\dots)\dot\vartheta = a\,(\varepsilon A_1+\dots)^2+a\,(\varepsilon A_1'+\dots)(\varepsilon B_1+\dots)=$$

$$=\varepsilon^2a\,(A_1^2+A_1'B_1)+\varepsilon^3\dots,\quad \left('=\frac{d}{d\vartheta}\right)$$

$$\dot a\dot\psi = a\,[\varepsilon\omega'A_1+\varepsilon^2(A_1B_1+\omega'A_2)+\varepsilon^3\dots],$$

$$\ddot\psi = (\varepsilon B_1'+\varepsilon^2B_2'+\dots)\,\dot\vartheta = \varepsilon^2B_1B_2+\varepsilon^3\dots,$$

$$\dot\psi^2 = \omega'^2+\varepsilon\cdot 2\omega'B_1+\varepsilon^2(B_1^2+2\omega'B_2)+\varepsilon^3\dots.$$

Therefore

$$
\begin{aligned}
\ddot{x} = a\Big\{&-\omega'^2\cos\psi+\varepsilon\left[-2\omega'(A_1\sin\psi+B_1\cos\psi)+S_1(u_1)\right]+\\
&+\varepsilon^2\Big[-(2\omega' A_2+B_1B_2)\sin\psi-2\omega' B_2\cos\psi+S_1(u_2)+\\
&+(A_1^2-B_1^2+A_1'B_1)\cos\psi-2A_1B_1\sin\psi+2\omega' A_1\frac{\partial u_1}{\partial\psi}+\\
&+2A_1\nu\frac{\partial u_1}{\partial\tau}+2\omega' B_1\frac{\partial^2 u_1}{\partial\psi^2}+2B_1\nu\frac{\partial^2 u_1}{\partial\tau\,\partial\psi}\Big]+\varepsilon^3\ldots\Big\}
\end{aligned}
$$

$$
\left\{S_1(u_i)=\left(\omega'\frac{\partial}{\partial\psi}+\nu\frac{\partial}{\partial\tau}\right)^2 u_i=\omega'^2\frac{\partial^2 u_i}{\partial\psi^2}+2\omega'\nu\frac{\partial^2 u_i}{\partial\tau\,\partial\psi}+\nu^2\frac{\partial^2 u_i}{\partial\tau^2},\qquad i=1,2,\ldots\right\}
$$

and

$$
\begin{aligned}
\ddot{x}+\omega'^2x = a\Big\{&\varepsilon\left[-2\omega'(A_1\sin\psi+B_1\cos\psi)+S_2(u_1)\right]+\varepsilon^2\Big[-(2\omega' A_2+B_1B_2)\sin\psi-\\
&-2\omega' B_2\cos\psi+S_2(u_2)+(A_1^2-B_1^2+A_1'B_1)\cos\psi-2A_1B_1\sin\psi+\\
&+2\omega' A_1\frac{\partial u_1}{\partial\psi}+2A_1\nu\frac{\partial u_1}{\partial\tau}+2\omega' B_1\frac{\partial^2 u_1}{\partial\psi^2}+2B_1\nu\frac{\partial^2 u_1}{\partial\tau\,\partial\psi}\Big]+\varepsilon^3\ldots\Big\}.
\end{aligned}
$$

Here

$$S_2(u_i)=S_1(u_i)+\omega'^2u_i.$$

The Taylor expansion of $\varepsilon[f(\nu t, x, \dot{x})-\Delta x]$ about the point $\Omega(\nu t, \cos\psi, -\omega'\sin\psi)$ reads as

$$\varepsilon[f(\nu t, x, \dot{x})-\Delta x]=\varepsilon a\left[f(\Omega)-\Delta\cos\psi\right]+\varepsilon^2 af_1(\nu t,\psi)+\varepsilon^3\ldots,$$

where

$$f_1(\nu t,\psi)=f_x(\Omega)u_1+f_{\dot{x}}(\Omega)\left(A_1\cos\psi-B_1\sin\psi+\omega'\frac{\partial u_1}{\partial\psi}+\nu\frac{\partial u_1}{\partial\tau}\right)-\Delta u_1.$$

Hence the equations of the first and second approximations are

$$
-2\omega' A_1\sin\psi-(2\omega' B_1-\Delta)\cos\psi+S_2(u_1)=f(\Omega)\overset{\text{def}}{\equiv}f_0(\nu t,\psi), \tag{32}
$$

$$
\begin{aligned}
&-(2\omega' A_2+B_1B_2)\sin\psi-2\omega' B_2\cos\psi+S_2(u_2)=f_1(\nu t,\psi)-\\
&-(A_1^2-B_1^2+A_1'B_1)\cos\psi+2A_1B_1\sin\psi-2\omega' A_1\frac{\partial u_1}{\partial\psi}-2A_1\nu\frac{\partial u_1}{\partial\tau}-\\
&-2\omega' B_1\frac{\partial^2 u_1}{\partial\psi^2}+2B_1\nu\frac{\partial^2 u_1}{\partial\tau\,\partial\psi}.
\end{aligned} \tag{33}
$$

Assume now u_1 and f_0 in the form

$$
\begin{aligned}
u_1(\nu t,\psi)&=\sum_{m,n=-\infty}^{\infty}\sum u_{mn}^{(1)}e^{i(m\psi+n\nu t)};\quad u_{mn}^{(1)}=\frac{1}{4\pi^2}\iint_0^{2\pi}u_1(\tau,\xi)e^{-(im\xi+n\tau)}\,d\xi\,d\tau,\\
f_0(\nu t,\psi)&=\sum_{m,n=-\infty}^{\infty}\sum f_{mn}^{(0)}e^{i(m\psi+n\nu t)};\quad f_{mn}^{(0)}=\frac{1}{4\pi^2}\iint_0^{2\pi}f_0(\tau,\xi)e^{-i(m\xi+n\tau)}\,d\xi\,d\tau.
\end{aligned} \tag{34}
$$

Then

$$S_2(u_1) = \sum_{m,n=-\infty}^{\infty}\sum (m, n) u_{mn}^{(1)} e^{i(m\psi+nvt)}$$ [6]

where

$$(m, n) = \overline{(m, n)}\, \frac{v^2}{q^2}, \quad \overline{(m, n)} = p^2-(mp+nq)^2$$

and (32) will have the form

$$(35) \quad -2\omega' A_1 \sin\psi-(2\omega' B_2-\varDelta)\cos\psi+\sum_{m,n=-\infty}^{\infty}\sum [(m, n)u_{mn}^{(1)}-f_{mn}^{(0)}]e^{i(m\psi+nvt)} = 0.$$

We have the implication $(m, n)=0 \Leftrightarrow (m\pm 1)p+nq=0$ which shows that if $(m, n)=0$, then $m\pm 1$ is a multiple of q, involving

$$m\pm 1 = \sigma q, \quad n = -\sigma p, \quad \sigma\in \mathbf{Z}$$

and σ varies from $-\infty$ to ∞, since m and n do the same. Thus being in this case $\left(\frac{p}{q} v=\omega'\right)$

$$m\psi+nvt = \mp\psi+(m\pm 1)\psi+nvt = \mp\psi+\sigma q\psi-\sigma pvt =$$

$$= \mp\psi+\sigma q\left(\psi-\frac{p}{q}vt\right) = \mp\psi+\sigma q\vartheta$$

we have from (35) the following equation to determine A_1 and B_1:

$$(36) \quad 2\omega' A_1 \sin\psi+(2\omega' B_1-\varDelta)\cos\psi+\sum_{\sigma=-\infty}^{\infty} f_{\sigma q\mp 1,\,-\sigma p}^{(0)} e^{i(\mp\psi+\sigma q\vartheta)} = 0.$$

The last term can be written in the form

$$\sum_{\sigma=-\infty}^{\infty} (\alpha_\sigma \cos\psi+\beta_\sigma \sin\psi)\, e^{i\sigma q\vartheta}$$

where

$$\alpha_\sigma = f_{\sigma q+1,\,-\sigma p}^{(0)}+f_{\sigma q-1,\,-\sigma p}^{(0)}, \quad \beta_\sigma = i(f_{\sigma q+1,\,-\sigma p}^{(0)}-f_{\sigma q-1,\,-\sigma p}^{(0)}).$$

Equation (36) can and will be satisfied by equating there the coefficients of $\cos\psi$ and $\sin\psi$ to 0. However, it can happen that it is not the unique possibility to satisfy (36), since the functions ψ and ϑ are not independent. Nevertheless, if the solution so obtained is an exact solution (the series are convergent or finite), then the uniqueness is assured by the uniqueness of the solution of (29) at given initial conditions. — The comparison of the coefficients gives

$$(37) \quad 2\omega' A_1 = -\sum_{\sigma=-\infty}^{\infty} \beta_\sigma e^{i\sigma q\vartheta}, \quad 2\omega' B_1-\varDelta = -\sum_{\sigma=-\infty}^{\infty} \alpha_\sigma e^{i\sigma q\vartheta}.$$

[6] Notice that in the formula of $f_0(vt, \psi)$ m is odd and in consequence so is in $u_1(vt, \psi)$, $S_2(u_1)$ etc., too.

By (34)

$$\alpha_\sigma = \frac{1}{4\pi^2}\iint\limits_0^{2\pi} f_0(\tau,\xi)[e^{-i[(\sigma q+1)\xi-\sigma p\tau]}+e^{-i[(\sigma q-1)\xi-\sigma p\tau]}]\,d\xi\,d\tau =$$

$$= \frac{1}{4\pi^2}\iint\limits_0^{2\pi} f_0(\tau,\xi)\,(e^{i\xi}+e^{-i\xi})e^{-i\sigma q\vartheta'}\,d\xi\,d\tau =$$

$$= \frac{1}{2\pi^2}\iint\limits_0^{2\pi} f_0(\tau,\xi)\cos\xi e^{-i\sigma q\vartheta'}\,d\xi\,d\tau \qquad \left(\vartheta' = \xi-\frac{p}{q}\tau\right)$$

and similarly

$$\beta_\sigma = \frac{1}{2\pi^2}\iint\limits_0^{2\pi} f_0(\tau,\xi)\sin\xi e^{-i\sigma q\vartheta'}\,d\xi\,d\tau.$$

The right sides of (37) will have the form

$$\sum_{\sigma=-\infty}^{\infty}\alpha_\sigma e^{i\sigma q\vartheta} = \frac{1}{2\pi^2}\sum_{\sigma=0}^{\infty}{}'\iint\limits_0^{2\pi} f_0(\tau,\xi)\cos\xi\,[e^{i\sigma q(\vartheta-\vartheta')}+e^{-i\sigma q(\xi-\xi')}]\,d\xi\,d\tau =$$

$$= \frac{1}{\pi^2}\sum_{\sigma=0}^{\infty}{}'\iint\limits_0^{2\pi} f_0(\tau,\xi)\cos\xi\cos\sigma q(\vartheta-\vartheta')\,d\xi\,d\tau,$$

$$\sum_{\sigma=-\infty}^{\infty}\beta_\sigma e^{i\sigma q\vartheta} = \frac{1}{\pi^2}\sum_{\sigma=0}^{\infty}{}'\iint\limits_0^{2\pi} f_0(\tau,\xi)\sin\xi\cdot\cos\sigma q(\vartheta-\vartheta')\,d\xi\,d\tau.$$

The primes attached to the signs Σ means that in the term corresponding to $\sigma=0$ the factor $\frac{1}{\pi^2}$ must be replaced by $\frac{1}{2\pi^2}$. Then A_1 and B_1 can be obtained from (37).

If $(m,n)\neq 0$, then from (35) $u_{mn}^{(1)} = \frac{1}{(m,n)}f_{mn}^{(0)}$ giving

$$u_1(vt,\psi) = \frac{q}{4\pi^2 v^2}\sum_{\substack{m,n=-\infty\\(m,n)\neq 0}}^{\infty}\sum\frac{e^{i(m\psi+nvt)}}{\overline{(m,n)}}\iint\limits_0^{2\pi} f(\tau,\xi)e^{-i(m\xi+n\tau)}\,d\xi\,d\tau =$$

$$= \frac{q}{4\pi^2 v^2}\sum_{\substack{m,n=0\\(m,n)\neq 0}}^{\infty}\sum\frac{1}{\overline{(m,n)}}\iint\limits_0^{2\pi} f_0(\tau,\xi)[e^{i[m(\psi-\xi)+n(vt-\tau)]}+e^{-[m(\psi-\xi)+n(vt-\tau)]}]\,d\xi\,d\tau =$$

$$= \frac{q}{2\pi^2 v^2}\sum_{\substack{m,n=0\\(m,n)\neq 0}}^{\infty}\sum\frac{1}{\overline{(m,n)}}\iint\limits_0^{2\pi} f_0(\tau,\xi)\cos[m(\psi-\xi)+n(vt-\tau)]\,d\xi\,d\tau.$$

Notice that the so called "small divisors" appear here. In consequence the series is not convergent in general. Otherwise, if the series is finite the above expression of u_1 is always meaningful. — The computation of the second approximation requires no new moments and may be omitted.

EXAMPLE. Letting $f(\tau, u, v)=\cos\tau \dfrac{u^3}{u^2+\dfrac{v^2}{\omega'^2}}$ we have

$$f_0(vt, \psi) = f(vt, \cos\psi, -\omega' \sin\psi) = \frac{3}{8}\cos(\psi - vt)+$$

$$+\frac{3}{8}\cos(\psi+vt)+\frac{1}{8}\cos(3\psi-vt)+\frac{1}{8}\cos(3\psi+vt).$$

Here (m, n) has the values

$$(1,-1)=\frac{2p-q}{q}v^2, \quad (3,-1)=\frac{(4p-q)(q-2p)}{q^2}v^2$$

$$(1,1)=-\frac{2p+q}{q}v^2, \quad (3,1)=-\frac{(4p+q)(2p+q)}{q^2}v^2.$$

Assuming $\frac{p}{q}>0$ the only possible values of $\frac{p}{q}$ for which $(m,n)=0$ are $\frac{1}{2}$ and $\frac{1}{4}$.

(i) If $\frac{p}{q}=\frac{1}{2}$ $(p=1, q=2)$, then $(1,-1)=(3,-1)=0$ and (being $vt=2\omega t= =2\psi-2\vartheta$)

$$\psi-vt=-\psi+2\vartheta, \quad \psi+vt=3\psi-2\vartheta, \quad 3\psi-vt=\psi+2\vartheta, \quad 3\psi+vt=5\psi-2\vartheta,$$

involving

$$u_1=-\frac{3}{16v^2}\cos(3\psi-2\vartheta)-\frac{1}{48v^2}\cos(5\psi-2\vartheta).$$

(ii) If $\frac{p}{q}=\frac{1}{4}$ $(p=1, q=4)$, then $(3,-1)=0$ and

$$\psi-vt=-3\psi+4\vartheta, \quad \psi+vt=5\psi-4\vartheta, \quad 3\psi-vt=-\psi+4\vartheta, \quad 3\psi+vt=7\psi-4\vartheta,$$

therefore

$$u_1=-\frac{3}{4v^2}\cos(3\psi-4\vartheta)-\frac{1}{4v^2}\cos(5\psi-4\vartheta)-\frac{1}{24v^2}\cos(7\psi-4\vartheta).$$

In the case (i) we have for the determination of A_1, B_1

$$vA_1\sin\psi+(vB_1-\Delta)\cos\psi=-\frac{3}{8}\cos(-\psi+2\vartheta)-\frac{1}{8}\cos(\psi+2\vartheta)=$$

$$=-\frac{1}{2}\cos\psi\cos 2\vartheta-\frac{1}{4}\sin\psi\sin 2\vartheta,$$

whence

$$A_1=-\frac{1}{4v}\sin 2\vartheta, \quad B_1=\alpha+\beta\cos 2\vartheta, \quad \alpha=\frac{\Delta}{v}, \quad \beta=-\frac{1}{2v}$$

and from (31′) in first approximation

$$\dot{\vartheta} = \varepsilon B_1(\vartheta) = \varepsilon(\alpha+\beta\cos 2\vartheta)$$

$$I(\vartheta) = \int \frac{d\vartheta}{\alpha+\beta\cos 2\vartheta} = \varepsilon(t+t_0).$$

Here

$$I(\vartheta) = \begin{cases} \dfrac{1}{\sqrt{\beta^2-\alpha^2}} \log \dfrac{(\beta-\alpha)\operatorname{tg}\vartheta+\sqrt{\beta^2-\alpha^2}}{(\beta-\alpha)\operatorname{tg}\vartheta-\sqrt{\beta^2-\alpha^2}} & (\beta^2>\alpha^2) \\ \dfrac{1}{\sqrt{\alpha^2-\beta^2}} \operatorname{arctg} \dfrac{(\alpha-\beta)\operatorname{tg}\vartheta}{\sqrt{\alpha^2-\beta^2}} & (\beta^2<\alpha^2), \end{cases}$$

$$I(\vartheta) = \begin{cases} \dfrac{\alpha}{2}\operatorname{tg}\vartheta \quad (\alpha=\beta) \\ -\dfrac{\alpha}{2}\operatorname{ctg}\vartheta \quad (\alpha=-\beta) \end{cases}$$

whence for $\Delta=0$

$$\operatorname{tg}\vartheta = \frac{c_0 e^{-\frac{\varepsilon t}{2\nu}}+1}{c_0 e^{-\frac{\varepsilon t}{2\nu}}-1} = f(t), \quad \vartheta = \operatorname{arctg} f(t), \quad c_0 = \text{const} \neq 0.$$

If $\Delta \neq 0$, then the computation goes in the same way. Furthermore

$$\dot{a} = a\varepsilon A_1(\vartheta) = -a\frac{\varepsilon}{4\nu}\sin 2\vartheta = -a\frac{\varepsilon}{2\nu}\frac{c_0^2 e^{-\frac{\varepsilon t}{\nu}}-1}{c_0^2 e^{-\frac{\varepsilon t}{\nu}}+1},$$

whence

$$a = a_0\sqrt{c_0^2 e^{-\frac{\varepsilon t}{2\nu}}+e^{\frac{\varepsilon t}{2\nu}}}, \qquad a_0 = \text{const} \neq 0.$$

As long as $\frac{\varepsilon t}{\nu} \ll 1$, ϑ and a are near constant.

In the case (ii) we have

$$\frac{\nu}{2}A_1\sin\psi+\left(\frac{\nu}{2}B_1-\Delta\right)\cos\psi = -\frac{1}{8}\cos(-\psi+4\vartheta) =$$

$$= -\frac{1}{8}\cos\psi\cos 4\vartheta-\frac{1}{8}\sin\psi\sin 4\vartheta,$$

whence

$$A_1 = -\frac{1}{4\nu}\sin 4\vartheta, \quad B_1 = \alpha+\beta\cos 4\vartheta, \quad \alpha = \frac{2\Delta}{\nu}, \quad \beta = -\frac{1}{4\nu}$$

and the further computation follows previous lines.

III. The convergence of the series of u_1 in (A)

The form of u_1 in the series of x is the following:

$$\sum_{m,n=-\infty}^{\infty}\sum \frac{e^{i(m\psi+nvt)}}{(m,n)}\alpha_{mn}, \quad \alpha_{mn}=\iint_0^{2\pi} f_0(\tau,\psi)e^{-i(m\psi+n\tau)}\,d\psi\,d\tau$$

and that of u_2 is a similar one. Here

$$(m,n)=\omega^2-(m\omega+nv)^2=-[nv+(m+1)\omega][nv+(m-1)\omega].$$

Now the case $\dfrac{v}{\omega}=-\dfrac{m\pm1}{n}$ is excluded, i.e., $\dfrac{v}{\omega}$ cannot be rational, but it is not sufficient, since $|(m,n)|$ can and will be arbitrarily small for pairs m, n in infinite number. Nevertheless — as it is well-known — (see, e.g., in [1] pp. 177—178) to almost all real numbers $\dfrac{v}{\omega}$ there can be found such numbers $C>0$, $\delta>0$ for which

$$\left|\frac{v}{\omega}-\frac{p}{q}\right| \geqq \frac{c}{(|p|+|q|)^{2+\delta}}$$

holds for every integer p and q. Therefore, in the present case

$$\left|\frac{v}{\omega}+\frac{m\pm1}{n}\right| \geqq \frac{c}{(|n|+|m\pm1|)^{2+\delta}},$$

whence

$$|nv+(m\pm1)\omega| \geqq \frac{c\omega|n|}{(|n|+|m\pm1|)^{2+\delta}},$$

and

$$|(m,n)| \geqq \frac{c^2\omega^2n^2}{(|n|+|m+1|)^{4+2\delta}} \quad (m>1)$$

or

$$\frac{1}{|(m,n)|} \leqq \frac{(|n|+|m+1|)^{4+2\delta}}{c^2\omega^2n^2}.$$

On the other hand it is also known that for the Fourier coefficients α_{mn} the following estimate holds (see [2], vol. III, p. 494):

$$|\alpha_{mn}|=\frac{|\alpha_{mn}^{(k,l)}|}{m^kn^l}$$

where $\alpha_{mn}^{(k,l)}$ is the Fourier coefficient of $\dfrac{\partial^{k+l}f_0}{\partial\psi^k\partial\theta^l}$ and $k, l\in\mathbf{Z}$. This partial derivative exists, moreover it is analytic, since f_0 is. Thus the terms of the above series are dominated by

$$\frac{(|n|+|m+1|)^{4+2\delta}}{c^2\omega^2 n^2}\frac{|\alpha_{mn}^{(k,l)}|}{m^k n^l}$$

where k, l are arbitrary large integers, consequently, the series is convergent for almost all $\frac{\nu}{\omega}$.

REFERENCES

[1] Bogoliubov, N. N.—Mitropolski, J. A., *Asymptotische Methoden in der Theorie der nichtlinearen Schwingungen,* Berlin, 1965.

[2] Fichtenholz, G. M., *Differential- und Integralrechnung*, Berlin, 1972.

Mathematical Institute of the Hungarian Academy of Sciences
Reáltanoda u. 13—15, H—1053 Budapest

(Received March 10, 1978)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 319—326.

ON DIFFERENCE SETS

by

I. Z. RUZSA

1. Introduction

Let $S \subset \mathbf{N}$ and

$$D = \Delta_1 S = S - S = \{s-t\colon\ s, t \in S\} \tag{1.1}$$

its difference-set. It is known that if S has a positive upper density, then D does not contain arbitrarily large gaps and it has a positive lower density, moreover

$$\underline{d}(D) \geqq \overline{d}(S). \tag{1.2}$$

(The definition of $\underline{d}$ and $\overline{d}$ see in Section 2.) See RUZSA [1]; the first result is due to ERDŐS and SÁRKÖZY.

Recently STEWART and TIJDEMAN [2] generalized these results for the intersection of several difference sets. Instead of Δ_1 they considered

$$\Delta_2(S) = \{z\colon\ z+s \in S \text{ for infinitely many } s \in S\}. \tag{1.3}$$

Let $S_1, \dots, S_k$ be sequences of positive upper density. They proved that there exists another set S satisfying

$$\underline{d}(S) \geqq c_k \prod_{i=1}^{k} \overline{d}(S_i), \quad c_k = \bigl(5 \log (k+1)\bigr)^{-k} \tag{1.4}$$

and

$$\Delta_1 S \subset D = \bigcap \Delta_2 S_i,$$

which immediately yields a lower bound for $\underline{d}(D)$ and implies that D does not contain arbitrarily large gaps. Moreover they proved that there are integers $n_1, \dots, n_l$,

$$l \leqq \bigl(c_k \prod \overline{d}(S_i)\bigr)^{-\log 3/\log 2} \tag{1.5}$$

such that

$$\bigcup_{j=1}^{l} (D+n_j) = \mathbf{Z}.$$

Our aim is to improve these inequalities. We shall remove the constant c_k from (1.4), the constant c_k and the exponent $\log 3/\log 2$ from (1.5).

2. Results

By a "sequence" we shall mean a set $S \subset \mathbf{Z}$. First we define some kinds of densities. Let

$$S(x) = |S \cap [1, x]| \tag{2.1}$$

be the "counting function" of S, and let

$$\bar{d}(S) = \limsup \frac{S(x)}{x}, \quad \underline{d}(S) = \liminf \frac{S(x)}{x}, \tag{2.2}$$

the upper, resp. lower density of S. If they coincide, their common value is called the asymptotic density of S and it is denoted by $d(S)$. (Note that in these definitions the negative elements of S have been neglected.)

Let

$$S^*(x) = \max_{a \in \mathbf{Z}} |S \cap [a+1, a+x]| \tag{2.3}$$

and

$$S_*(x) = \min_{a \in \mathbf{Z}} |S \cap [a+1, a+x]|, \tag{2.4}$$

the upper, resp. lower counting function of S. Evidently, S^* is subadditive, S_* superadditive, that is for all $x, y \in \mathbf{N}$ we have

$$S^*(x+y) \leqq S^*(x) + S^*(y), \quad S_*(x+y) \geqq S_*(x) + S_*(y).$$

These properties imply that there exists

$$d^*(S) = \lim \frac{S^*(x)}{x} = \inf \frac{S^*(x)}{x} \tag{2.5}$$

and

$$d_*(S) = \lim \frac{S_*(x)}{x} = \sup \frac{S_*(x)}{x}, \tag{2.6}$$

which we shall call the strong upper, resp. lower density of S.

Evidently, we always have

$$d_*(S) \leqq \underline{d}(S) \leqq \bar{d}(S) \leqq d^*(S).$$

Note also that $d_*(S) > 0$ if and only if S does not contain arbitrarily large gaps.

Next we define three kinds of difference sets. Two of them have been defined in the previous section, but for sake of completeness we repeat them. Let

$$\Delta_1(S) = \{g\colon\ S \cap (S+g) \neq \emptyset\}, \tag{2.7}$$

$$\Delta_2(S) = \{g\colon\ |S \cap (S+g)| = \infty\}, \tag{2.8}$$

$$\Delta_3(S) = \{g\colon\ \bar{d}(S \cap (S+g)) > 0\}. \tag{2.9}$$

Their names are the ordinary difference set, the infinite difference set and the density difference set of S.

THEOREM 1. *For arbitrary sequences $S_1, \ldots, S_k$ there exist two other sequences S and S' such that*

a) $$\Delta_1(S) \subset \bigcap \Delta_2 S_i, \quad \underline{d}(S) \geqq \prod d^*(S_i);$$

b) $$\Delta_1 S' \subset \bigcap \Delta_3 S_i, \quad \underline{d}(S') \geqq \prod \bar{d}(S_i).$$

THEOREM 2. *Let S be an arbitrary sequence.*
a) *If $d^*(S)>0$, then there exist $k=[1/d^*(S)]$ numbers $a_1, \ldots, a_k$ such that*

$$\bigcup \big(\Delta_2(S)+a_i\big) = \mathbf{Z}.$$

b) *If $\bar{d}(S)>0$, then there exist $l=[1/\bar{d}(S)]$ numbers $b_1, \ldots, b_l$ such that*

$$\bigcup \big(\Delta_3(S)+b_j\big) = \mathbf{Z}.$$

(2.10) COROLLARY.
a) *If $d^*(S)>0$, then $\Delta_2(S)$ does not contain arbitrarily large gaps and we have*

$$\underline{d}(\Delta_2 S) \geqq d_*(\Delta_2 S) \geqq 1/k, \quad k = [1/d^*(S)].$$

b) *If $\bar{d}(S)>0$, then $\Delta_3 S$ does not contain arbitrarily large gaps and we have*

$$\underline{d}(\Delta_3 S) \geqq d_*(\Delta_3 S) \geqq 1/l, \quad l = [1/\bar{d}(S)].$$

Combining Theorem 1 and Corollary (2.10) we get:

(2.11) COROLLARY. *Let $S_1, \ldots, S_n$ be arbitrary sequences and*

$$D = \bigcap \Delta_2 S_j, \quad D' = \bigcap \Delta_3 S_j.$$

We have

$$d_*(D) \geqq \prod d^*(S_j),$$

$$d_*(D') \geqq \prod \bar{d}(S_j).$$

Moreover $\mathbf{Z}$ *can be covered by $\left[\prod d^*(S_j)^{-1}\right]$ translations of D and $\left[\prod \bar{d}(S_j)^{-1}\right]$ translations of D'.*

Theorem 1 and this Corollary contain the improvements of Stewart and Tijdeman's results mentioned in the introduction.

3. Homogeneous systems

Our main tool in achieving the results stated in the previous section will be the notion of a homogeneous system.

(3.1) DEFINITION. Let H be a set of finite sets of integers. H is called a homogeneous system, if for every $A \in H$ all the subsets and translations of A belong to H as well.

To a sequence S one can associate a homogeneous system in several ways; we shall deal with three of them. Let

(3.2) $h_1(S) = \{A\colon A+z\in S \text{ for some } z\in \mathbf{Z}\}$,

(3.3) $h_2(S) = \{A\colon A+z\in S \text{ for infinitely many } z\in \mathbf{Z}\}$,

(3.4) $h_3(S) = \{A\colon A+z\in S \text{ for a sequence of } z\text{'s, having positive upper density}\}$.

Given a homogeneous system H, we define its difference set by

$$\Delta H = \bigcup_{A\in H} (A-A). \tag{3.5}$$

(Note that ΔH is an ordinary sequence.) Evidently, we have

$$\Delta\bigl(h_j(S)\bigr) = \Delta_j S \qquad (j=1,2,3); \tag{3.6}$$

this is the main connection between sequences and homogeneous systems.

For a homogeneous system H we define its counting function by

$$H(x) = \max_{A\in H} |A\cap[1,x]|. \tag{3.7}$$

Evidently $H(x)$ is subadditive, which implies that there exists

$$dH = \lim_{x\to\infty} H(x)/x = \inf_{x\in\mathbf{N}} H(x)/x. \tag{3.8}$$

THEOREM 3. *For every sequence S we have*

$$d\bigl(h_1(S)\bigr) = d\bigl(h_2(S)\bigr) = d^*(S), \tag{3.9}$$

$$d\bigl(h_3(S)\bigr) \geqq \bar{d}(S). \tag{3.10}$$

PROOF. Let $H_j = h_j(S)$ $(j=1,2,3)$. Obviously, we have

$$H_2(x) \leqq H_1(x) = S^*(x),$$

therefore

$$d(H_2) \leqq d(H_1) = d^*(S).$$

We have to show that

$$d(H_2) \geqq d^*(S). \tag{3.11}$$

Regard the sets

$$A_z = [1,x]\cap(S-z).$$

Call z "average" if $A_w = A_z$ for infinitely many w's, and "special" in the other case. Evidently, there is only a finite number of special z's. Let

$$S_1 = S\setminus\bigcup\{(A_z+z)\colon z \text{ is special}\}. \tag{3.12}$$

Since S and S_1 differ only by a finite number of elements, we have

$$d^*(S) = d^*(S_1).$$

On the other hand, if z is average, then $A_z\in H_2$, therefore

$$H_2(x) \geqq \max\{|A_z|\colon z \text{ is average}\} \geqq S_1^*(x),$$

which implies

$$\frac{H_2(x)}{x} \geqq \frac{S_1^*(x)}{x} \geqq \inf \frac{S_1^*(t)}{t} = d^*(S_1) = d^*(S);$$

making $x \to \infty$ we get (3.11), which completes the proof of (3.9).

(3.10) can be proved similarly. This time we call z average, if $A_w = A_z$ for a sequence of w's having positive upper density and special in the other case. Again we define S_1 by (3.12). Evidently, the sequence of special z's has density 0, hence

$$\bar{d}(S) = \bar{d}(S_1),$$

and we have

$$H_3(x) = \max \{|A_z|:\ z \text{ is average}\},$$

which implies

$$\frac{H_3(x)}{x} \geqq \frac{S_1^*(x)}{x} \geqq d^*(S_1) \geqq \bar{d}(S_1) = \bar{d}(S).$$

This yields (3.10) immediately. □

Now we prove a converse of Theorem 3.

THEOREM 4. *Given an arbitrary homogeneous system H, there exists a sequence S such that*

$$h_1(S) \subset H, \quad d(S) = d(H).$$

(3.13) LEMMA. *Let H be a homogeneous system. For arbitrary x there exists a set $A \in H$, $A \subset [1, x]$ satisfying*

(3.14) $$|A \cap [1, y]| \geqq y d(H) \quad \textit{for all} \quad 1 \leqq y \leqq x.$$

PROOF. Choose a large z and a set $B \subset [1, z]$, $B \in H$ such that

$$|B| = H(z) \geqq z d(H).$$

Let

$$f(j) = |B \cap [1, j]| - \frac{jz}{x+z} d(H) \quad (j \in \mathbf{Z}).$$

Let k be the place of minimum of f in $[0, z]$. We have

(3.15) $$f(k+y) \geqq f(k) \qquad (1 \leqq y \leqq x),$$

since for $k+y \leqq z$ this follows from the definition of k, and for $z+1 \leqq k+y \leqq z+x$ we have

$$f(k+y) = |B| - \frac{(k+y)z}{x+z} d(H) \geqq 0$$

and

$$f(k) \leqq f(0) = 0.$$

(3.15) means

$$0 \leqq f(k+y) - f(k) = |B \cap [k+1, k+y]| - \frac{yz}{z+x} d(H),$$

that is, $A_z=(B-k)\cap[1, x]$ satisfies

$$|A_z\cap[1, y]| \geqq y\frac{z}{z+x}d(H) \qquad (1\leqq y\leqq x). \tag{3.16}$$

Since for $A_z\subset[1, x]$ there are only finitely many possibilities, there must be a set A which occurs infinitely many times as an A_z. This A must satisfy (3.16) for arbitrary large values of z, which just means (3.14). □

Proof of Theorem 4. Let C_x be the set of A's satisfying (3.14). We define an ordered tree. The points at the k'th level are the elements of C_k; $A\in C_k$ and $B\in C_{k+1}$ are joined by an edge if $A\subset B$. Each level is finite, and from each point there is an edge downwards, for $B\in C_{k+1}$ is connected with

$$A = B\cap[1, k]\in C_k.$$

A well-known theorem of D. Kőnig implies that there is a chain $(A_k)_{k=1}^{\infty}$ such that $A_k\in C_k$ and $A_k\subset A_{k+1}$ for all k. The sequence

$$S=\cup A_k$$

will fulfill our requirements. Namely $h_1(S)\subset H$ is obvious from the construction. Moreover

$$S(x)\geqq|A_x|\geqq d(H)x\Rightarrow \underline{d}(S)\geqq d(H);$$

on the other hand

$$\bar{d}(S)\leqq d^*(S)=d(h_1(S))\leqq d(H)$$

by Theorem 3, and these two inequalities yield

$$d(S)=d(H). \quad □$$

Evidently, the intersection of homogeneous systems is also a homogeneous system. The main advantage of homogeneous systems over sequences is the following property.

Theorem 5. *For arbitrary homogeneous systems $H_1, \dots, H_k$ we have*

$$d(H_1\cap\ldots\cap H_k)\geqq\prod d(H_j).$$

Proof. Evidently, it is sufficient to prove the statement for $k=2$. Let $H=H_1\cap H_2$ and choose sets

$$A\subset[1, x],\quad A\in H_1,\quad |A|=H_1(x) \tag{3.17}$$

and

$$B\subset[1, y],\quad B\in H_2,\quad |B|=H_2(y). \tag{3.18}$$

For

$$X_j=A\cap(B-j)$$

we have

$$X_j\in H,\quad X_j\subset[1, x],$$

hence

$$|X_j|\leqq H(x). \tag{3.19}$$

$|X_j|$ is the number of solutions of $a=b-j$, $a\in A$, $b\in B$, so we have

$$\sum |X_j| = |A|\,|B|. \tag{3.20}$$

Moreover, $X_j=\emptyset$ unless $1-x\leqq j\leqq y-1$, therefore (3.20) contains at most $x+y-1$ summands, different from zero. This together with (3.17—20) yields

$$H_1(x)\,H_2(y) = |A|\,|B| = \sum |X_j| \leqq (x+y-1)\,H(x),$$

that is

$$H(x) \geqq H_1(x)\frac{H_2(y)}{x+y-1}. \tag{3.21}$$

Making $y\to\infty$ we get

$$H(x) \geqq d(H_2)\,H_1(x);$$

dividing by x and making $x\to\infty$ this gives the desired inequality

$$d(H) \geqq d(H_1)\,d(H_2). \quad \square$$

4. The difference set of homogeneous systems

THEOREM 6. *Let H be a homogeneous system satisfying $d(H)>0$ and let $D=\Delta H$. There are integers $b_1, \dots, b_k$, $k\leqq 1/d(H)$ such that*

$$\bigcup (D+b_j) = \mathbf{Z}. \tag{4.1}$$

PROOF. Let $b_1, \dots, b_k$ be a maximal set of integers such that the sets $A+b_1, \dots, A+b_k$ are disjoint for every $A\in H$. Choosing a set $A\in H$, $A\subset[1, x]$, $|A|=H(x)$ and denoting

$$b = \max |b_j|$$

we get

$$kH(x) \leqq x+2b, \tag{4.2}$$

since all the sets $A+b_j$ lie within the interval $[1-b, x+b]$. Dividing by x and making $x\to\infty$ (4.2) yields $k\leqq 1/d(H)$ as wanted.

On the other hand, let $n\in\mathbf{Z}$ be arbitrary; we show that $n\in D+b_j$ for some j. Since $b_1, \dots, b_k$ was a maximal set, there must be an $A\in H$ and a j such that

$$(A+b_j)\cap(A+n) \neq \emptyset,$$

that is,

$$a+b_j = a'+n \qquad (a, a'\in A).$$

Hence

$$n = (a-a')+b_j\in D+b_j. \quad \square$$

Now we prove the results of Section 2.

PROOF of Theorem 1. Let

$$H_i = h_2(S_i), \quad H = \bigcap H_i,$$

$$H_i' = h_3(S_i), \quad H' = \bigcap H_i'.$$

We have

$$d(H) \geqq \prod d(H_i) = \prod d^*(S_i)$$

and

$$d(H') \geqq \prod d(H_i') \geqq \prod \bar{d}(S_i)$$

by Theorems 5 and 3.

Let S, resp. S' be sequences satisfying

$$h_1(S) \subset H, \quad d(S) = d(H),$$

$$h_1(S') \subset H', \quad d(S') = d(H');$$

their existence is guaranteed by Theorem 4. We have

$$\Delta_1(S) = \Delta(h_1(S)) \subset \Delta(H) = \Delta(\bigcap H_j) \subset \bigcap(\Delta(H_j)) = \bigcap(\Delta_2(S_j))$$

and similarly

$$\Delta_1(S') \subset \bigcap(\Delta_3(S_j)),$$

as required. □

PROOF of Theorem 2. Apply Theorem 6 to $H=h_2(S)$ in Case a), and to $H=h_3(S)$ in Case b). □

REFERENCES

[1] RUZSA, I. Z., On difference-sequences, *Acta Arithmetica* **25** (1974), 151—157.
[2] STEWART, C. L. and TIJDEMAN, R., On infinite difference sets, *Canad. J. Math.* **31** (1979), 897—910.

Mathematical Institute of the Hungarian Academy of Sciences
Reáltanoda u. 13—15, H—1053 Budapest

(Received May 29, 1978)

FUNCTORIAL PROPERTIES OF THE MIXED LIMIT FUNCTORS

by

C. G. CHEHATA and I. A. ASSEM

Introduction

The mixed limit functors were introduced in [1]. Some of their functorial properties were studied in [1] and [2]. The present work is a continuation of these two papers. We start by summarizing briefly the results of [1], from which we deduce directly that the mixed limit functors have no adjoints. It is known that projective and inductive limits do not change by restriction of a directed index set to a cofinal subset. We show that the same result holds for mixed limits. We then show that any equivalence functor commutes with the mixed limit functors. The results of the last section are a generalization of the results of [2].

1. Preliminaries

Definition 1.1. A category $\mathscr{C}$ will be called *I-complete (I-cocomplete)* if every projective (inductive) system in $\mathscr{C}$ over the pre-ordered set I has a limit. It is simply called *complete (cocomplete)* if every projective (inductive) system in $\mathscr{C}$ over any index set has a limit.

Definition 1.2. Let I and L be pre-ordered sets. A *mixed system* $(E_\alpha^\lambda, f_{\alpha\beta}^{\mu\lambda})_{I\times L}$ in a category $\mathscr{C}$ over $I\times L$ is defined by the following:

— To every pair $(\alpha, \lambda)\in I\times L$ corresponds an object E_α^λ of $\mathscr{C}$.

— To every couple of pairs (α, λ) and (β, μ) of $I\times L$ with $\alpha\leqq\beta$ and $\lambda\leqq\mu$ corresponds a morphism $f_{\alpha\beta}^{\mu\lambda}\colon E_\beta^\lambda\to E_\alpha^\mu$ such that:

(MS1) For every $(\alpha, \lambda)\in I\times L$, we have

$$f_{\alpha\alpha}^{\lambda\lambda} = 1_{E_\alpha^\lambda}.$$

(MS2) If $\alpha\leqq\beta\leqq\gamma$ in I, and $\lambda\leqq\mu\leqq\nu$ in L,

$$f_{\alpha\gamma}^{\nu\lambda} = f_{\alpha\beta}^{\nu\mu} f_{\beta\gamma}^{\mu\lambda}.$$

To simplify notations, we shall denote the morphisms $f_{\alpha\alpha}^{\mu\lambda}$ by $h_\alpha^{\mu\lambda}$ and the morphisms $f_{\alpha\beta}^{\lambda\lambda}$ by $g_{\alpha\beta}^\lambda$.

A mixed system $(E_\alpha^\lambda, f_{\alpha\beta}^{\mu\lambda})_{I\times L}$ over an I-complete and L-cocomplete category $\mathscr{C}$ has two limits which we construct as follows: the system $(E_\alpha^\lambda, h_\alpha^{\mu\lambda})_L$ is inductive, with limit

$$(E_\alpha, h_\alpha^\lambda) = \varinjlim_{\lambda\in L} (E_\alpha^\lambda, h_\alpha^{\mu\lambda}),$$

while $(E_\alpha^\lambda, g_{\alpha\beta}^\lambda)_I$ is projective with limit

$$(E^\lambda, g_\alpha^\lambda) = \varprojlim_{\alpha\in I} (E_\alpha^\lambda, g_{\alpha\beta}^\lambda).$$

The family of morphisms $(g_{\alpha\beta}^\lambda)_{\lambda\in L}$ $(\alpha\leqq\beta)$ forms an inductive system of morphisms from $(E_\beta^\lambda, h_\beta^{\mu\lambda})_L$ into $(E_\alpha^\lambda, h_\alpha^{\mu\lambda})_L$. Let:

$$g_{\alpha\beta} = \varinjlim_{\lambda\in L} g_{\alpha\beta}^\lambda.$$

Similarly, $(h_\alpha^{\mu\lambda})_{\alpha\in I}$ $(\lambda\leqq\mu)$ forms a projective system of morphisms from $(E_\alpha^\lambda, g_{\alpha\beta}^\lambda)_I$ into $(E_\alpha^\mu, g_{\alpha\beta}^\mu)_I$, with limit

$$h^{\mu\lambda} = \varprojlim_{\alpha\in I} h_\alpha^{\mu\lambda}.$$

Finally, the systems $(E_\alpha, g_{\alpha\beta})_I$ and $(E^\lambda, h^{\mu\lambda})_L$ are respectively projective and inductive. We shall denote their limits by

$$(E, g_\alpha) = \varprojlim_{\alpha\in I} (E_\alpha, g_{\alpha\beta}),$$

$$(F, h^\lambda) = \varinjlim_{\lambda\in L} (E^\lambda, h^{\mu\lambda}).$$

It was shown in [1] that there exists a unique morphism $f\colon F\to E$ such that

$$(\forall(\alpha,\lambda)\in I\times L)\colon\ g_\alpha f h^\lambda = h_\alpha^\lambda g_\alpha^\lambda.$$

E and F are called the limits of the system and f is the canonical morphism.

Definition 1.3. Let $\mathscr{E}=(E_\alpha^\lambda, f_{\alpha\beta}^{\mu\lambda})_{I\times L}$ be a mixed system in the I-complete and L-cocomplete category $\mathscr{C}$, and let

$$(E_\alpha, h_\alpha^\lambda) = \varinjlim_{\lambda\in L} (E_\alpha^\lambda, f_{\alpha\alpha}^{\mu\lambda}), \quad g_{\alpha\beta} = \varinjlim_{\lambda\in L} f_{\alpha\beta}^{\lambda\lambda},$$

$$(E^\lambda, g_\alpha^\lambda) = \varprojlim_{\alpha\in I} (E_\alpha^\lambda, f_{\alpha\beta}^{\lambda\lambda}), \quad h^{\mu\lambda} = \varprojlim_{\alpha\in I} f_{\alpha\alpha}^{\mu\lambda}.$$

A triple (E, F, f) when E and F are objects of $\mathscr{C}$ and $f\colon F\to E$ a morphism, will be called a *limit triple of* $\mathscr{E}$ if

(ML1) There exist two families of morphisms $(g_\alpha\colon E\to E_\alpha)_{\alpha\in I}$ and $(h^\lambda\colon E^\lambda\to F)_{\lambda\in L}$ such that

(i) $g_\alpha = g_{\alpha\beta} g_\beta$, if $\alpha\leqq\beta$ in I.

(ii) $h^\lambda = h^\mu h^{\mu\lambda}$, if $\lambda\leqq\mu$ in L.

(iii) $g_\alpha f h^\lambda = h_\alpha^\lambda g_\alpha^\lambda$, for any $(\alpha,\lambda)\in I\times L$.

(ML2) If (E', F', f') is another triple, with E', F' objects in $\mathscr{C}$ and $f'\colon F'\to E'$, and there exist families $(g'_\alpha\colon E'\to E_\alpha)_{\alpha\in I}$ and $(h'^\lambda\colon E^\lambda\to F')_{\lambda\in L}$ of morphisms of $\mathscr{C}$ which satisfy:

(i) $g'_\alpha = g_{\alpha\beta} g'_\beta$, if $\alpha\leqq\beta$ in I.

(ii) $h'^\lambda = h'^\mu h^{\mu\lambda}$, if $\lambda\leqq\mu$ in L.

(iii) $g'_\alpha f' h'^\lambda = h_\alpha^\lambda g_\alpha^\lambda$, for any $(\alpha,\lambda)\in I\times L$.

Then there exists a unique pair (g, h) of morphisms of $\mathscr{C}$ such that:

$$f = gf'h.$$

$$\begin{tikzcd}
E & F \arrow[l, "f"'] \arrow[d, "h"] \\
E' \arrow[u, "g"] & F' \arrow[l, "f'"]
\end{tikzcd}$$

Figure (1.1)

It is proved (cf. [1]) that the triple (E, F, f), where E and F are the limits as defined above and f is the canonical morphism, is the only (up to isomorphism) limit triple of $(E_\alpha^\lambda, f_{\alpha\beta}^{\mu\lambda})_{I\times L}$.

DEFINITION 1.4. Let $\mathscr{E}=(E_\alpha^\lambda, f_{\alpha\beta}^{\mu\lambda})_{I\times L}$ and $\mathscr{E}'=(E_\alpha'^\lambda, f_{\alpha\beta}'^{\mu\lambda})_{I\times L}$ be mixed systems in $\mathscr{C}$ over $I\times L$. A family $u=(u_\alpha^\lambda)_{I\times L}$ of morphisms of $\mathscr{C}$ will be called a *mixed system of morphisms of $\mathscr{E}$ into $\mathscr{E}'$* if $\alpha\leqq\beta$ in I and $\lambda\leqq\mu$ in L imply

$$u_\alpha^\mu f_{\alpha\beta}^{\mu\lambda} = f_{\alpha\beta}'^{\mu\lambda} u_\beta^\lambda.$$

$$\begin{tikzcd}
E_\beta^\lambda \arrow[r, "f_{\alpha\beta}^{\mu\lambda}"] \arrow[d, "u_\beta^\lambda"'] & E_\alpha^\mu \arrow[d, "u_\alpha^\mu"] \\
E_\beta'^\lambda \arrow[r, "f_{\alpha\beta}'^{\mu\lambda}"] & E_\alpha'^\mu
\end{tikzcd}$$

Figure (1.2)

DEFINITION 1.5. The category $\mathfrak{M}(I\times L, \mathscr{C})$ of mixed systems in $\mathscr{C}$ over $I\times L$ has for objects all such systems, and for morphisms, the mixed systems of morphisms. The product of two morphisms $u=(u_\alpha^\lambda)_{I\times L}\colon \mathscr{E}\to\mathscr{E}'$ and $v=(v_\alpha^\lambda)_{I\times L}\colon \mathscr{E}'\to\mathscr{E}''$ is defined to be $vu=(v_\alpha^\lambda u_\alpha^\lambda)_{I\times L}$.

It was shown in [1] that a mixed system of morphisms defines two limit morphisms

$$u_- = \underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}\,\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}\, u_\alpha^\lambda\colon E\to E',$$

$$u_+ = \underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}\,\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}\, u_\alpha^\lambda\colon F\to F'.$$

This allows us to define two covariant functors of $\mathfrak{M}(I\times L, \mathscr{C})$ into $\mathscr{C}$ as follows:

(i) $l_+(-)=\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}\,\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}\,(-)$ associates to a mixed system $\mathscr{E}$ its limit $F=\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}\,\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}\, E_\alpha^\lambda$ and to a morphism $u\colon \mathscr{E}\to\mathscr{E}'$ its limit $u_+=\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}\,\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}\, u_\alpha^\lambda\colon F\to F'$.

(ii) $l_-(-)=\varprojlim_{\alpha\in I}\varinjlim_{\lambda\in L}(-)$ associates to $\mathscr{E}$ its limit $E=\varprojlim_{\alpha\in I}\varinjlim_{\lambda\in L}E_\alpha^\lambda$ and to u its limit $u_-=\varprojlim_{\alpha\in I}\varinjlim_{\lambda\in L}u_\alpha^\lambda\colon E\to E'$.

Moreover the canonical morphism $f\colon l_+(\mathscr{E})\to l_-(\mathscr{E})$ is a functorial morphism (cf. [1]).

It is known (cf. [4]) that projective and inductive limits do not in general commute. This implies:

Proposition 1.1. *The functors l_+ and l_- have no adjoint functors.*

For if this were the case, l_+ and l_- would commute with products and sums, which are particular projective and inductive limits.

2. Cofinal subsets of I and L

Theorem 2.1. *Let I and L be directed sets, and $J\subseteq I$, $K\subseteq L$ be cofinal subsets. Let $(E_\alpha^\lambda, f_{\alpha\beta}^{\mu\lambda})_{I\times L}$ be a mixed system in $\mathscr{C}$ over $I\times L$, (E, F, f) a limit triple with associated morphisms $(g_\alpha\colon E\to E_\alpha)_{\alpha\in I}$ and $(h^\lambda\colon E^\lambda\to F)_{\lambda\in L}$. Then $(E_\alpha^\lambda, f_{\alpha\beta}^{\mu\lambda})_{J\times K}$ is a mixed system in $\mathscr{C}$ over $J\times K$, of which a limit triple is (E, F, f) with associated morphisms $(g_\alpha)_{\alpha\in J}$ and $(h^\lambda)_{\lambda\in K}$.*

Proof. This result can be proved using the corresponding results in the case of projective and inductive limits, we shall prove it directly using Definition (1.3).

It is obvious that $(E_\alpha^\lambda, f_{\alpha\beta}^{\mu\lambda})_{J\times K}$ is a mixed system in $\mathscr{C}$ over $J\times K$ and that the triple (E, F, f) with the morphisms $(g_\alpha)_{\alpha\in J}$ and $(h^\lambda)_{\lambda\in K}$ satisfies (ML1). We let (E', F', f') be another triple in $\mathscr{C}$, with morphisms $(g'_\alpha)_{\alpha\in J}$ and $(h'^\lambda)_{\lambda\in K}$ satisfying the conditions of (ML2).

Let $\alpha\in I$, there exists a $\beta\in J$ such that $\beta\geqq\alpha$. The directedness of I implies that the morphism $g_{\alpha\beta}g'_\beta\colon E'\to E_\alpha$ is independent of the index β. We shall set $g'_\alpha=g_{\alpha\beta}g'_\beta$. In this way, we have defined a family of morphisms $(g'_\alpha)_{\alpha\in I}$ and it is easily seen that

$$g'_\alpha = g_{\alpha\beta}\,g'_\beta \quad\text{whenever}\quad \alpha\leqq\beta \quad\text{in } I.$$

Similarly, let $\lambda\in L$, there exists a $\mu\in K$ with $\mu\geqq\lambda$, and the morphism $h'^\mu h^{\mu\lambda}$ is independent of μ. We shall denote it by h'^λ. The family $(h'^\lambda)_{\lambda\in L}$ thus defined is clearly seen to satisfy

$$h'^\lambda = h'^\mu h^{\mu\lambda} \quad\text{whenever}\quad \lambda\leqq\mu \quad\text{in } L.$$

Finally, let $(\alpha,\lambda)\in I\times L$. There exists $(\beta,\mu)\in J\times K$ such that

$$g'_\alpha = g_{\alpha\beta}\,g'_\beta,\quad h'^\lambda = h'^\mu h^{\mu\lambda}.$$

Then:

$$\begin{aligned} g'_\alpha f' h'^\lambda &= g_{\alpha\beta}\,g'_\beta f' h'^\mu h^{\mu\lambda} =\\ &= g_{\alpha\beta}\,h_\beta^\mu\, g_\beta^\mu\, h^{\mu\lambda} =\\ &= h_\alpha^\mu\, g_{\alpha\beta}^\mu\, g_\beta^\mu\, h^{\mu\lambda} =\\ &= h_\alpha^\mu\, g_\alpha^\mu\, h^{\mu\lambda} =\\ &= h_\alpha^\mu\, h_\alpha^{\mu\lambda}\, g_\alpha^\lambda =\\ &= h_\alpha^\lambda\, g_\alpha^\lambda. \end{aligned}$$

Therefore there exist unique morphisms $g\colon E'\to E$ and $h\colon F\to F'$ such that

$$f = gf'h.$$

This completes the proof.

COROLLARY 2.1. *If either I or L (or both) has a maximal element, and the other index set is directed, the mixed limits are isomorphic.*

PROOF. Assume I has a maximal element α_0, the mixed system $(E_\alpha^\lambda, f_{\alpha\beta}^{\mu\lambda})_{I\times L}$ has the same limit triple as $(E_{\alpha_0}^\lambda, f_{\alpha_0\alpha_0}^{\mu\lambda})_{\{\alpha_0\}\times L}$ by Theorem (2.1). Now the last system is an inductive system, a limit triple of which is $(E_{\alpha_0}, E_{\alpha_0}, 1_{E_{\alpha_0}})$. Hence the result.

3. Equivalence functors and mixed limits

We recall the following:

DEFINITION 3.1. Let $\mathscr{C}$ and $\mathscr{D}$ be categories, $G\colon \mathscr{C}\to\mathscr{D}$ a covariant functor. G is called an *equivalence* if one of the following equivalent conditions is satisfied:

1) There exists a functor $H\colon \mathscr{D}\to\mathscr{C}$ and functorial isomorphisms from HG into $1_{\mathscr{C}}$ and from GH into $1_{\mathscr{D}}$.

2) G is full and faithful, and for any object D of $\mathscr{D}$, there exists an object C of $\mathscr{C}$ such that $G(C)$ is isomorphic to D.

THEOREM 3.1. *Let $\mathscr{C}$ and $\mathscr{D}$ be I-complete and L-cocomplete categories, and $G\colon\mathscr{C}\to\mathscr{D}$ a full and faithful functor. In order that (E, F, f) with associated morphisms $(g_\alpha\colon E\to E_\alpha)_{\alpha\in I}$ and $(h^\lambda\colon E^\lambda\to F)_{\lambda\in L}$ be a limit triple of the mixed system $(E_\alpha^\lambda, f_{\alpha\beta}^{\mu\lambda})_{I\times L}$ in $\mathscr{C}$, it is sufficient that $\big(G(E), G(F), G(f)\big)$ with the morphisms $\big(G(g_\alpha)\big)_{\alpha\in I}$ and $\big(G(h^\lambda)\big)_{\lambda\in L}$ be a limit triple of the mixed system $\big(G(E_\alpha^\lambda), G(f_{\alpha\beta}^{\mu\lambda})\big)_{I\times L}$.*

PROOF. The equalities

$$G(g_\alpha) = G(g_{\alpha\beta})G(g_\beta), \quad \text{if} \quad \alpha \leqq \beta \text{ in } I,$$

$$G(h^\lambda) = G(h^\mu)G(h^{\mu\lambda}), \quad \text{if} \quad \lambda \leqq \mu \text{ in } L,$$

and $G(g_\alpha)G(f)G(h^\lambda) = G(h_\alpha^\lambda)G(g_\alpha^\lambda)$, for any $(\alpha,\lambda)\in I\times L$ imply respectively

$$g_\alpha = g_{\alpha\beta}\, g_\beta, \quad \text{if} \quad \alpha \leqq \beta \text{ in } I,$$

$$h^\lambda = h^\mu h^{\mu\lambda}, \quad \text{if} \quad \lambda \leqq \mu \text{ in } L$$

and

$$g_\alpha f h^\lambda = h_\alpha^\lambda g_\alpha^\lambda, \quad \text{for any} \quad (\alpha,\lambda)\in I\times L$$

since G is faithful. Thus (E, F, f) satisfies (ML1).

Let now (E', F', f') be another triple with associated families of morphisms $(g'_\alpha\colon E'\to E_\alpha)_{\alpha\in I}$ and $(h'^\lambda\colon E^\lambda\to F')_{\lambda\in L}$ which satisfies the conditions of (ML2). Then we have:

$$G(g'_\alpha) = G(g_{\alpha\beta})G(g'_\beta), \quad \text{if} \quad \alpha \leqq \beta \text{ in } I,$$

$$G(h'^\lambda) = G(h'^\mu)G(h^{\mu\lambda}), \quad \text{if} \quad \lambda \leqq \mu \text{ in } L,$$

and

$$G(g'_\alpha)G(f')G(h'^\lambda)=G(h^\lambda_\alpha)G(g^\lambda_\alpha),\quad \text{for any}\quad (\alpha,\lambda)\in I\times L.$$

Then, by hypothesis, there exist unique morphisms g_0: $G(E')\to G(E)$ and h_0: $G(F)\to G(F')$ such that

$$G(f)=g_0G(f')h_0.$$

Since G is full and faithful, this implies the existence of unique morphisms $g\colon E'\to E$ and h: $F\to F'$ such that

$$g_0=G(g),\quad h_0=G(h).$$

Then $G(f)=G(g)G(f')G(h)$ implies, by the faithfulness of G

$$f=gf'h.$$

THEOREM 3.2. *If* G: $\mathscr{C}\to\mathscr{D}$ *is an equivalence, then* (E, F, f) *with associated morphism* $(g_\alpha)_{\alpha\in I}$ *and* $(h^\lambda)_{\lambda\in L}$ *is a limit triple of the system* $(E^\lambda_\alpha, f^{\mu\lambda}_{\alpha\beta})_{I\times L}$ *if and only if* $(G(E), G(F), G(f))$ *with morphisms* $G(g_\alpha)_{\alpha\in I}$ *and* $G(h^\lambda)_{\lambda\in L}$ *is a limit triple of* $(G(E^\lambda_\alpha), G(f^{\mu\lambda}_{\alpha\beta}))_{I\times L}$.

PROOF. It remains to show that if $G\colon \mathscr{C}\to\mathscr{D}$ is an equivalence, and (E, F, f) is a limit triple of $(E^\lambda_\alpha, f^{\mu\lambda}_{\alpha\beta})_{I\times L}$, then $(G(E^\lambda_\alpha), G(f^{\mu\lambda}_{\alpha\beta}))_{I\times L}$ has $(G(E), G(F), G(f))$ as a limit triple.

It is obvious that (ML1) is satisfied. Assume that $(\bar{E}', \bar{F}', \bar{f}')$ is a triple of $\mathscr{D}$ and $(\bar{g}'\colon \bar{E}'\to G(E_\alpha))_{\alpha\in I}$ and $(\bar{h}'\colon G(E^\lambda)\to\bar{F}')_{\lambda\in L}$ are such that

$$\bar{g}'_\alpha=G(g_{\alpha\beta})\bar{g}'_\beta,\quad \text{if}\quad \alpha\leqq\beta \text{ in } I,$$

$$\bar{h}'^\lambda=\bar{h}'^\mu G(h^{\mu\lambda}),\quad \text{if}\quad \lambda\leqq\mu \text{ in } L,$$

and

$$\bar{g}'_\alpha\bar{f}'\bar{h}'^\lambda=G(h^\lambda_\alpha)G(g^\lambda_\alpha),\quad \text{for any}\quad (\alpha,\lambda)\in I\times L.$$

Since G is an equivalence, there exist objects E' and F' in $\mathscr{C}$ and isomorphisms j_1: $G(E')\to\bar{E}'$ and j_2: $G(F')\to\bar{F}'$. Let us write:

$$\hat{g}'_\alpha=\bar{g}'_\alpha j_1,$$

$$\hat{f}'=j_1^{-1}\bar{f}'j_2,$$

$$\hat{h}'^\lambda=j_2^{-1}\bar{h}'^\lambda.$$

Obviously, we have

$$\hat{g}'_\alpha=G(g_{\alpha\beta})\hat{g}'_\beta,\quad \text{if}\quad \alpha\leqq\beta \text{ in } I,$$

$$\hat{h}'^\lambda=\hat{h}'^\mu G(h^{\mu\lambda}),\quad \text{if}\quad \lambda\leqq\mu \text{ in } L,$$

and

$$\hat{g}'_\alpha\hat{f}'\hat{h}'^\lambda=G(h^\lambda_\alpha)G(g^\lambda_\alpha),\quad \text{for any}\quad (\alpha,\lambda)\in I\times L.$$

$$\begin{tikzcd}
& G(E) \arrow[dl, "G(g_\alpha)"'] & G(F) \arrow[l, "G(f)"'] & \\
G(E_\alpha) & \bar{E}' \arrow[l, "\bar{g}'_\alpha"'] & \bar{F}' \arrow[l, "\bar{f}'"'] & G(E^\lambda) \arrow[l, "\bar{h}'^\lambda"'] \arrow[ul, "G(h^\lambda)"'] \arrow[dl, "\hat{h}'^\lambda"] \\
& G(E') \arrow[ul, "\hat{g}'_\alpha"] \arrow[u, "j_1"] & G(F') \arrow[l, "\hat{f}'"] \arrow[u, "j_2"'] &
\end{tikzcd}$$

Figure (3.1)

Since G is full and faithful, there exist unique morphisms $f'\colon F' \to E'$, $(g'_\alpha\colon E' \to E_\alpha)_{\alpha \in I}$ and $(h'^\lambda\colon E^\lambda \to F')_{\lambda \in L}$ such that

$$G(f') = \hat{f}',$$

$$G(g'_\alpha) = \hat{g}'_\alpha, \quad (\forall \alpha \in I),$$

and

$$G(h'^\lambda) = \hat{h}'^\lambda, \quad (\forall \lambda \in L).$$

Again the faithfulness of G implies that (E', F', f') satisfies the conditions of (ML2), then there exist unique morphisms $g\colon E' \to E$ and $h\colon F \to F'$ such that

$$\begin{tikzcd}
G(E) & & & G(F) \arrow[lll, "G(f)"'] \arrow[dl, dashed, "\bar{h}"'] \arrow[dd, "G(h)"] \\
& \bar{E}' \arrow[ul, dashed, "\bar{g}"'] & \bar{F}' \arrow[l, "\bar{f}'"'] & \\
G(E') \arrow[uu, "G(g)"] \arrow[ur, "j_1"'] & & & G(F') \arrow[lll, "\hat{f}'"] \arrow[ul, "j_2"]
\end{tikzcd}$$

Figure (3.2)

Then

$$G(f) = G(g)\hat{f}'G(h) = G(g)j_1^{-1}\bar{f}'j_2G(h).$$

There only remains to show that $\bar{g} = G(g)j_1^{-1}$ and $\bar{h} = j_2G(h)$ are unique. Let $(\bar{g}_0, \bar{h}_0)$ be a pair of morphisms such that

$$G(f) = \bar{g}_0\bar{f}'\bar{h}_0.$$

Again, since G is full and faithful, there exist morphisms $g_0\colon E' \to E$ and $h_0\colon F \to F'$ such that

$$G(g_0) = \bar{g}_0 j_1, \quad G(h_0) = j_2^{-1}\bar{h}_0.$$

But then

$$\begin{aligned} G(f) &= \bar{g}_0 \bar{f}' \bar{h}_0 = \\ &= G(g_0) j_1^{-1} \bar{f}' j_2 G(h_0) = \\ &= G(g_0) \hat{f}' G(h_0) = \\ &= G(g_0) G(f') G(h_0) \end{aligned}$$

and therefore

$$f = g_0 f' h_0.$$

The uniqueness of the pair (g, h) implies that $g=g_0$ and $h=h_0$. Then $\bar{g}=\bar{g}_0$ and $\bar{h}=\bar{h}_0$.

We have the obvious corollaries

COROLLARY 3.1. *If $\mathscr{C}$ is I-complete and L-cocomplete, so is every equivalent category $\mathscr{D}$.*

COROLLARY 3.2. *Every equivalence functor G commutes with the mixed limit functors*

$$Gl_+ = l_+G, \quad Gl_- = l_-G.$$

This can also be proved directly: indeed it is known that a functor which admits a left adjoint (right adjoint) commutes with projective (inductive) limits. Now an equivalence has both a left and a right adjoint (cf. [4]). Hence the result.

4. Commutation of functors with l_+ and l_-

As proofs in this section follow the lines of the proofs of [2], they will be omitted. We start with some remarks on terminology.

Let $(E_\alpha, g_{\alpha\beta})_I$ be a projective system in an I-complete category $\mathscr{C}$ and let $(E, g_\alpha) = \underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}\,(E_\alpha, g_{\alpha\beta})$.

Let now $\mathscr{D}$ be another I-complete category and $G: \mathscr{C} \to \mathscr{D}$ be a covariant functor. Then $\big(G(E_\alpha), G(g_{\alpha\beta})\big)_I$ is also a projective system. Let $\big(k_\alpha: \underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}\, G(E_\alpha) \to G(E_\alpha)\big)_{\alpha\in I}$ denote its canonical projections. By definition of projective limits, there exists a unique $s: G(E) \to \underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}\, G(E_\alpha)$ such that, for all $\alpha\in I$,

$$k_\alpha s = G(g_\alpha).$$

It is easy to see that s is functorial.

$$\begin{tikzcd} G(E_\alpha) & \\ \underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}\, G(E_\alpha) \arrow[u, "k_\alpha"] & G(E) \arrow[l, "s"'] \arrow[ul, "G(g_\alpha)"'] \end{tikzcd}$$

Figure (4.1)

DEFINITION 4.1. The functor F will be said to *commute with* $\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}$ if s is an isomorphism.

LEMMA 4.1. *The family* $\big(G(g_\alpha)\big)_{\alpha\in I}$ *is a monomorphic family.*

If now G is contravariant and $\mathscr{D}$ is I-cocomplete $\big(G(E_\alpha), G(g_{\alpha\beta})\big)_I$ is an inductive system.

Let $\big(k^\alpha\colon G(E_\alpha)\to \underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}\, G(I_\alpha)\big)_{\alpha\in I}$ denote its canonical morphisms. There exists a unique morphism

$$s\colon \underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}\, G(E_\alpha)\to G(E)$$

such that, for all $\alpha\in I$,

$$sk^\alpha = G(g_\alpha).$$

Again, s is a functorial morphism.

$$\begin{tikzcd}
G(E_\alpha) \arrow[d, "k^\alpha"'] \arrow[dr, "G(g_\alpha)"] & \\
\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}\, G(E_\alpha) \arrow[r, "s"] & G(E)
\end{tikzcd}$$

Figure (4.2)

DEFINITION 4.2. The functor G will be said to *transform* $\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}$ into $\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}$ if s is an isomorphism.

LEMMA 4.2. *The family* $\big((g_\alpha)\big)_{\alpha\in I}$ *is epimorphic.*

Dually, let $(E^\lambda, h^{\mu\lambda})_L$ be an inductive system in an L-cocomplete category $\mathscr{C}$, of limit (E, h^λ). Let G be a covariant functor of $\mathscr{C}$ into another L-cocomplete category $\mathscr{D}$. Then $\big(G(E^\lambda), G(h^{\mu\lambda})\big)_L$ is also an inductive system.

Let $\big(l^\lambda\colon G(E^\lambda)\to \underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}\, G(E^\lambda)\big)_{\lambda\in L}$ be its canonical morphisms. Then there exists a unique $r\colon \underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}\, G(E^\lambda)\to G(E)$ such that, for all $\lambda\in L$,

$$rl^\lambda = G(h^\lambda).$$

$$\begin{tikzcd}
G(E^\lambda) \arrow[d, "l^\lambda"'] \arrow[dr, "G(h^\lambda)"] & \\
\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}\, G(E^\lambda) \arrow[r, "r"] & G(E)
\end{tikzcd}$$

Figure (4.3)

Definition 4.3. The functor G will be said to *commute with* $\varinjlim_{\lambda\in L}$ if r is an isomorphism.

Lemma 4.3. *The family* $(G(h^{\lambda}))_{\lambda\in L}$ *is epimorphic.*

Finally, if G is contravariant and $\mathscr{D}$ is L-complete, $(G(E^{\lambda}), G(h^{\mu\lambda}))_L$ is now a projective system and there exists a unique $r\colon G(E)\to\varprojlim_{\lambda\in L} G(E^{\lambda})$ such that, for all $\lambda\in L$,

$$l_{\lambda} r = G(h^{\lambda}),$$

where $(l_{\lambda}\colon \varprojlim_{\lambda\in L} G(E^{\lambda})\to G(E^{\lambda}))_{\lambda\in L}$ denote the canonical projections.

$$\begin{tikzcd}
G(E^{\lambda}) & \\
\varprojlim\limits_{\lambda\in L} G(E^{\lambda}) \arrow[u, "l_{\lambda}"] & G(E) \arrow[l, "r"'] \arrow[ul, "G(h^{\lambda})"']
\end{tikzcd}$$

Figure (4.4)

Definition 4.5. The functor G will be said to *transform* $\varinjlim_{\lambda\in L}$ into $\varprojlim_{\lambda\in L}$ if r is an isomorphism.

Lemma 4.4. *The family* $(G(h^{\lambda}))_{\lambda\in L}$ *is monomorphic.*

Examples. (i) Let $\mathscr{C}$ be any I-complete category then, given a fixed object G of $\mathscr{C}$, the covariant functor $\mathrm{Hom}_{\mathscr{C}}(G, -)$ into the category of sets and maps commutes with $\varprojlim_{\alpha\in I}$, while the contravariant functor $\mathrm{Hom}_{\mathscr{C}}(-, G)$ transforms $\varinjlim_{\lambda\in L}$ into $\varprojlim_{\lambda\in L}$ (cf. [5]).

(ii) Let $\mathscr{C}$ be the category of compact pairs, then the Čech homology functor commutes with $\varprojlim_{\alpha\in I}$ while the Čech cohomology functor transforms $\varprojlim_{\alpha\in I}$ into $\varinjlim_{\alpha\in I}$. These are the so-called continuity theorems (cf. [6]).

(iii) Let G be an arbitrary right (left) R-module the functor $G\underset{R}{\otimes}-$ of the category ${}_R\mathrm{Mod}$ of left R-modules ($-\underset{R}{\otimes} G$ of the category Mod_R of right R-modules) into the category of abelian groups commutes with $\varinjlim_{\lambda\in L}$ (cf. (5)). Also let G be an arbitrary abelian group, then the endofunctors $\mathrm{Tor}(G, -)$ and $\mathrm{Tor}(-, G)$ of the category of abelian groups commute with $\varinjlim_{\lambda\in L}$ in case L is directed (cf. [3]).

Theorem 4.1. *Let* $\mathscr{C}$ *and* $\mathscr{D}$ *be* I*-complete and* L*-cocomplete categories and* $G\colon\mathscr{C}\to\mathscr{D}$ *a covariant functor which commutes with* $\varprojlim_{\alpha\in I}$. *Then we have a commutative diagram* (4.5) *of covariant functors and of functorial morphisms of* $\mathfrak{M}(I\times L, \mathscr{C})$ *into* $\mathscr{D}$.

$$\begin{tikzcd}
Gl_+ \arrow[d, "G(f)"'] & & l_+G \arrow[ll, "q"'] \arrow[d, "g"] \\
Gl_- & & l_-G \arrow[ll, "p"']
\end{tikzcd}$$

Figure (4.5)

THEOREM 4.2. *Let $\mathscr{C}$ and $\mathscr{D}$ be I-complete and L-cocomplete categories and $G\colon \mathscr{C}\to\mathscr{D}$ a covariant functor which commutes with $\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}$. Then we have a commutative diagram* (4.6) *of covariant functors and of functorial morphisms of $\mathfrak{M}(I\times L,\mathscr{C})$ into $\mathscr{D}$.*

$$\begin{tikzcd}
Gl_+ \arrow[rr, "q"] \arrow[d, "G(f)"'] & & l_+G \arrow[d, "g"] \\
Gl_- \arrow[rr, "p"] & & l_-G
\end{tikzcd}$$

Figure (4.6)

THEOREM 4.3. *Let $\mathscr{C}$ be an I-complete and L-cocomplete category, $\mathscr{D}$ be an I-cocomplete and L-complete category, and $G\colon \mathscr{C}\to\mathscr{D}$ a contravariant functor which transforms $\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}$ into $\underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}$. Then we have a commutative diagram* (4.7) *of contravariant functors and of functorial morphisms of $\mathfrak{M}(I\times L,\mathscr{C})$ into $\mathscr{D}$*

$$\begin{tikzcd}
l^+G \arrow[rr, "q"] \arrow[d, "g"'] & & Gl_- \arrow[d, "G(f)"] \\
l^-G \arrow[rr, "p"] & & Gl_+
\end{tikzcd}$$

Figure (4.7)

where $l^+=\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}\,\underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}$ *and* $l^-=\underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}\,\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}$.

THEOREM 4.4. *Let $\mathscr{C}$ be an I-complete and L-cocomplete category, $\mathscr{D}$ an I-complete and L-cocomplete category, and $G\colon \mathscr{C}\to\mathscr{D}$ a contravariant functor which transforms $\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}$ into $\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}$. Then we have a commutative diagram* (4.8) *of contravariant functors and of functorial morphisms of $\mathfrak{M}(I\times L,\mathscr{C})$ into $\mathscr{D}$.*

$$\begin{tikzcd}
l^{+}G \arrow[d, "g"'] & & Gl_{-} \arrow[ll, "q"'] \arrow[d, "G(f)"] \\
l^{-}G & & Gl_{+} \arrow[ll, "p"']
\end{tikzcd}$$

Figure (4.8)

We now turn to the case of bifunctors:

THEOREM 4.5. *Let $\mathscr{C}_1$ be an I-complete category, $\mathscr{C}_2$ an L-cocomplete category, and F a bifunctor of $\mathscr{C}_1\times\mathscr{C}_2$ into the I-complete and L-cocomplete category $\mathscr{C}$ which is covariant in both variables. Then:*

(i) *If F commutes (in the first variable) with* $\varprojlim_{\alpha\in I}$, *we have a commutative diagram* (4.9) *of bifunctors of $\mathscr{C}_1\times\mathscr{C}_2$ into $\mathscr{C}$ and of functorial morphisms.*

$$\begin{tikzcd}
l_{-}F(-,-) \arrow[dr, "u"'] & & l_{+}F(-,-) \arrow[ll, "f"'] \arrow[dl, "v"] \\
& F\left(\varprojlim_{\alpha\in I}(-), \varinjlim_{\lambda\in L}(-)\right) &
\end{tikzcd}$$

Figure (4.9)

(ii) *If F commutes (in the second variable) with* $\varinjlim_{\lambda\in L}$, *we have a commutative diagram* (4.10) *of bifunctors of $\mathscr{C}_1\times\mathscr{C}_2$ into $\mathscr{C}$ and of functorial morphisms:*

$$\begin{tikzcd}
l_{-}F(-,-) & & l_{+}F(-,-) \arrow[ll, "f"'] \\
& F\left(\varprojlim_{\alpha\in I}(-), \varinjlim_{\lambda\in L}(-)\right) \arrow[ul, "u'"] \arrow[ur, "v'"'] &
\end{tikzcd}$$

Figure (4.10)

COROLLARY 4.1. *Under the above conditions, if F commutes in the first variable with* $\varprojlim_{\alpha\in I}$ *and in the second variable with* $\varinjlim_{\lambda\in L}$, *then f is a functorial isomorphism*

$$l_{-}F \cong l_{+}F.$$

Dually,

THEOREM 4.6. *Let $\mathscr{C}_1$ be an I-cocomplete category, $\mathscr{C}_2$ an L-complete category and F a bifunctor of $\mathscr{C}_1\times\mathscr{C}_2$ into the I-complete and L-cocomplete category $\mathscr{C}$ which is contravariant in both variables, then:*

(i) *If F (in the first variable) transforms* $\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}$ *into* $\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}$, *we have a commutation diagram* (4.11) *of bifunctors of* $\mathscr{C}_1\times\mathscr{C}_2$ *into* $\mathscr{C}$ *and of functorial morphisms.*

$$\begin{tikzcd}
l_-F(-,-) \arrow[dr, "u"'] & & l_+F(-,-) \arrow[ll, "f"'] \arrow[dl, "v"] \\
& F\left(\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}(-), \underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}(-)\right) &
\end{tikzcd}$$

Figure (4.11)

(ii) *If F (in the second variable) transforms* $\underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}$ *into* $\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}$, *then we have a commutation diagram* (4.12) *of bifunctors of* $\mathscr{C}_1\times\mathscr{C}_2$ *into* $\mathscr{C}$ *and of functorial morphisms.*

$$\begin{tikzcd}
l_-F(-,-) & & l_+F(-,-) \arrow[ll, "f"'] \\
& F\left(\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}(-), \underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}(-)\right) \arrow[ul, "u'"] \arrow[ur, "v'"'] &
\end{tikzcd}$$

Figure (4.12)

COROLLARY 4.2. *Under the above conditions, if F transforms (in the first variable)* $\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}$ *into* $\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}$ *and (in the second variable)* $\underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}$ *into* $\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}$, *then f is a functorial isomorphism*

$$l_-F\cong l_+F.$$

THEOREM 4.7. *Let* $\mathscr{C}_1$ *be an I-cocomplete,* $\mathscr{C}_2$ *an L-complete category, and F a bifunctor of* $\mathscr{C}_1\times\mathscr{C}_2$ *into the I-complete and L-cocomplete category* $\mathscr{C}$ *which is contravariant in the first and covariant in the second variable, then:*

(i) *If F (in the first variable) transforms* $\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}$ *into* $\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}$, *we have a commutative diagram* (4.13) *of bifunctors of* $\mathscr{C}_1\times\mathscr{C}_2$ *into* $\mathscr{C}$ *and of functorial morphisms.*

$$\begin{tikzcd}
l_-F(-,-) \arrow[dr, "u"'] & & l_+F(-,-) \arrow[ll, "f"'] \arrow[dl, "v"] \\
& F\left(\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}(-), \underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}(-)\right) &
\end{tikzcd}$$

Figure (4.13)

(ii) *If F commutes (in the second variable) with* $\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}$, *we have a commutative diagram* (4.14) *of bifunctors of* $\mathscr{C}_1\times\mathscr{C}_2$ *into* $\mathscr{C}$ *and of functorial morphisms.*

$$\begin{tikzcd}
l_-F(-,-) & & l_+F(-,-) \arrow[ll, "f"'] \\
& F\left(\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}(-), \underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}(-)\right) \arrow[ul, "u'"] \arrow[ur, "v'"'] &
\end{tikzcd}$$

Figure (4.14)

COROLLARY 4.3. *Under the above conditions, if F transforms (in the first variable)* $\underset{\alpha\in I}{\underrightarrow{\mathrm{Lim}}}$ *into* $\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}$, *and (in the second variable) commutes with* $\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}$, *then f is a functorial isomorphism:*

$$l_-F\cong l_+F.$$

THEOREM 4.8. *Let $\mathscr{C}_1$ be an I-complete, $\mathscr{C}_2$ an L-cocomplete category, and F a bifunctor of $\mathscr{C}_1\times\mathscr{C}_2$ into the I-complete and L-cocomplete category $\mathscr{C}$ which is contravariant in the first and covariant in the second variable, then:*

(i) *If F commutes (in the second variable) with* $\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}$, *we have a commutative diagram* (4.15) *of bifunctors of* $\mathscr{C}_1\times\mathscr{C}_2$ *into* $\mathscr{C}$ *and of functorial morphisms:*

$$\begin{tikzcd}
l_-F(-,-) \arrow[dr, "u"'] & & l_+F(-,-) \arrow[ll, "f"'] \arrow[dl, "v"] \\
& F\left(\underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}(-), \underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}(-)\right) &
\end{tikzcd}$$

Figure (4.15)

(ii) *If F transforms (in the first variable)* $\underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}$ *into* $\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}$, *we have a commutative diagram* (4.16) *of bifunctors of* $\mathscr{C}_1\times\mathscr{C}_2$ *into* $\mathscr{C}$ *and of functorial morphisms:*

$$\begin{tikzcd}
l_-F(-,-) & & l_+F(-,-) \arrow[ll, "f"'] \\
& F\left(\underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}(-), \underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}(-)\right) \arrow[ul, "u'"] \arrow[ur, "v'"'] &
\end{tikzcd}$$

Figure (4.16)

Corollary 4.4. *Under the above conditions, if F transforms (in the first variable) $\underset{\lambda\in L}{\underleftarrow{\mathrm{Lim}}}$ into $\underset{\lambda\in L}{\underrightarrow{\mathrm{Lim}}}$ and commutes (in the second variable) with $\underset{\alpha\in I}{\underleftarrow{\mathrm{Lim}}}$, then f is a functorial isomorphism*

$$l_- F \cong l_+ F.$$

REFERENCES

[1] Chehata, C. G., El-Gendy, M. A. and Assem, I. A., Limits of bifunctors into a category (to appear in *Acta Math. Acad. Sci. Hungar*).

[2] Chehata, C. G., El-Gendy, M. A. and Assem, I. A., The Hom and tensor functor with mixed systems (to appear).

[3] Fuchs, L., *Infinite abelian groups,* Academic Press, 1970.

[4] Jaffard, P. et Poitou, G., *Introduction aux catégories et aux problèmes universels,* Ediscience, Paris, 1971.

[5] Lafon, J. P., *Les formalismes fondamentaux de l'algèbre commutative,* Hermann, Paris, 1974.

[6] Wallace, A. H., *Algebraic topology,* Benjamin, 1970.

Faculty of Science, University of Alexandria,
Alexandria, Egypt

Department of Mathematics, Carleton University,
Ottawa, Canada

(Received June 2, 1978)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 343—352.

О НЕКОТОРЫХ ОЦЕНКАХ ДЛЯ РАЗНОСТНЫХ ОПЕРАТОРОВ, АППРОКСИМИРУЮЩИХ ДИФФЕРЕНЦИАЛЬНЫЕ ОПЕРАТОРЫ ЭЛЛИПТИЧЕСКОГО ТИПА В n-МЕРНОМ ПРОСТРАНСТВЕ

NGUYEN TUONG

В работах [3], [4] при доказательстве разностного аналога теоремы вложения 2, ограничились частными случаями когда число измерений пространства равно 2 и 3 ($n=2, 3$), получили некоторые оценки, которые в свою очередь послужили основной для доказательства устойчивости и сходимости разностных схем для уравнений эллиптического типа с двумя или тремя независимыми переменными.

Возникает вопрос: Имеет ли место разностный аналог теоремы вложения в общем случае когда $n>3$? Эта данная статья отвечает на этот вопрос. Она состоит из двух частей.

В первой части рассматривается задача на собственные значения и собственные функции разностного оператора Лапласа в n-мерном пространстве.

В второй части будем получить разностный аналог теоремы вложения в общем соболевском пространстве $\overset{\circ}{W}{}_2^m$ [2] и некоторые другие оценки.

В этой статьи используем обозначения, введённые в книге А. А. Самарского [1].

I. Задача на собственные значения и собственные функции разностного оператора Лапласа

Задача на собственные значения и собственные функции разностного оператора Лапласа.

Пусть на n-мерном пространстве задана область:

$$\bar{G} = \{x = (x_1, x_2, \ldots, x_n);\ 0 \leqq x_\alpha \leqq l_\alpha;\ \alpha = 1, 2, \ldots, n\}$$

с границей Γ и оператор Лапласа:

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2}.$$

Введем в области $\bar{G}$ разностную сетку таким образом, чтобы гиперплоскости, образующие границу Γ, принадлежали классу гиперплоскостей, образующих сетку:

$$\bar{\omega}_h = \{x = (x_1, x_2, \ldots, x_n) \in \bar{G};\ x_\alpha = i_\alpha h_\alpha;\ i_\alpha = 0, 1, 2, \ldots, N_\alpha;\ \alpha = 1, 2, \ldots, n\}.$$

Обозначим: $\bar{\omega}_h = \omega_h + \gamma_h$, где γ_h: множество граничных узлов сетки, ω_h: множество внутренних узлов сетки.

Рассмотрим простейщую аппроксимацию оператора Лапласа:

$$\mathring{\Lambda} v = v_{\bar{x}_1 x_1} + v_{\bar{x}_2 x_2} + \ldots + v_{\bar{x}_n x_n},$$

где

$$v_{\bar{x}_\alpha x_\alpha} = \frac{1}{h_\alpha^2} [v(x_1, \ldots, x_\alpha + h_\alpha, \ldots, x_n) -$$

$$- 2v(x_1, \ldots, x_\alpha, \ldots, x_n) + v(x_1, \ldots, x_\alpha - h_\alpha, \ldots, x_n)].$$

Легко видеть, что он аппроксимирует Δv с погрешностью $O(|h|^2)$, $|h|^2 = \sum_{\alpha=1}^{n} h_\alpha^2$.

Решим следующую задачу на собственные значения:

$$\mathring{\Lambda} v + \lambda v = 0, \quad x \in \omega_h; \quad v(x) = 0, \quad x \in \gamma_h. \tag{1}$$

Будем искать нетривиальные, т. е. не равные нулю тождественно, решения уравнения (1) вида:

$$v(x_1, x_2, \ldots, x_n) = X^{(1)}(x_1) X^{(2)}(x_2) \ldots X^{(n)}(x_n) \tag{2}$$

удовлетворяющие граничным условиям. Мы здесь считаем, что $X^{(\alpha)}(x_\alpha)$ зависит только от x_α.

Подставив правую часть (2) вместо v в уравнение (1), получаем

$$(X^{(2)} \ldots X^{(n)}) X^{(1)}_{\bar{x}_1 x_1} + (X^{(1)} \ldots X^{(n)}) X^{(2)}_{\bar{x}_2 x_2} + \ldots + (X^{(1)} \ldots X^{(n-1)}) X^{(n)}_{\bar{x}_n x_n} + \lambda X^{(1)} \ldots X^{(n)} = 0,$$

так как мы ищем нетривиальные решения задачи (1), то можно разделить обе части этого уравнения на $X^{(1)} \ldots X^{(n)}$. В результате получим:

$$\frac{X^{(1)}_{\bar{x}_1 x_1}}{X^{(1)}} + \frac{X^{(2)}_{\bar{x}_2 x_2}}{X^{(2)}} + \ldots + \frac{X^{(n)}_{\bar{x}_n x_n}}{X^{(n)}} + \lambda = 0,$$

или

$$\frac{X^{(1)}_{\bar{x}_1 x_1}}{X^{(1)}} = -\left(\frac{X^{(2)}_{\bar{x}_2 x_2}}{X^{(2)}} + \ldots + \frac{X^{(n)}_{\bar{x}_n x_n}}{X^{(n)}} + \lambda \right) = -\lambda^{(1)}$$

причем $\lambda^{(1)}$ не зависит ни от x_1, ни от x_α $(\alpha = 2, 3, \ldots, n)$.

Тем самым для $X^{(1)}$ получаем простейщую задачу на собственные значения:

$$X^{(1)}_{\bar{x}_1 x_1} + \lambda^{(1)} X^{(1)} = 0, \quad x_1 = h_1, \ldots, l_1 - h_1; \; X^{(1)}_0 = X^{(1)}_{N_1} = 0.$$

Легко видеть, что решением этой задачи является [см. 1]:

$$X^{(1)}_{K_1}(x_1) = \sqrt{\frac{2}{l_1}} \sin \frac{K_1 \pi x_1}{l_1},$$

и

$$\lambda^{(1)}_{K_1} = \frac{4}{h_1^2} \sin^2 \frac{K_1 \pi h_1}{2 l_1}, \quad K_1 = 1, 2, \ldots, N_1 - 1.$$

Аналогичную задачу получаем для $X^{(2)}$:

$$\frac{X^{(2)}_{\bar{x}_2x_2}}{X^{(2)}} = -\left(\frac{X^{(3)}_{\bar{x}_3x_3}}{X^{(3)}}+\ldots+\frac{X^{(n)}_{\bar{x}_nx_n}}{X^{(n)}}+\lambda-\lambda^{(1)}\right) = -\lambda^{(2)},$$

причем $\lambda^{(2)}$ не зависит ни от x_2, ни от x_α ($\alpha=1, 3, 4, \ldots, n$). Тем самым для $X^{(2)}$ получаем выше аналогичную задачу:

$$X^{(2)}_{\bar{x}_2x_2}+\lambda^{(2)}X^{(2)} = 0, \quad x_2 = h_2, \ldots, l_2-h_2; \; X^{(2)}_0 = X^{(2)}_{N_2} = 0,$$

и её решением является:

$$X^{(2)}_{K_2}(x_2) = \sqrt{\frac{2}{l_2}}\sin\frac{K_2\pi x_2}{l_2},$$

и

$$\lambda^{(2)}_{K_2} = \frac{4}{h_2^2}\sin^2\frac{K_2\pi h_2}{2l_2}, \quad K_2 = 1, 2, \ldots, N_2-1.$$

Продолжая это процесс, после $(n-1)$-го шага перейдем к уравнению для $X^{(n)}$:

$$\frac{X^{(n)}_{\bar{x}_nx_n}}{X^{(n)}} = -(\lambda-\lambda^{(1)}-\lambda^{(2)}-\ldots-\lambda^{(n-1)}) = -\lambda^{(n)}.$$

Обозначим: $\lambda-\lambda^{(1)}-\lambda^{(2)}-\ldots-\lambda^{(n-1)}=\lambda^{(n)}$, или

$$\lambda = \lambda^{(1)}+\lambda^{(2)}+\ldots+\lambda^{(n)} \tag{3}$$

причем $\lambda^{(n)}$ не зависит от x_α ($\alpha=1, 2, \ldots, n$). Получаем задачу на собственные значения для $X^{(n)}$:

$$X^{(n)}_{\bar{x}_nx_n}+\lambda^{(n)}X^{(n)} = 0, \quad x_n = h_n, \ldots, l_n-h_n; \; X^{(n)}_0 = X^{(n)}_{N_n} = 0,$$

и решением является:

$$X^{(n)}_{K_n}(x_n) = \sqrt{\frac{2}{l_n}}\sin\frac{K_n\pi x_n}{l_n},$$

и

$$\lambda^{(n)}_{K_n} = \frac{4}{h_n^2}\sin^2\frac{K_n\pi h_n}{2l_n}, \quad K_n = 1, 2, \ldots, N_n-1.$$

Наконец в силу (2) и (3) получаем собственные функции и собственные значения задачи (1):

$$v = v_{K_1\ldots K_n}(x_1, \ldots, x_n) = \frac{2^{n/2}}{\sqrt{l_1\ldots l_n}}\sin\frac{K_1\pi x_1}{l_1}\ldots\sin\frac{K_n\pi x_n}{l_n}, \tag{4}$$

и

$$\lambda = \lambda_{K_1\ldots K_n} = 4\left(\frac{1}{h_1^2}\sin^2\frac{K_1\pi h_1}{2l_1}+\ldots+\frac{1}{h_n^2}\sin^2\frac{K_n\pi h_n}{2l_n}\right) \tag{5}$$

где $K_\alpha=1, 2, \ldots, N_\alpha-1$; $\alpha=1, 2, \ldots, n$.

Перечислим их свойства:

1. Собственные значения $\lambda_{K_1 \dots K_n}$ перенумерованы в порядке возрастания и для всей совокупности $\{\lambda_{K_1 \dots K_n}\}$ справедливы следующие оценки:

$$\delta_0 = 8\left(\frac{1}{l_1^2}+\dots+\frac{1}{l_n^2}\right) \leqq \lambda_{1\dots 1} \leqq \dots \leqq \lambda_{N_1-1\dots N_n-1} \leqq 4\left(\frac{1}{h_1^2}+\dots+\frac{1}{h_n^2}\right) = \Delta_0.$$

2. Собственные функции $v_{K_1 \dots K_n}$ образуют ортонормированную систему в смысле следующего скалярного произведения:

$$(v, w) = \sum_{i_1=1}^{N_1-1} \dots \sum_{i_n=1}^{N_n-1} v(i_1 h_1, \dots, i_n h_n) w(i_1 h_1, \dots, i_n h_n) h_1 \dots h_n; \quad \|v\|^2 = (v, v).$$

3. Первые разностные производные от собственных функций, имеющие вид (например по x_1):

$$\bar{v} = (v_{K_1 \dots K_n})_{\bar{x}_1} = \sqrt{\frac{2\lambda_{K_1}^{(1)}}{l_1}} X^{(2)} \dots X^{(n)} \cos \frac{K_1 \pi (x_1 - 0{,}5h_1)}{l_1},$$

ортогональны в смысле скалярного произведения (,], т. е.:

$$(\bar{v}, \bar{w}]_1 = \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2-1} \dots \sum_{i_n=1}^{N_n-1} \bar{v}(i_1 h_1, \dots, i_n h_n) \bar{w}(i_1 h_1, \dots, i_n h_n) h_1 \dots h_n,$$

и кроме того

$$|(\bar{v}]|^2 = \lambda_{K_1}^{(1)}.$$

4. Пусть на сетке $\bar{\omega}_h$ задана сеточная функция $f(x_1, \dots, x_n)$, причем $f(x)|_{\gamma_h} = 0$. Тогда, очевидно, она представима в виде суммы по собственным функциям задачи (1):

$$f(x) = \sum_{K_1=1}^{N_1-1} \dots \sum_{K_n=1}^{N_n-1} c_{K_1 \dots K_n} v_{K_1 \dots K_n}(x).$$

При этом справедливо равенство:

$$\|f(x)\|^2 = \sum_{K\ =1}^{N_1-1} \dots \sum_{K_n=1}^{N_n-1} c^2_{K_1 \dots K_n}.$$

II. Априорные оценки

В этой части получено три теоремы, которые являются хорошим основанием для доказательства равномерной сходимости разностных схем для уравнения эллиптического типа.

Теорема 1. *Для всякой сеточной функции $v(x)$, заданной на сетке $\bar{\omega}_h$ и обращающейся в нуль на границе γ_h $(v|_{\gamma_h} = 0)$, если $m > \frac{n}{4}$ (m: целое число, n: число измерений пространства), тогда имеет место разностный аналог теоремы вложения:*

$$\|v\|_C \leqq M \|\mathring{A}^m v\|,$$

где

$$\|v\|_C = \max_{x \in \omega_h} |v(x)|, \quad \mathring{A} = -\Lambda, \ \mathring{A}^m = \underbrace{\mathring{A}\mathring{A}\ldots\mathring{A}}_{m \text{ раз}},$$

M—постоянная не зависищая от функции $v(x)$*:*

$$M = \frac{l_0^{2m}}{2^{2m-n/2}\sqrt{l_1 \ldots l_n}} \sqrt{\frac{1}{n^{2m}} + \frac{S_1}{4m-n}}, \quad l_0 = \max\{l_1, \ldots, l_n\},$$

$S_1 = \dfrac{2\pi^{\frac{n}{2}}}{\Gamma\left(\dfrac{n}{2}\right)}$*—площадь поверхности единичной сферы.*

Доказательство. Прежде всего разложим $v(x)$ по системе собственных функций $\{v_{K_1 \ldots K_n}(x)\}$:

$$v(x) = \sum_{K_1=1}^{N_1-1} \cdots \sum_{K_n=1}^{N_n-1} c_{K_1 \ldots K_n} v_{K_1 \ldots K_n}(x),$$

отсюда получаем

$$\mathring{A}v = \sum_{K_1 \ldots K_n} c_{K_1 \ldots K_n} \lambda_{K_1 \ldots K_n} v_{K_1 \ldots K_n}(x), \quad (\text{так как } \mathring{A}v_{K_1 \ldots K_n} = \lambda_{K_1 \ldots K_n} v_{K_1 \ldots K_n}(x)).$$

Умножая это равенство ещё раз на разностный оператор $\mathring{A}$ получаем:

$$\mathring{A}^2 v = \sum_{K_1 \ldots K_n} c_{K_1 \ldots K_n} \lambda^2_{K_1 \ldots K_n} v_{K_1 \ldots K_n}(x),$$

или после m умножений, получим:

$$\mathring{A}^m v = \sum_{K_1 \ldots K_n} c_{K_1 \ldots K_n} \lambda^m_{K_1 \ldots K_n} v_{K_1 \ldots K_n}(x),$$

и следовательно

$$\|\mathring{A}^m v\|^2 = \sum_{K_1 \ldots K_n} c^2_{K_1 \ldots K_n} \lambda^{2m}_{K_1 \ldots K_n}.$$

Оценим теперь функцию $v(x)$ следующим образом:

$$|v(x)| \leqq \Big(\sum_{K_1 \ldots K_n} |c_{K_1 \ldots K_n}|\Big) \max_{K_1 \ldots K_n} |v_{K_1 \ldots K_n}(x)|.$$

В силу (4) получаем

$$|v_{K_1 \ldots K_n}(x)| \leqq \frac{2^{n/2}}{\sqrt{l_1 \ldots l_n}},$$

так что

$$|v(x)| \leqq \frac{2^{n/2}}{\sqrt{l_1 \ldots l_n}} \sum_{K_1 \ldots K_n} |c_{K_1 \ldots K_n}|,$$

$$|v(x)| \leqq \frac{2^{n/2}}{\sqrt{l_1 \ldots l_n}} \sum_{K_1 \ldots K_n} |(c_{K_1 \ldots K_n} \lambda^m_{K_1 \ldots K_n})(\lambda^{-m}_{K_1 \ldots K_n})|.$$

Используя неравенство Буняковского для сумм, получаем

$$|v(x)| \leqq \frac{2^{n/2}}{\sqrt{l_1 \dots l_n}} \Big(\sum_{K_1 \dots K_n} c^2_{K_1 \dots K_n} \lambda^{2m}_{K_1 \dots K_n}\Big)^{1/2} \Big(\sum_{K_1 \dots K_n} \lambda^{-2m}_{K_1 \dots K_n}\Big)^{1/2},$$

или

$$|v(x)| \leqq \frac{2^{n/2}}{\sqrt{l_1 \dots l_n}} \|\mathring{A}^m v\| \Big(\sum_{K_1 \dots K_n} \lambda^{-2m}_{K_1 \dots K_n}\Big)^{1/2}. \tag{6}$$

Для завершения доказательства нам нужно оценить:

$$\sum_{K_1 \dots K_n} (\lambda^{-2m}_{K_1 \dots K_n}).$$

Обращаясь к формуле (5) и учитывая, что $\sin x \geqq \frac{2x}{\pi}$ при $0 \leqq x \leqq \frac{\pi}{2}$, получаем:

$$\lambda_{K_1 \dots K_n} \geqq 4\left(\frac{K_1^2}{l_1^2} + \dots + \frac{K_n^2}{l_n^2}\right) \geqq \frac{4}{l_0^2}(K_1^2 + \dots + K_n^2),$$

где $l_0 = \max\{l_1, \dots, l_n\}$. Следовательно:

$$\sum_{K_1 \dots K_n} \lambda^{-2m}_{K_1 \dots K_n} \leqq \frac{l_0^{4m}}{2^{4m}} \sum_{K_1=1}^{N_1-1} \dots \sum_{K_n=1}^{N_n-1} \frac{1}{(K_1^2 + \dots + K_n^2)^{2m}}. \tag{7}$$

Имеем оценку

$$\sum_{K_1=1}^{N_1-1} \dots \sum_{K_n=1}^{N_n-1} \frac{1}{(K_1^2 + \dots + K_n^2)^{2m}} \leqq \frac{1}{(1^2 + \dots + 1^2)^{2m}} + \overbrace{\int \dots \int}^{n}_{r \geqq 1} \frac{d\tau}{r^{4m}}, \tag{8}$$

где $r^2 = K_1^2 + \dots + K_n^2$.

Известно, что последний несобственный интеграл будет сходиться если $4m > n$ или $m > \frac{n}{4}$. Но это удовлетворяется в силу условия теоремы.

Переходя в последнем интеграле к сферическим координатам n-мерного пространства, получаем

$$\overbrace{\int \dots \int}^{n}_{r \geqq 1} \frac{d\tau}{r^{4m}} = \int_1^\infty \frac{dr}{r^{4m+1-n}} \int_{S_1} d_1 s,$$

где

$$d\tau = r^{n-1} d_1 s \, dr,$$

$$d_1 s = \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \dots \sin\theta_{n-2} \, d\theta_1 \, d\theta_2 \dots d\theta_{n-2} \, d\varphi,$$

$$0 \leqq \theta_k \leqq \pi, \quad 0 \leqq \varphi \leqq 2\pi,$$

S_1 — поверхность единичной сферы. В результате, получим:

$$\overbrace{\int \dots \int}^{n}_{r \geqq 1} \frac{d\tau}{r^{4m}} = \frac{S_1}{4m - n}, \tag{9}$$

где $S_1=\dfrac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}$ — площадь поверхности единичной сферы. $\Gamma\left(\frac{n}{2}\right)$ — гамма функция.

Наконец, из полученных неравенств (6), (7), (8) и равенства (9) получим:

$$|v| \leqq \frac{2^{n/2}}{\sqrt{l_1 \ldots l_n}} \frac{l_0^{2m}}{2^{2m}} \sqrt{\frac{1}{n^{2m}}+\frac{S_1}{4m-n}}\, \|\mathring{A}^m v\|,$$

или

$$\|v\|_C \leqq \frac{l_0^{2m}}{2^{2m-n/2}\sqrt{l_1 \ldots l_n}} \sqrt{\frac{1}{n^{2m}}+\frac{S_1}{4m-n}}\, \|\mathring{A}^m v\|. \tag{10}$$

Обозначим

$$M=\frac{l_0^{2m}}{2^{2m-n/2}\sqrt{l_1 \ldots l_m}} \sqrt{\frac{1}{n^{2m}}+\frac{S_1}{4m-n}}\,.$$

Это завершает доказательство теоремы.

Замечание. Из общей формулы (10) в частных случаях получим следующие оценки, некоторые из них принадлежат авторам [3], [4].

1. При $n=1$ и $m=1$ $(S_1=2)$, получим:

$$\|v\|_C \leqq \frac{l^{3/2}}{2}\, \|\mathring{A} v\|.$$

2. При $n=2$ и $m=1$ $(S_1=2\pi)$, получим:

$$\|v\|_C \leqq \frac{l_0^2}{\sqrt{l_1 l_2}}\, \|\mathring{A} v\|.$$

3. При $n=3$ и $m=1$ $(S_1=4\pi)$, получим:

$$\|v\|_C \leqq \frac{3 l_0^2}{\sqrt{l_1 l_2 l_3}}\, \|\mathring{A} v\|.$$

4. При $n=2$ и $m=2$ $(S_1=2\pi)$, получим:

$$\|v\|_C \leqq \frac{l_0^4}{4\sqrt{2 l_1 l_2}}\, \|\mathring{A}^2 v\|$$

где $\mathring{A}^2$ — разностный оператор бигармонического оператора Δ^2:

$$\Delta^2 v=\frac{\partial^4 v}{\partial x_1^4}+2\frac{\partial^2 v}{\partial x_1^2 \partial x_2^2}+\frac{\partial^4 v}{\partial x_2^4},$$

$$\mathring{A}^2 v=v_{\bar{x}_1 x_1 \bar{x}_1 x_1}+2 v_{\bar{x}_1 x_1 \bar{x}_2 x_2}+v_{\bar{x}_2 x_2 \bar{x}_2 x_2}.$$

Теорема 2. *Для всякой сеточной функции $v(x)$, заданной на сетке $\overline{\omega}_h$ и обращающейся в нуль на границе γ_h $(v|_{\gamma_h}=0)$, имеют место неравенства:*

$$
(11)\qquad
\begin{aligned}
&\frac{1}{\delta_0^{m-2}}\|\mathring{A}^{m-1}v\|^2 \leqq (\mathring{A}^m v, v) \leqq \frac{1}{\Delta_0^{m-2}}\|\mathring{A}^{m-1}v\|^2 \quad \text{при} \quad m \leqq 2,\\
&\frac{1}{\Delta_0^{m-2}}\|\mathring{A}^{m-1}v\|^2 \leqq (\mathring{A}^m v, v) \leqq \frac{1}{\delta_0^{m-2}}\|\mathring{A}^{m-1}v\|^2 \quad \text{при} \quad m \geqq 3,
\end{aligned}
$$

где
$$\delta_0 = 8\left(\frac{1}{l_1^2}+\ldots+\frac{1}{l_n^2}\right), \quad \Delta_0 = 4\left(\frac{1}{h_1^2}+\ldots+\frac{1}{h_n^2}\right),$$

m: целое положительное число и положено $\mathring{A}^0=1$.

Доказательство. Разложим теперь функцию $v(x)$ по системе собственных функций $\{v_{K_1\ldots K_n}(x)\}$:

$$v(x) = \sum_{K_1=1}^{N_1-1}\cdots\sum_{K_n=1}^{N_n-1} c_{K_1\ldots K_n} v_{K_1\ldots K_n}(x); \quad \|v\|^2 = \sum_{K_1\ldots K_n} c_{K_1\ldots K_n}^2.$$

Отсюда находим

$$\mathring{A}^{m-1}v = \sum_{K_1\ldots K_n} c_{K_1\ldots K_n}\lambda_{K_1\ldots K_n}^{m-1} v_{K_1\ldots K_n}(x), \quad \|\mathring{A}^{m-1}v\|^2 = \sum_{K_1\ldots K_n} c_{K_1\ldots K_n}^2 \lambda_{K_1\ldots K_n}^{2(m-1)},$$

и

$$\mathring{A}^m v = \sum_{K_1\ldots K_n} c_{K_1\ldots K_n}\lambda_{K_1\ldots K_n}^{m} v_{K_1\ldots K_n}(x).$$

Вычисляя скалярное произведение $(\mathring{A}^m v, v)$, получаем

$$(\mathring{A}^m v, v) = \sum_{K_1\ldots K_n} c_{K_1\ldots K_n}^2 \lambda_{K_1\ldots K_n}^{m},$$

или

$$(\mathring{A}^m v, v) = \sum_{K_1\ldots K_n} (c_{K_1\ldots K_n}^2 \lambda_{K_1\ldots K_n}^{2(m-1)} \lambda_{K_1\ldots K_n}^{-(m-2)}).$$

Нам осталось оценить сумму, стоящую в правой части. Для этой цели из-под знака суммы вынесем максимальное и минимальное значение третьего множителя:

$$\min_{K_1\ldots K_n} \lambda_{K_1\ldots K_n}^{-(m-2)}\|\mathring{A}^{m-1}v\|^2 \leqq (\mathring{A}^m v, v) \leqq \max_{K_1\ldots K_n} \lambda^{-(m-2)}\|\mathring{A}^{m-1}v\|^2.$$

Отсюда следует, что при $m \leqq 2$:

$$\big(\min_{K_1\ldots K_n} \lambda_{K_1\ldots K_n}\big)^{-(m-2)}\|\mathring{A}^{m-1}v\|^2 \leqq (\mathring{A}^m v, v) \leqq \big(\max_{K_1\ldots K_n} \lambda_{K_1\cdots K_n}\big)^{-(m-2)}\|\mathring{A}^{m-1}v\|^2,$$

и при $m \geqq 3$:

$$\big(\max_{K_1\ldots K_n} \lambda_{K_1\ldots K_n}\big)^{-(m-2)}\|\mathring{A}^{m-1}v\|^2 \leqq (\mathring{A}^m v, v) \leqq \big(\min_{K_1\ldots K_n} \lambda_{K_1\ldots K_n}\big)^{-(m-2)}\|\mathring{A}^{m-1}v\|^2.$$

Значения верхней и нижней границ собственных чисел известны (1-ое свойство собственных значений):

$$\min_{K_1\ldots K_n} \lambda_{K_1\ldots K_n} \geqq \delta_0 = 8\left(\frac{1}{l_1^2}+\ldots+\frac{1}{l_n^2}\right),$$

$$\max_{K_1\ldots K_n} \lambda_{K_1\ldots K_n} \leqq \varDelta_0 = 4\left(\frac{1}{h_1^2}+\ldots+\frac{1}{h_n^2}\right).$$

Следовательно

$$\begin{aligned} &\frac{1}{\delta_0^{m-2}}\|\mathring{A}^{m-1}v\|^2 \leqq (\mathring{A}^m v, v) \leqq \frac{1}{\varDelta_0^{m-2}}\|\mathring{A}^{m-1}v\|^2 \quad \text{при} \quad m \leqq 2, \\ &\frac{1}{\varDelta_0^{m-2}}\|\mathring{A}^{m-1}v\|^2 \leqq (\mathring{A}^m v, v) \leqq \frac{1}{\delta_0^{m-2}}\|\mathring{A}^{m-1}v\|^2 \quad \text{при} \quad m \geqq 3. \end{aligned} \tag{11}$$

Теорема 2 доказана.

Замечание. В частных случаях из общей формулы (11) получим следующие оценки (см. [1]).

1. При $m=1$ получим оценку:

$$\delta_0\|v\|^2 \leqq (\mathring{A}v, v) \leqq \varDelta_0\|v\|^2.$$

По определению, эти неравенства показывают, что разностный оператор $\mathring{A}$ — положительно определен.

2. При $m=1$ и $n=1$ получим:

$$\frac{8}{l^2}\|v\|^2 \leqq \|v_{\bar{x}}]|^2 \leqq \frac{4}{h^2}\|v\|^2,$$

или

$$\frac{h^2}{4}\|v_{\bar{x}}]|^2 \leqq \|v\|^2 \leqq \frac{l^2}{8}\|v_{\bar{x}}]|^2.$$

3. При $m=2$ получим:

$$\|\mathring{A}v\|^2 \leqq (\mathring{A}^2 v, v) \leqq \|\mathring{A}v\|^2,$$

т. е.

$$(\mathring{A}^2 v, v) = \|\mathring{A}v\|^2.$$

Это равенство показывает, что разностный оператор $\mathring{A}$ — самосопряжен $\big((\mathring{A}v, w)=(v, \mathring{A}w)\big)$.

Теорема 3. *Для всякой сеточной функции $v(x)$, заданной на сетке $\bar{\omega}_h$ и обращающейся в нуль на границе γ_h, имеют неравенства:*

$$\delta_0\|\mathring{A}^{m-1}v\| \leqq \|\mathring{A}^m v\| \leqq \varDelta_0\|\mathring{A}^{m-1}v\|, \tag{12}$$

где

$$\delta_0 = 8\left(\frac{1}{l_1^2}+\ldots+\frac{1}{l_n^2}\right),$$

$$\Delta_0 = 4\left(\frac{1}{h_1^2}+\ldots+\frac{1}{h_n^2}\right),$$

m: *целое положительное число и положено* $\mathring{A}^0=1$.

Доказательство идёт совершенно так же, как и в предыдущем случае.

Замечание. Из формулы (12) следует при $m=1$ известная формула:

$$\delta_0\|v\| \leqq \|\mathring{A}v\| \leqq \Delta_0\|v\|.$$

ЛИТЕРАТУРА

[1] Самарский, А. А., *Введение в теорию разностных схем*, «Наука» (1971).

[2] Соболев, С. Л., *Некоторые применения функционального анализа в математической физике*, Изд-во СО АН СССР. Новосибирск (1962).

[3] Nitsche, J. and Nitsche, J. C. C., Error estimates for the numerical solution of elliptic differential equations, *Arch. Rat. Mech. and Anal.* **5** № 4 (1960), 293—306.

[4] Андреев, В. Б., О равномерной сходимости некоторых разностных схем, *Ж. В. М. и М. Ф.* **6** № 2 (1966), 238—250.

Математический Институт Академии Наук Венгрии, Будапешт

(Поступила 7-ого января 1976 г.)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 353—359.

NOTE TO THE LAW OF THE ITERATED LOGARITHM

by

J. BECK

1. Introduction

Let $X_1, X_2, \ldots, X_n, \ldots$ be independent, identically distributed random variables (i.i.d.r.v.),

$$\mathsf{E}X_1 = 0, \quad \mathsf{E}X_1^2 = \sigma^2 < +\infty, \quad S_n = X_1 + \ldots + X_n.$$

The famous law of the iterated logarithm states that

$$\limsup_{n\to\infty} \frac{|S_n|}{(n \log\log n)^{1/2}} = \sigma\sqrt{2}$$

almost surely (a.s.). This version is due to HARTMAN and WINTNER, following the basic work of KHINTCHINE and KOLMOGOROV. STRASSEN [1] proved that this theorem remains valid also in the case $\sigma = +\infty$, i.e., under the condition $\mathsf{E}X_1^2 = +\infty$ we have

$$\limsup_{n\to\infty} \frac{|S_n|}{(n \log\log n)^{1/2}} = +\infty \quad \text{a.s.}$$

Strassen's result, together with the Hartman—Wintner theorem, seem to be complete. Indeed, these two theorems together exactly characterize how the variance affects the oscillations of successive sums of r.v. Nevertheless, the following question arises: What is the true order of magnitude of the fluctuations of successive sums of i.i.d. r.v. with infinite variance in terms of the common distribution function. In the sequel we shall define a function $f(n) = f(n, F)$ (where $F(y) = \mathsf{P}(X_1 < y)$ is the common distribution function) such that

$$\limsup_{n\to\infty} \frac{|S_n|}{f(n)} > 0 \quad \text{a.s.}$$

and

$$\frac{f(n)}{(n \log\log n)^{1/2}} \to +\infty \quad \text{if} \quad \int_{-\infty}^{+\infty} y^2\, dF(y) = +\infty.$$

Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d.r.v., $S_n = X_1 + \ldots + X_n$. Let $\tilde{F}(y)$ denote the distribution function of $X_1 - X_2$ (symmetrisation of $F(y) = \mathsf{P}(X_1 < y)$). Let

$$|a|_\alpha^\beta = \begin{cases} \beta, & \text{if} \quad a > \beta \\ \alpha, & \text{if} \quad a < \alpha \\ a, & \text{if} \quad \alpha \leqq a \leqq \beta \end{cases}$$

and let $\tilde{F}_\varepsilon(y)=|\tilde{F}(y)|_\varepsilon^{1-\varepsilon}$, i.e., $\tilde{F}_\varepsilon$ is obtained from $\tilde{F}$ by a quantile-type truncation method. Let

$$\sigma^2(\varepsilon)=\int_{-\infty}^{+\infty} y^2\,d\tilde{F}_\varepsilon(y),$$

i.e., $\sigma^2(\varepsilon)$ is a "truncated variance", and let

$$A_n=\sum_{i=3}^{n}\sigma^2\left(\frac{\log\log i}{4i}\right).$$

THEOREM. *If* $\mathsf{E}X_1=0$, *then*

$$\limsup_{n\to\infty}\frac{|S_n|}{(A_n\log\log n)^{1/2}}>0 \quad a.s.$$

We remark that Strassen's theorem is a corollary of our result. Indeed, if $\mathsf{E}X_1^2=+\infty$, then

$$\sigma^2\left(\frac{\log\log i}{4i}\right)\to\infty \quad \text{as} \quad i\to\infty.$$

2. Proof of the Theorem

Let $X_1, X_1', X_2, X_2', \dots, X_n, X_n', \dots$ be i.i.d.r.v., $\mathsf{E}X_1=0$. Let $Y_i=X_i-X_i'$, $\tilde{F}(y)=\mathsf{P}(Y_1<y)$. It suffices to show that

$$\limsup_{n\to\infty}\frac{\sum_{i=1}^{n}Y_i}{(A_n\log\log n)^{1/2}}>0 \quad \text{a.s.}$$

We can obviously restrict ourselves to the case when the distribution function $\tilde{F}(y)$ is continuous. Thus, we assume that there exist positive real numbers M_i, $i=3,4,\dots,n$, such that

$$\mathsf{P}(Y_i>M_i)=\frac{\log\log i}{4i}.$$

Let

$$\sigma_i^2=\sigma^2\left(\frac{\log\log i}{4i}\right)=\int_{-M_i}^{M_i} y^2\,d\tilde{F}(y).$$

Let $\mathbf{N}$ denote the set of integers and for an arbitrary subset $N^*\subset\mathbf{N}$ let $\bar{d}(N^*)$ denote the upper density of N^*, i.e.,

$$\bar{d}(N^*)=\limsup_{n\to\infty}\frac{1}{n}\,|\{i\leq n:\ i\in N^*\}|.$$

Finally, let

$$N_1=\{n\in\mathbf{N}:\ \sigma_{2^{n+1}}\geqq 4\sigma_{2^n}\}.$$

We prove first the following inequality

$$\bar{d}(N_1) \leqq \frac{1}{2}. \tag{1}$$

Indeed, if $\bar{d}(N_1) > \frac{1}{2}$, then $\limsup_{n\to\infty} \frac{\sigma_n}{n} > 0$. On the other hand, the condition $\mathsf{E}|Y_1| < \infty$ ($\mathsf{E}X_1 = 0$, $Y_1 = X_1 - X_1'$) implies

$$M_n \frac{\log\log n}{4n} \leqq \int_{M_n}^{+\infty} y\, d\tilde{F}(y) = o(1),$$

thus

$$\sigma_n \leqq M_n = o\left(\frac{n}{\log\log n}\right),$$

a contradiction.

From (1) we obtain

$$\bar{d}(N_2) > 0, \tag{2}$$

where

$$N_2 = \mathbf{N} \setminus N_1 = \{n \in \mathbf{N}:\ \sigma_{2^{n+1}} < 4\sigma_{2^n}\}.$$

Let

$$B_n = A_{2n} - A_n = \sum_{i=n+1}^{2n} \sigma_i^2.$$

The proof will be based on the following estimate: If $n \in N_2$, n is sufficiently large and $\delta > 0$ is sufficiently small, then

$$\mathsf{P}\left\{\sum_{i=2^n+1}^{2^{n+1}} Y_i \geqq \delta(B_{2^n} \log n)^{\frac{1}{2}}\right\} \geqq \frac{1}{2n}. \tag{3}$$

Indeed, using (2) and (3) we can complete the proof as follows.

Let $S_n = \sum_{i=1}^{n} Y_i$. Clearly, $B_n \geqq \frac{1}{2} A_{2n}$, therefore by (3) we have

$$\mathsf{P}\left\{S_{2^{n+1}} - S_{2^n} \geqq \delta(A_{2^{n+1}} \log n)^{\frac{1}{2}}\right\} \geqq \frac{1}{2n}$$

if $n \in N_2$, $n \geqq n_0$, $\delta \leqq \delta_0$. The sums $S_{2^{n+1}} - S_{2^n}$, being non-overlapping sums of independent r.v., are independent and by the Borel—Cantelli criterion, it suffices to prove that $\sum_{n \in N_2} \frac{1}{n} = +\infty$. But this is a simple consequence of (2).

To prove (3) we distinguish two cases. Let

$$I_n = \left\{i:\ 2^n < i \leqq 2^{n+1} \quad \text{and} \quad M_i \leqq \left(\frac{B_{2^n}}{\log n}\right)^{1/2}\right\}$$

and

$$I_n^c = [2^n + 1, 2^{n+1}] \setminus I_n.$$

Case 1. $|I_n|\geqq 2^{n-1}$.
Clearly

$$\mathsf{P}\Big\{\sum_{i=2^n+1}^{2^{n+1}} Y_i \geqq \delta(B_{2^n}\log n)^{\frac{1}{2}}\Big\} \geqq$$

$$\geqq \mathsf{P}\Big\{\sum_{i\in I_n} Y_i \geqq \delta(B_{2^n}\log n)^{1/2}/\prod_{i\in I_n} C_i\Big\}\times \mathsf{P}\Big\{\prod_{i\in I_n} C_i\Big\}\mathsf{P}\Big\{\sum_{i\in I_n^c} Y_i \geqq 0\Big\},$$

where $C_i=\{|Y_i|\leqq M_i\}$. Thus, we have to estimate the terms

$$\mathsf{P}\Big\{\sum_{i\in I_n} Y_i \geqq \delta(B_{2^n}\log n)^{1/2}/\prod_{i\in I_n} C_i\Big\},$$

$$\mathsf{P}\Big\{\prod_{i\in I_n} C_i\Big\} \quad \text{and} \quad \mathsf{P}\Big\{\sum_{i\in I_n^c} Y_i \geqq 0\Big\}.$$

The r.v. $\sum_{i\in I_n^c} Y_i$ is symmetrically distributed, therefore

$$\mathsf{P}\Big\{\sum_{i\in I_n^c} Y_i \geqq 0\Big\} \geqq \frac{1}{2}. \tag{4}$$

On the other hand, the events C_i, $i=1, 2, \ldots$ are independent, so we have

$$\begin{aligned}\mathsf{P}\Big\{\prod_{i\in I_n} C_i\Big\} &= \prod_{i\in I_n} \mathsf{P}\{C_i\} \geqq \prod_{i=2^n+1}^{2^{n+1}} \mathsf{P}\{C_i\} = \\ &= \prod_{i=2^n+1}^{2^{n+1}}\left(1-\frac{\log\log i}{2i}\right) \geqq \frac{1}{\sqrt{n}}.\end{aligned} \tag{5}$$

To estimate the first term we shall apply an inequality due to Kolmogorov [2].

Kolmogorov's inequality. Let $Z_1, \ldots, Z_r$ be independent r.v., $\mathsf{E}Z_i=0$, $i=1, \ldots, r$. Let $t>0$ and assume that the inequalities

$$|Z_i| \leqq \frac{\varepsilon}{t}\left(\sum_{i=1}^{r} \mathsf{E}Z_i^2\right)^{1/2} \quad i=1, \ldots, r$$

hold. Given $\gamma>0$, if $\varepsilon\leqq\varepsilon(\gamma)$ and $t\geqq t(\gamma)$, then

$$\mathsf{P}\left\{\sum_{i=1}^{r} Z_i \geqq t\left(\sum_{i=1}^{r}\mathsf{E}Z_i^2\right)^{1/2}\right\} \geqq \exp\left[-\frac{t^2}{2}(1+\gamma)\right].$$

First observe that

$$\mathsf{E}\Big(Y_i^2/\prod_{i\in I_n} C_i\Big) = \mathsf{E}(Y_i^2/C_i) = \frac{\sigma_i^2}{\mathsf{P}(C_i)} \geqq \sigma_i^2.$$

Therefore, $n\in N_2$ and $|I_n|\geqq 2^{n-1}$ imply

$$\sum_{i\in I_n}\mathsf{E}(Y_i^2/C_i) \geqq \sum_{i\in I_n}\sigma_i^2 \geqq 2^{n-1}\sigma_{2^n}^2 > 2^{n-5}\sigma_{2^{n+1}}^2 \geqq 2^{-5}B_{2^n},$$

furthermore, by $i\in I_n$

$$M_i \leqq \left(\frac{B_{2^n}}{\log n}\right)^{1/2} \leqq \left(\frac{\sum_{i\in I_n}\mathsf{E}(Y_i^2/C_i)}{2^{-5}\log n}\right)^{1/2}. \tag{6}$$

According to Kolmogorov's inequality and (6)

$$\mathsf{P}\Big\{\sum_{i\in I_n} Y_i \geqq \delta(B_{2^n}\log n)^{1/2}\Big/\prod_{i\in I_n} C_i\Big\} \geqq \frac{1}{\sqrt{n}} \tag{7}$$

if δ is sufficiently small and n is sufficiently large. By (4), (5) and (7) we obtain

$$\mathsf{P}\Big\{\sum_{i=2^n+1}^{2^{n+1}} Y_i \geqq \delta\,(B_{2^n}\log n)^{1/2}\Big\} \geqq \frac{1}{2n}.$$

Case 2. $|I_n^c|\geqq 2^{n-1}$.

Clearly

$$\mathsf{P}\Big\{\sum_{i=2^n+1}^{2^{n+1}} Y_i \geqq \delta(B_{2^n}\log n)^{1/2}\Big\} \geqq$$

$$\geqq \mathsf{P}\Big\{\sum_{i\in I_n^c} Y_i \geqq \delta(B_{2^n}\log n)^{1/2}\Big\}\mathsf{P}\Big\{\sum_{i\in I_n} Y_i \geqq 0\Big\} \geqq$$

$$\geqq \frac{1}{2}\mathsf{P}\Big\{\sum_{i\in I_n} Y_i \geqq \delta(B_{2^n}\log n)^{1/2}\Big\}.$$

We want to give a good lower bound for the last term.

Lemma. *Let $Z_1, Z_2, \dots, Z_r$ be independent, symmetrically distributed r.v.,* $\mathsf{E}Z_i=0$, $i=1, \dots, r$. *Assume that* $\mathsf{P}(Z_i>m_i)=p_i\leqq\frac{1}{10}$ *and* $\sum_{i=1}^{r} p_i\geqq 16$. *Then*

$$\mathsf{P}\left\{\sum_{i=1}^{r} Z_i \geqq m\frac{\sum_{i=1}^{r} p_i}{2}\right\} \geqq \frac{3}{8}\exp\left[-2\sum_{i=1}^{r} p_i\right],$$

where $m=\min_{1\leqq i\leqq r} m_i$.

Applying the lemma we get

$$\mathsf{P}\Big\{\sum_{i\in I_n^c} Y_i \geqq \delta(B_{2^n}\log n)^{1/2}\Big\} \geqq \mathsf{P}\left\{\sum_{i\in I_n^c} Y_i \geqq \frac{t}{2}\left(\frac{B_{2^n}}{\log n}\right)^{1/2}\right\} \geqq \frac{3e^{-2t}}{8} \geqq \frac{3}{8\sqrt{n}}$$

where $t=\sum_{i\in I_n^c}\mathsf{P}(Y_i>M_i)$, $\delta\leqq\delta_0$ and $n\geqq n_0$.

In the last step we used the estimates:

$$\sum_{i\in I_n^c}\mathsf{P}(Y_i\geqq M_i) > |I_n^c|\,\frac{\log\log(2^{n+1})}{2^{n+3}} \geqq \frac{\log n}{16}-O(1)$$

and

$$\sum_{i\in I_n^c}\mathsf{P}(Y_i>M_i) \leqq \sum_{i=2^n+1}^{2^{n+1}}\frac{\log\log i}{4i} \leqq \frac{1}{4}\log n.$$

Summarizing we have

$$\mathsf{P}\left\{\sum_{i=2^n+1}^{2^{n+1}} Y_i \geqq \delta(B_{2^n}\log n)^{1/2}\right\} \geqq \frac{3}{16\sqrt{n}} > \frac{1}{2n},$$

if n is sufficiently large and δ is sufficiently small. Thus (3) holds, which was to be proved.

3. Proof of the Lemma

For arbitrary $H\subset[1, r]=\{i\colon 1\leqq i\leqq r\}$ we define the events

$$C_H^1 = \prod_{i\in H}\{Z_i > m_i\},$$

$$C_H^2 = \prod_{i\in H^c}\{|Z_i| \leqq m_i\},$$

$$C_H^3 = \Big\{\sum_{i\in H^c} Z_i \geqq 0\Big\},$$

$$C_H^4 = \prod_{i\in H^c}\{Z_i \leqq m_i\},$$

where $H^c=[1, r]\setminus H$. Obviously,

$$\begin{aligned}\mathsf{P}\left\{\sum_{i=1}^{r} Z_i \geqq tm\right\} &\geqq \sum_{H\subset[1,r]\colon |H|\geqq t} \mathsf{P}(C_H^1)\mathsf{P}(C_H^2\cap C_H^3) \geqq \\ &\geqq \frac{1}{2}\sum_{H\subset[1,r]\colon |H|\geqq t} \mathsf{P}(C_H^1)\mathsf{P}(C_H^2).\end{aligned}\tag{8}$$

We can compute the exact value of $\mathsf{P}(C_H^2)$ and $\mathsf{P}(C_H^4)$:

$$\mathsf{P}(C_H^2) = \prod_{i\in H^c}(1-2p_i) \quad\text{and}\quad \mathsf{P}(C_H^4) = \prod_{i\in H^c}(1-p_i).$$

Therefore we get

$$\mathsf{P}(C_H^2) = \left(\prod_{i\in H^c}\frac{1-2p_i}{1-p_i}\right)\mathsf{P}(C_H^4).$$

Replacing it to (8) we obtain

$$\begin{aligned}\mathsf{P}\left\{\sum_{i=1}^{r} Z_i \geqq tm\right\} &\geqq \frac{1}{2}\sum_{H\subset[1,r]\colon|H|\geqq t}\mathsf{P}(C_H^1)\left(\prod_{i\in H^c}\frac{1-2p_i}{1-p_i}\right)\mathsf{P}(C_H^4) \geqq \\ &\geqq \frac{1}{2}\prod_{i=1}^{r}\frac{1-2p_i}{1-p_i}\Big(\sum_{H\subset[1,r]\colon|H|\geqq t}\mathsf{P}(C_H^1\cap C_H^4)\Big) = \\ &= \frac{1}{2}\prod_{i=1}^{r}\frac{1-2p_i}{1-p_i}\mathsf{P}\left\{\sum_{i=1}^{r}\chi_{\{Z_i>m_i\}} \geqq t\right\},\end{aligned}$$

where $\chi_{\{Z_i>m_i\}}$ denotes the indicator variable of the event $\{Z_i>m_i\}$, i.e., $\chi_{\{Z_i>m_i\}}=1$ if $\omega\in\{Z_i>m_i\}$ and $\chi_{\{Z_i>m_i\}}=0$ if $\omega\notin\{Z_{\cdot}>m_i\}$. Since $p_i\leqq\frac{1}{10}$,

$$\prod_{i=1}^{r}\frac{1-2p_i}{1-p_i}\geqq\exp\left[-2\sum_{i=1}^{r}p_i\right].$$

On the other hand, let $\xi=\sum_{i=1}^{r}\chi_{\{Z_i>m_i\}}$. By definition $\mathsf{E}\xi=\sum_{i=1}^{r}p_i$ and

$$\sigma^2\xi=\mathsf{E}(\xi-\mathsf{E}\xi)^2=\sum_{i=1}^{r}(p_i-p_i^2),$$

so we have

$$\mathsf{E}\xi\geqq 4\sigma\xi\quad\text{if}\quad\sum_{i=1}^{r}p_i\geqq 16. \tag{9}$$

By Chebyshev's inequality and (9)

$$\mathsf{P}\left\{\xi\geqq\frac{1}{2}\sum_{i=1}^{r}p_i\right\}=\mathsf{P}\left\{\xi\geqq\frac{1}{2}\mathsf{E}\xi\right\}\geqq$$

$$\geqq\mathsf{P}\{|\xi-\mathsf{E}\xi|\leqq 2\sigma\xi\}\geqq 1-\frac{1}{4}=\frac{3}{4}.$$

Summarizing,

$$\mathsf{P}\left\{\sum_{i=1}^{r}Z_i\geqq m\frac{\sum_{i=1}^{r}p_i}{2}\right\}\geqq\frac{3}{8}\exp\left[-2\sum_{i=1}^{r}p_i\right],$$

thus the lemma is proved.

REFERENCES

[1] Strassen, V., A converse to the law of the iterated logarithm, *Z. Wahrscheinlichkeitsrechnung* **4** (1965), 265—268.

[2] Kolmogorov, A., Über das Gesetz des iterierten Logarithmus, *Math. Ann.* **101** (1929), 126—136.

Mathematical Institute of the Hungarian Academy of Sciences
Reáltanoda u. 13—15, H—1053 Budapest

(Received June 19, 1978; in revised form November 12, 1979)

ŁOŚ LEMMA HOLDS IN EVERY CATEGORY

by

HAJNAL ANDRÉKA and ISTVÁN NÉMETI

§ 1. Introduction

There are a steadily growing number of versions of model theory and universal algebra which differ from classical model theory not only in syntax but first of all in semantics, i.e., the mathematical objects called models and their homomorphisms are different. One of the examples is *partial algebras:* The universal algebra of partial algebras is basically different from the classical one not because of syntax but because the models and their homomorphisms are a quite new category (cf. [7], [2], [18]).

Abstract model theory was started to unify at least some of the different model theories (cf. [14], [6] p. 45). In a version of abstract model theory stress is laid on the case when the models themselves and their homomorphisms are allowed to vary. In this approach, by a language we understand a triple

$$L = \langle F, M, \models \rangle,$$

where F and M are *arbitrary classes* and $\models$ is a relation between them, i.e., $\models \subseteqq M \times F$. F is *called* "the class of formulas" of L, M is called "the class of models" of L and $\models$ is called "validity relation" of L. Such an arbitrary triple is said to be a language if certain additional axioms are satisfied. We shall not quote these axioms. What is important here is that M is *not* required to consist of certain prespecified constructs (e.g., to consist of classical first order models). Instead, M is an abstract category and we do not know what its specific ingredients are. They can be classical models, partial algebras, nondeterministic algebras, intensional models, Kripke models etc. The present paper is an approach to attacking Problem 9 raised in Makowsky [15] about abstract model theory. For a detailed introduction to abstract model theory see, e.g., [20].

Independently, another approach was started to generalize parts of universal algebra and model theory to category theory with the aim of applying the resulting theory, e.g., to partial algebras, relational systems or models, etc. It appears that the area of application has much in common with that of abstract model theory. Concerning this approach, see, e.g., [13]; note that the aims here are completely different from those of algebraic theories in the sense of Lawvere.

AMS (MOS) subject classifications (1979). Primary 03C95, 18C10; Secondary 18A15, 03G30, 18C99, 03C20.

Key words and phrases. Category theory, category theoretic logic, abstract model theory, formulas and injectivity, finitely presented objects, ultraproducts.

An important idea in this approach is that universal (strict) Horn-formulas, i.e., quasi-equations can be represented by certain arrows (morphisms) and then validity corresponds to *injectivity*. I.e., in classical model theory, to every universal Horn-formula $\chi\in F$, there is an arrow $\bar{\chi}\in \mathrm{Mor}(M)$ of the category M of models such that for any model $\mathfrak{A}\in \mathrm{Ob}(M)$ we have: $\mathfrak{A}\models\chi$ if and only if $\mathfrak{A}$ is injective with respect to $\bar{\chi}$. (In the sequel we shall abbreviate "with respect to" by w.r.t.) For more detail and examples cf. [5] or [19]. In [19] and in [17] injectivity was extended to cones (from arrows) in order to represent all universally quantified formulas.

In the present paper *injectivity is extended to represent* validity of *all first order formulas*. For any category $\mathscr{C}$ we shall define the class $\mathrm{STr}(\mathscr{C})$ of small trees of $\mathscr{C}$. (Arrows are special trees.) Then we extend the definition of injectivity from arrows to trees. Now, if M is the category of classical models (or universal algebras) then to any first order formula φ there corresponds a small tree $\bar{\varphi}\in \mathrm{STr}(M)$ such that for every $\mathfrak{A}\in \mathrm{Ob}(M)$ we have:

$$\mathfrak{A}\models\varphi \quad \text{iff} \quad \mathfrak{A} \text{ is injective w.r.t. the tree } \bar{\varphi}.$$

Also to any small tree there corresponds a first order formula in exactly the same way. I.e.: If $\langle F, M, \models\rangle$ is a classical first order language, then the triple

$$\langle \mathrm{STr}(M),\ M,\ \text{"injectivity"}\rangle$$

is a language equivalent to $\langle F, M, \models\rangle$. All this is proved in [4].

Using these definitions, in Theorem 1 we shall prove that the Łoś lemma holds for any category. Therefore, if $\mathscr{C}$ is an arbitrary category with ultraproducts, then the language $\langle \mathrm{STr}(\mathscr{C}), \mathrm{Ob}(\mathscr{C}), \text{"injectivity"}\rangle$ is compact.

ILLUSTRATION 1 (for motivation). In order to facilitate the understanding of this general injectivity, here we quote some examples of how *trees* and *formulas* can be correlated to each other in the case of classical first order model theory (as a reference for the latter see [8]).

Let $\mathscr{C}$ be the category of relational structures with two binary relations R and S.

Recall from [19], Def. 3, that an object $\mathfrak{C}$ is injective w.r.t. the cone

$$\begin{tikzcd} & \mathscr{L}_0 \\ \mathfrak{A} \arrow[ur, "h_0"] \arrow[dr, "h_n"'] & \vdots \\ & \mathscr{L}_n \end{tikzcd}$$

iff every morphism $\mathfrak{A}\xrightarrow{k}\mathfrak{C}$ factors through the cone, i.e. to k there exist $i\leqq n$ and k' such that

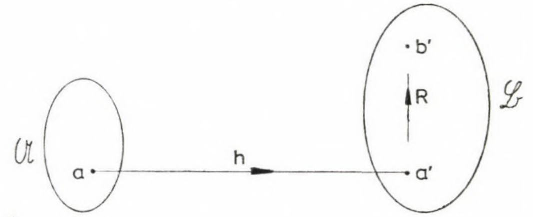
$$\begin{tikzcd}
\mathfrak{A} \arrow[rr, "h_i"] \arrow[dr, "k"'] & & \mathcal{L}_i \arrow[dl, "k'"] \\
& \mathcal{L} &
\end{tikzcd}$$

commutes.

Note that a single morphism $\mathfrak{A}\xrightarrow{h}\mathfrak{B}$ is a special cone, namely a one-member one. Thus the above definition yields the usual relation of injectivity between objects and morphisms, cf. 31.6 and 37.8 of [12].

1. Consider a tree consisting of a single arrow $\mathfrak{A}\xrightarrow{h}\mathfrak{B}$ where $\mathfrak{A}$ is a one-point set $\{a\}$ with R and S empty, $\mathfrak{B}$ is $\{a', b'\}$ with $R=\{(a', b')\}$ and S empty, further $h(a)=a'$.

The tree



corresponds to the formula

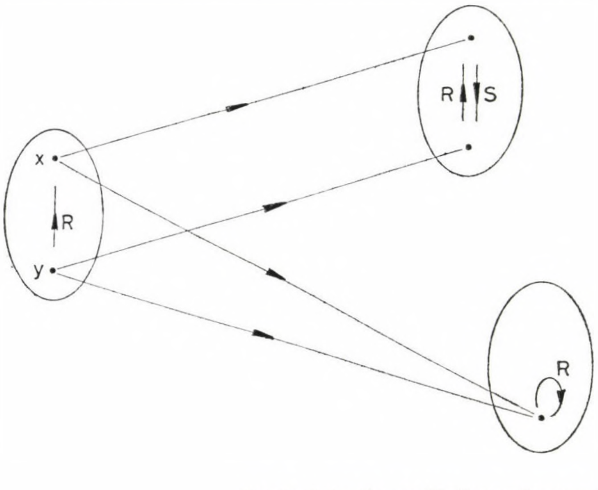
$$\forall x\, \exists y\, R(x, y).$$

This means, as is easy to check, that for any relational structure $\mathfrak{C}\in \mathrm{Ob}(\mathscr{C})$,

$$\mathfrak{C} \models \forall x\, \exists y\, R(x, y) \quad \text{iff} \quad \mathfrak{C} \text{ is injective w.r.t. } h.$$

(All the existing elements and relations are indicated in the figure.)

2. Let the next tree be a proper cone instead of a single arrow. The tree



corresponds to the formula

$$\forall xy\big(R(x, y) \to (S(y, x) \vee x = y)\big).$$

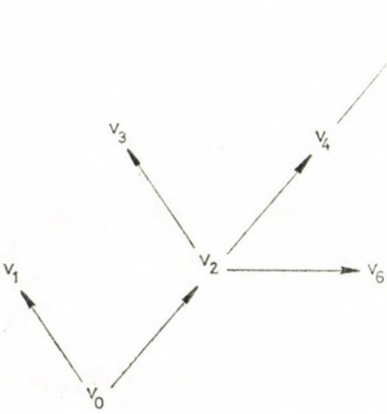
This means that for any relational structure $\mathfrak{C} \in \mathrm{Ob}(\mathscr{C})$, denoting $\forall xy\big(R(x, y) \to (S(y, x) \vee x = y)\big)$ by ψ, $\mathfrak{C} \models \psi$ iff $\mathfrak{C}$ is injective w.r.t. the above cone. This example was explained in [19].

§ 2. Definition of trees and injectivity of trees

We shall give two equivalent definitions of trees and their injectivity in a category. The first one is more intuitive while the second one is more adequate for working with.

§ 2.1. First definition

By a tree we understand a poset (i.e. a special category) which is a finite rooted tree in the usual sense. E.g., the category



is a tree. Here we say that v_4 is a successor of v_2. v_5 is not a successor of v_2 and v_5 has no successors. We say that v_i is a *successor* of v_j iff, in poset theoretical terms, v_i covers v_j, see [10]. The *root* is v_0.

DEFINITION 1. A tree of the category $\mathscr{C}$ is a diagram indexed by a finite rooted tree. More precisely, a tree of $\mathscr{C}$ is a functor $T \xrightarrow{F} \mathscr{C}$ in which T is a finite rooted tree. □

NOTATION. If v is a vertex then $F(v)$ is denoted by F_v and if w is a successor of v then F_{vw} is the obvious arrow $F_v \xrightarrow{F_{vw}} F_w$. (I.e., if $h_{vw} \in \mathrm{Hom}_T(v, w)$ then $F_{vw} = F(h_{vw})$.)

Now we turn to define *injectivity* of objects w.r.t. trees as a generalization of the usual injectivity w.r.t. arrows. The notation $a \models F$ will be used to denote that the object a is injective w.r.t. the tree F (and all this is meant in the category $\mathscr{C}$). For this end, first we define a game. To see how Definition 2 below corresponds to the usual notion of validity, the reader is referred to Illustration 2 and Definition 4 below, to [4], and [6] p. 254.

DEFINITION 2. Let $a \in \mathrm{Ob}(\mathscr{C})$ and $T \xrightarrow{F} \mathscr{C}$ be a tree of $\mathscr{C}$. Then the (a, F)-game is played by two players "$\forall$" and "$\exists$" who move alternately, and "$\forall$" moves first.

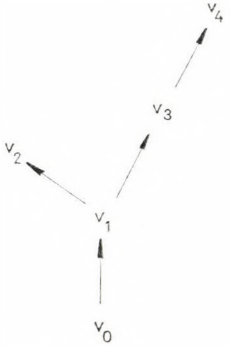
A move in the (a, F)-game is to choose a pair (v, k) such that $F_v \xrightarrow{k} a$ is a morphism of $\mathscr{C}$. The first move is that "$\forall$" chooses a pair (v_0, k_0) such that v_0 is the root of the tree. Chosen (v_n, k_n) in the n-th move, the $(n+1)$-th move is to choose a pair (v_{n+1}, k_{n+1}) such that v_{n+1} is a successor of v_n and

$$\begin{tikzcd}
F_{v_n} \arrow[rr, "F_{v_n v_{n+1}}"] \arrow[dr, "k_n"'] & & F_{v_{n+1}} \arrow[dl, "k_{n+1}"] \\
& a &
\end{tikzcd}$$

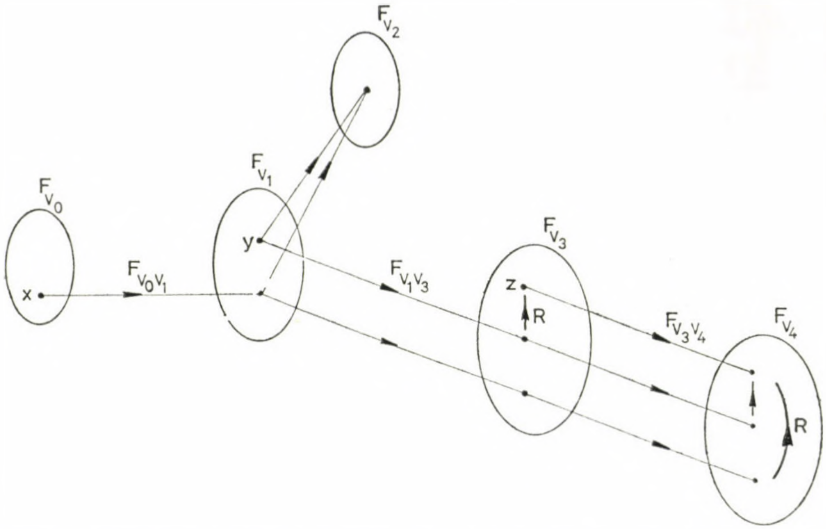
commutes.

Starting at the root v_0 the two participants "climb up" the tree. A participant looses if he cannot move. Now: we define $a \models F$ to hold if there is a winning strategy for "$\exists$" in the (a, F)-game. □

ILLUSTRATION 2. Let $\mathscr{C}$ be the category of relational structures with one binary relation R. Let T be the following tree:



Let the functor $F: T \to \mathscr{C}$ be the following:



Now F is a tree of $\mathscr{C}$ and for any $\mathfrak{A}\in\mathrm{Ob}(\mathscr{C})$, $\mathfrak{A}\models F$ if and only if

$$\mathfrak{A}\models\forall x\,\exists y\,(x\neq y\wedge\forall z(R(y,z)\to R(x,z))).$$

§ 2.2. *Second definition*

DEFINITION 3. $\mathrm{Tr}\,(\mathscr{C})$ is the smallest class $\mathscr{T}$ of pairs with the following closure property:

For any finite sequence $((d_i,\sigma_i))_{i<n}$ of elements of $\mathscr{T}$ and any cone $(d\xrightarrow{h_i}d_i)_{i<n}$ of $\mathscr{C}$, also the pair $(d,(h_i,(d_i,\sigma_i))_{i<n})$ is an element of $\mathscr{T}$. Note that we may choose $n=0$ and thus the empty sequence (which is a sequence of elements of $\mathscr{T}$), and this yields $(d,0)\in\mathscr{T}$ for any $d\in\mathrm{Ob}\,(\mathscr{C})$.

For any pair $(d,\sigma)\in\mathrm{Tr}\,(\mathscr{C})$ we define $\mathrm{dom}\,((d,\sigma))\overset{\mathrm{d}}{=}d$. The elements of $\mathrm{Tr}\,(\mathscr{C})$ are called the *trees of* $\mathscr{C}$. □

NOTATION. The following notations will be used throughout this paper. $\varphi\overset{\mathrm{d}}{=}(d,(h_i,\varphi_i)_{i<n})$ is a tree of $\mathscr{C}$, its "depth-one subtrees" are $\varphi_i\overset{\mathrm{d}}{=}(d_i,(g^i_j,\psi^i_j)_{j<m(i)})$, the ψ^i_j being the "depth-two subtrees" of $\mathscr{C}$ (see Fig. 1). Note that whenever n or $m(i)$ occurs in what follows, it will always be the n or $m(i)$ belonging to the tree φ under consideration at that place.

The sets of the h's, d's, φ's, g's, and ψ's, respectively, may be empty as well. These cases correspond to $n=0$ and $m=0$, respectively.

Induction principle. Our inductions will go along the depth-two subtrees of a tree. More precisely: The transitive closure of the binary relation "is a depth-two subtree of" is a well-founded partial order on the class $\mathrm{Tr}\,(\mathscr{C})$. Therefore we

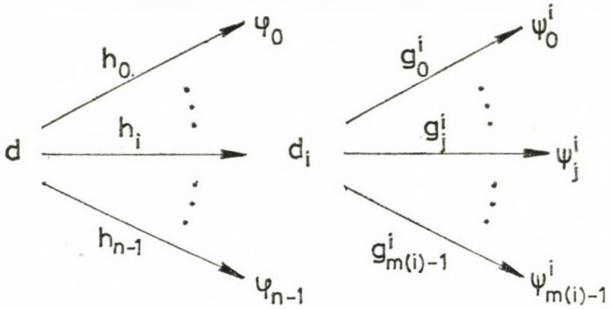


Fig. 1

can use the "Noether induction principle" as described, e.g., in [9] § 4 (following A.11). This principle reads as follows:

To prove that every element of Tr ($\mathscr{C}$) has property "P" it is enough to prove that: For any $\varphi \in \mathrm{Tr}(\mathscr{C})$, the hypothesis that every depth-two subtree of φ has property "P" implies that also φ itself has property "P".

We shall use this principle for recursive definitions, too. Now we are going to define a relation $\models \subseteq \mathrm{Ob}(\mathscr{C}) \times \mathrm{Tr}(\mathscr{C})$. We read out "$a \models \varphi$" as: "the object a is *injective* w.r.t. φ".

DEFINITION 4. Let $a \in \mathrm{Ob}(\mathscr{C})$ and $\varphi \in \mathrm{Tr}(\mathscr{C})$ be arbitrary.

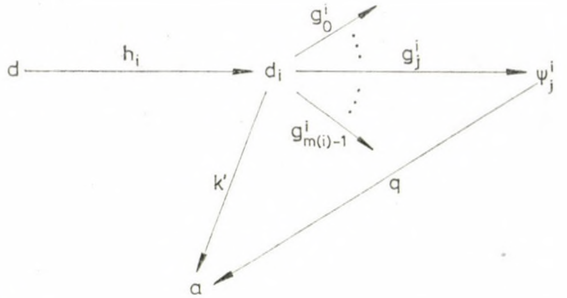
(i) First we define for arbitrary $\mathrm{dom}(\varphi) \xrightarrow{k} a$ the ternary relation "$a \models \varphi[k]$", which reads as "the arrow k *factors through* the tree φ" (or "the tree φ is true in the object a under the evaluation k"). Recall the notation:

$$\varphi \overset{\mathrm{d}}{=} \big(d, (h_i, (d_i, (g^i_j, \psi^i_j)_{j<m(i)}))_{i \leqq n-1}\big).$$

Suppose that for every $\mathrm{dom}(\psi^i_j) \xrightarrow{q} a$ the truth of $a \models \psi^i_j[q]$ is already defined. Let $d \xrightarrow{k} a$ be arbitrary. We define $a \models \varphi[k]$ to hold iff there are $i<n$ and a commutative completion

$$\begin{tikzcd}
d \arrow[rr, "h_i"] \arrow[dr, "k"'] & & d_i \arrow[dl, dashed, "k'"] \\
& a &
\end{tikzcd}$$

such that for every $j<m(i)$ and every commutative diagram



we have: $a \models \psi^i_j[q]$.

(ii) We define $a \models \varphi$ to hold iff every $d \xrightarrow{k} a$ factors through φ. (I.e. $a \models \varphi$ iff $\forall k \in \mathrm{Hom}(d, a)$ $a \models \varphi[k]$.) □

It is easy to see that this definition of injectivity with respect to a tree is equivalent with Definition 2.

Remark. Let us briefly return to the abstract model-theoretic terminology of the introduction.

Let $\mathscr{C}$ be an arbitrary category. *So far we have defined a language* $\mathrm{Tr} \overset{d}{=} \langle \mathrm{Tr}(\mathscr{C}), \mathrm{Ob}(\mathscr{C}), \models \rangle$. On the basis of the motivating examples given in Illustrations 1 and 2 the following can be proved:

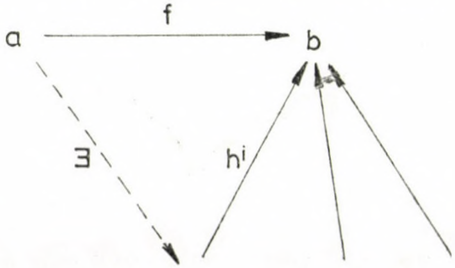
If $\mathscr{C}$ is a similarity class of classical models (cf. [8]*) then* Tr *contains (the language of) classical first order logic. More precisely, every classical first order formula corresponds to an element of* Tr ($\mathscr{C}$)*, validity of the first being equivalent with injectivity of the latter (see* [4]*).*

Now we are going to define a sublanguage STr of Tr. In the preceding case of a similarity class of classical models, this STr can be shown to be equivalent with the classical first order language. First we recall the definition of an s. small (strongly small) object of $\mathscr{C}$.

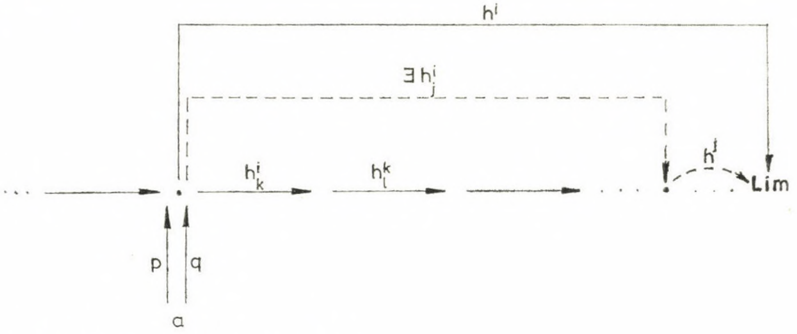
Definition 5 ([12] 22E). The object $a \in \mathrm{Ob}(\mathscr{C})$ is *s. small* if the functor $\mathrm{Hom}(a, -)\colon \mathscr{C} \to \mathrm{Set}$ preserves small direct limits. (By a direct limit we understand the colimit of a directed diagram.) □

Remark ([16]) $a \in \mathrm{Ob}(\mathscr{C})$ is s. small iff for any set I and for any directed diagram $(h^i_j\colon i \leq j \in I)$ with colimiting cocone $c = ((h^i)_{i \in I}, b)$ conditions (i) and (ii) below are satisfied.

(i): Every arrow $a \xrightarrow{f} b$ cofactors through the cocone c.



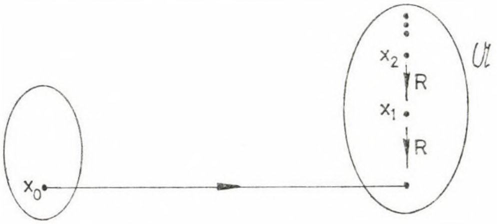
(ii): To any pair $a \overset{p}{\underset{q}{\rightrightarrows}}$ such that $ph^i = qh^i$ for some $i \in I$ there exists a $j \in I$ such that $ph^i_j = qh^i_j$.



The proof of this can be found in [16].

EXAMPLE. In categories of relational structures or partial algebras the s. small objects are exactly the ones with a finite universe in which all but a finite number of relational symbols (operations) are interpreted to be empty. In categories of total universal algebras or heterogeneous algebras the s. small objects are exactly the finitely presented ones.

In any of the above categories, if a tree contains only s. small objects then it can be shown to correspond to a finitary first order formula. Therefore s. small objects can be used to exclude trees representing infinitary formulas. For example,



represents the infinitary formula

$$\forall x_0 \exists x_1 \exists x_2 \dots \exists x_i \dots \bigwedge_{i \in \omega} R(x_{i+1}, x_i).$$

Here the object $\mathfrak{A}$ is not a small object of the category of relational structures.

DEFINITION 6. A tree of $\mathscr{C}$ is *small* if all the objects occurring in it are s. small. The class of small trees of $\mathscr{C}$ is denoted by STr $(\mathscr{C})$. □

Now we recall the definition of an *ultraproduct* of objects of $\mathscr{C}$. (Cf., e.g., [16], [1].)

DEFINITION 7. Let I be a set and let $(a_i)_{i \in I}$ be a family of objects of $\mathscr{C}$ (indexed by I). Let U be a set of subsets of I. Now consider the products $\underset{i \in X}{\mathrm{P}}\, a_i$ for the sets $X \in U$. If $X, Y \in U$ and $Y \supseteq X$ then the morphism induced by the cone of projections of $\underset{i\in Y}{\mathrm{P}} a_i$ into the product $\underset{i\in X}{\mathrm{P}} a_i$ is denoted by π_X^Y. By this we have a diagram

$$\Big(\underset{i\in Y}{\mathrm{P}} a_i \xrightarrow{\pi_X^Y} \underset{i\in X}{\mathrm{P}} a_i \colon X, Y\in U, Y \supseteqq X\Big)$$

of products and projections, which is indexed by the poset $(U, \subseteqq)$. The colimit of this diagram is denoted by

$$\big(\underset{i\in Y}{\mathrm{P}} a_i \xrightarrow{\pi^Y} (\underset{i\in I}{\mathrm{P}} a_i/U)\big)_{Y\in U}.$$

This colimit is called the *U-reduced product* of the family $(a_i)_{i\in I}$. If U is an ultrafilter then $\underset{i\in I}{\mathrm{P}} a_i/U$ is an *ultraproduct.* □

For the equivalence of the category theoretic (Def. 7) and the set theoretic definitions of reduced products see [11] Chap 3. Exercise 100 and [6] Sec. A. 3.2 p. 109. (P. Burmeister proved that the equivalence of these two notions is equivalent to the axiom of choice.)

§ 3. Łoś lemma

THEOREM 1. *Let $\mathscr{C}$ be an arbitrary category, $\varphi\in \mathrm{STr}(\mathscr{C})$ be a small tree. Let U be an ultrafilter over the set I and let $(a_i)_{i\in I}$ be a family of objects of $\mathscr{C}$. Then statements* (i) *and* (ii) *below are equivalent.*

(i) $\underset{i\in I}{\mathrm{P}} a_i/U \models \varphi$,

(ii) *there is a $Y\in U$ such that for every $i\in Y$*

$$a_i \models \varphi.$$

PROOF. We shall prove this theorem as a consequence of Theorem 2.

NOTATION. By the definition of a product, a cone $\big(d \xrightarrow{f_i} a_i \colon i\in I\big)$ induces a unique morphism $d \to \underset{i\in I}{\mathrm{P}} a_i$ to the product of the objects $(a_i)_{i\in I}$ such that f commutes with the projections π_i^I of the product cone, i.e., $f\pi_i^I = f_i$ for every $i\in I$. We denote this induced morphism by $\dot{\prod\limits_{i\in I}} f_i$.

THEOREM 2. *Let $\mathscr{C}$ be an arbitrary category, $\varphi\in \mathrm{STr}(\mathscr{C})$ be a small tree. Let U be an ultrafilter over the set I and let $(a_i)_{i\in I}$ be a family of objects of $\mathscr{C}$. Let further $\big(d \xrightarrow{k_i} a_i\big)_{i\in R}$ be an arbitrary cone of $\mathscr{C}$, where $R\subseteqq I$. Then statements* (i) *and* (ii) *below are equivalent.*

(i) $\underset{i\in I}{\mathrm{P}} a_i/U \models \varphi\Big[\Big(\dot{\prod\limits_{i\in Z}} k_i\Big)\pi^Z\Big]$ *for some $R \supseteqq Z\in U$,*

(ii) *there is an $R\supseteqq Y\in U$ such that for every $i\in Y$ $a_i \models \varphi[k_i]$.*

PROOF. The proof goes by Noether induction along the depth-two subtrees.

Let

$$\varphi = \big(d, (h_r, (d_r, (g^r_j, \psi^r_j)_{j<m(r)}))_{r<n}\big)$$

be a small tree of $\mathscr{C}$ and let $(d \xrightarrow{k_i} a_i)_{i\in R}$ be a cone of "evaluations".

A. Suppose that (ii)⇒(i) holds for every depth-two subtree ψ^r_j of φ and for any cone of "evaluations"

$$\big(\mathrm{dom}\,(\psi^r_j) \xrightarrow{q_i} a_i\big)_{i\in X}$$

where $X\subseteq I$. We prove that (ii)⇒(i) holds for φ, too.

Let (ii) hold for φ and $(k_i)_{i\in R}$, that is: Let $R\supseteq Y\in U$ be such that $(\forall i\in Y)a_i\models\varphi[k_i]$. For every $i\in Y$, let

$$\begin{tikzcd} d \arrow[rr, "h_{r(i)}"] \arrow[dr, "k_i"'] & & d_{r(i)} \arrow[dl, "k'_i"] \\ & a & \end{tikzcd}$$

be a commutative triangle the existence of which is provided by the definition of $a_i\models\varphi[k_i]$; this definition also says that for every $j<m(r(i))$ and every q_i making

$$\begin{tikzcd} d_{r(i)} \arrow[rr, "g^{r(i)}_j"] \arrow[dr, "k'_i"'] & & \mathrm{dom}(\psi^{r(i)}_j) \arrow[dl, "q_i"] \\ & a_i & \end{tikzcd}$$

commutative, it holds that $a_i\models\psi^{r(i)}_j[q_i]$. Since U is an ultrafilter and n is finite, there are $Y\supseteq Z\in U$ and $r<n$ for which $(\forall i\in Z)r(i)=r$. Let us fix this r and Z. Now we show that for this $R\supseteq Z\in U$

$$\underset{i\in I}{\mathrm{P}}\, a_i/U \models \varphi\Big[\Big(\dot{\prod_{i\in Z}}\, k_i\Big)\pi^Z\Big].$$

(This is (i) for φ and $(k_i)_{i\in R}$.)

NOTATION.

$$k \overset{\mathrm{d}}{=} \Big(\dot{\prod_{i\in Z}}\, k_i\Big)\pi^Z$$

and

$$k' \overset{\mathrm{d}}{=} \Big(\dot{\prod_{i\in Z}}\, k'_i\Big)\pi^Z.$$

See Figure 2 below.

We show that k' satisfies the requirements in the definition of $a\models\varphi[k]$.

1. $k=h_r k'$ follows by the definition of products, since for every $i\in Z$, $k_i=h_r k'_i$.

2. Let $j<m(r)$ and the arrow $\mathrm{dom}\,(\psi^r_j)\xrightarrow{q}\big(\underset{i\in I}{\mathrm{P}}\, a_i/U\big)$ be such that $k'=g^r_j q$. We have to prove $\underset{i\in I}{\mathrm{P}}\, a_i/U\models\psi^r_j[q]$.

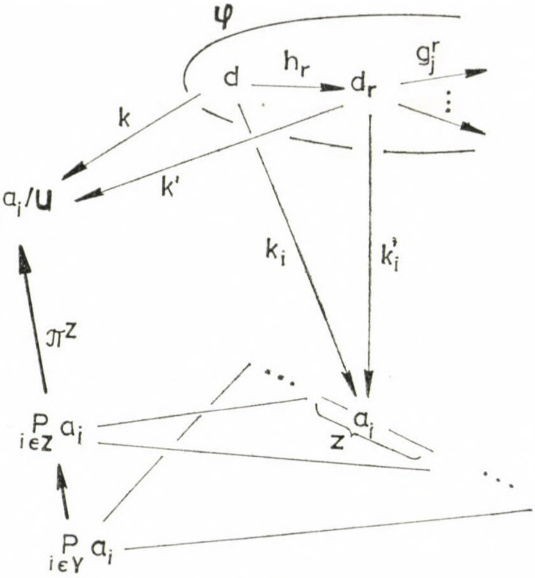

φ
d
h_r
d_r
g_j^r
k
k'
P_{i∈I} a_i/U
k_i
k_i'
π^Z
P_{i∈Z} a_i
a_i
Z
P_{i∈Y} a_i

Fig. 2

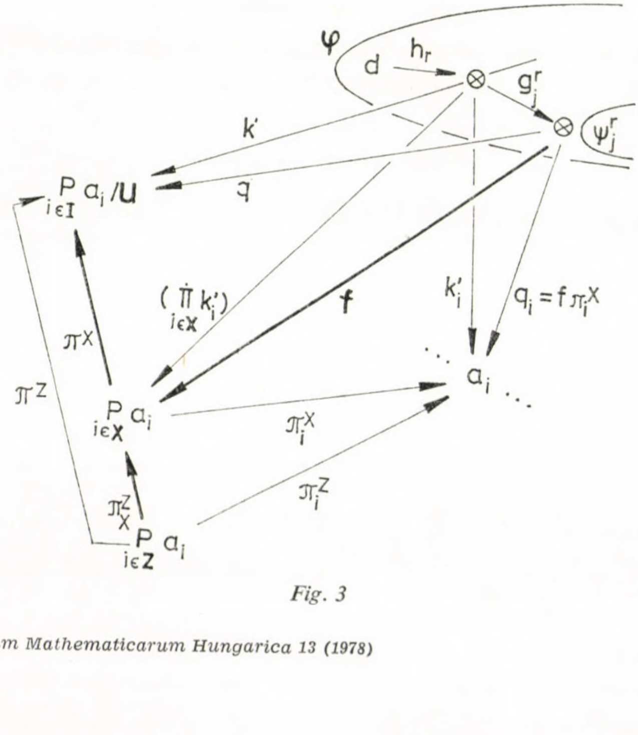

φ
d
h_r
g_j^r
ψ_j^r
k'
q
P_{i∈I} a_i/U
(Π_{i∈X} k_i')
f
k_i'
q_i = fπ_i^X
π^X
π^Z
a_i
P_{i∈X} a_i
π_i^X
π_i^Z
π_X^Z
P_{i∈Z} a_i

Fig. 3

Since φ is small, $\operatorname{cod}(g_j^r)$ is s. small and therefore there is a $Z\supseteq X\in U$ together with f such that $q=f\pi^X$. Since φ is small, also $\operatorname{dom}(g_j^r)$ is s. small and therefore $(g_j^r f)\pi^X=k'=(\dot{\prod\limits_{i\in X}} k_i')\pi^X$ implies the existence of an $X\supseteq X'\in U$ for which $(g_j^r f)\pi^X_{X'}=$ $=(\dot{\prod\limits_{i\in X}} k_i')\pi^X_{X'}$. To avoid cumbersome notation, we can assume that $X=X'$ and thus $g_j^r f=(\dot{\prod\limits_{i\in X}} k_i')$. See Figure 3.

NOTATION. $q_i\overset{\mathrm{d}}{=}f\pi_i^X$ for every $i\in X$.

Now, $k_i'=g_j^r q_i$ and therefore $a_i\models\psi_j^r[q_i]$ by the definition of k_i'. This holds for every $i\in X$, i.e., (ii) holds for the tree ψ_j^r and the cone $(q_i)_{i\in X}$. Thus by the induction hypothesis

$$\mathop{\mathrm{P}}_{i\in I} a_i/U\models\psi_j^r\left[\left(\dot{\prod_{i\in W}} q_i\right)\pi^W\right]$$

for some $X\supseteq W\in U$. Since $q=(\dot{\prod\limits_{i\in W}} q_i)\pi^W$, this means that (ii)⇒(i) is proved.

B. Next we prove (i)⇒(ii) for our fixed tree φ and cone $(k_i)_{i\in R}$. Suppose that (i)⇒(ii) holds for every depth-two subtree ψ_j^i of φ and for any cone of "evaluations" $(\operatorname{dom}(\psi_j^i)\xrightarrow{q_i}a_i)_{i\in Q}$. We prove that (i)⇒(ii) holds for φ, too.

Let (i) hold for φ and $(k_i)_{i\in R}$, i.e., let $R\supseteq Z\in U$, $k\overset{\mathrm{d}}{=}(\dot{\prod\limits_{i\in Z}} k_i)\pi^Z$ and $\mathop{\mathrm{P}}\limits_{i\in I} a_i/U\models$ $\models\varphi[k]$. Let

$$\begin{tikzcd}
d \arrow[rr, "h_r"] \arrow[dr, "k"'] & & d_r \arrow[dl, "k'"] \\
& \mathop{\mathrm{P}}\limits_{i\in I} a_i/U &
\end{tikzcd}$$

be a commutative diagram the existence of which is provided by the definition of $\mathop{\mathrm{P}}\limits_{i\in I} a_i/U\models\varphi[k]$.

Then for every $j<m(r)$ and q satisfying $k'=g_j^r q$ it holds that $\mathop{\mathrm{P}}\limits_{i\in I} a_i/U\models\psi_j^r[q]$.

Since φ is small, $\operatorname{dom}(h_r)$ as well as $\operatorname{cod}(h_r)$ are s. small objects, and therefore there are $Z\supseteq X\in U$ and f such that $k'=f\pi^X$ and further $(\dot{\prod\limits_{i\in X}} k_i)=h_r f$. See Figure 4.

We are going to show that $Y\overset{\mathrm{d}}{=}\{i\in X\colon a_i\models\varphi[k_i]\}\in U$. This will complete the proof of (i)⇒(ii).

The proof goes indirectly. Suppose that $Y\notin U$, i.e., that $W\overset{\mathrm{df}}{=}(X\setminus Y)\in U$.

NOTATION. $k_i'\overset{\mathrm{d}}{=}f\pi_i^X$ for every $i\in X$.

Obviously $k_i=h_r k_i'$ for every $i\in X$.

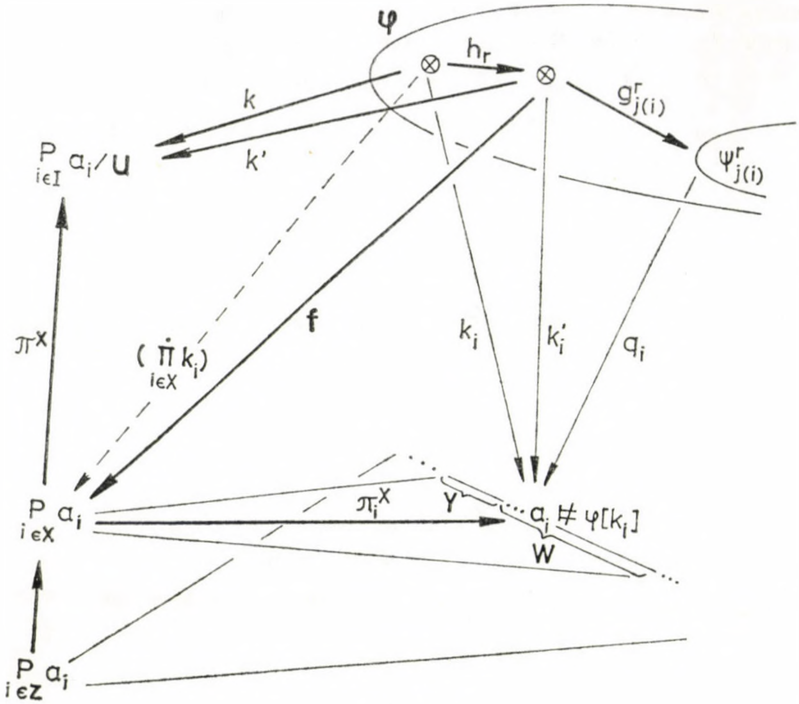

φ
h_r
$g^r_{j(i)}$
k
k'
$\underset{i\in I}{\mathrm{P}}\, a_i/U$
$\psi^r_{j(i)}$
π^X
$(\dot{\prod\limits_{i\in X}} k_i)$
f
k_i
k'_i
q_i
$\underset{i\in X}{\mathrm{P}}\, a_i$
π_i^X
Y
$a_i \not\models \varphi[k_i]$
W
$\underset{i\in Z}{\mathrm{P}}\, a_i$

Fig. 4

By the definition of $a_i \not\models \varphi[k_i]$, for every $i \in W$ there are a $j(i) < m(r)$ and a commutative triangle

$$\begin{tikzcd}
{} \arrow[rr, "g^r_{j(i)}"] \arrow[dr, "k'_i"'] & & {} \arrow[dl, "q_i"] \\
& a_i &
\end{tikzcd}$$

such that $a_i \not\models \psi^r_{j(i)}[q_i]$. Since U is an ultrafilter and $m(r)$ is finite, there are $W \supseteq Q \in U$ and $j < m(r)$ such that $(\forall i \in Q)\, j(i) = j$. By the induction hypothesis (i)⇒(ii) holds for ψ^r_j. Therefore $(\forall i \in Q \in U)$ $a_i \not\models \psi^r_j[q_i]$ implies

$$\underset{i\in I}{\mathrm{P}}\, a_i/U \not\models \psi^r_j\left[\left(\dot{\prod_{i\in Q'}} q_i\right)\pi^{Q'}\right]$$

for every $Q \supseteq Q' \in U$. This contradicts the definition of k', since

$$k' = g^r_j\left(\left(\dot{\prod_{i\in Q'}} q_i\right)\pi^{Q'}\right).$$

QED (Th 2)

Proof of Theorem 1.

1. Suppose that $Y\in U$ and $(\forall i\in Y)a_i\models\varphi$. We have to prove $\underset{i\in I}{\mathrm{P}}\, a_i/U\models\varphi$. Let $\mathrm{dom}\,(\varphi)\xrightarrow{k}(\underset{i\in I}{\mathrm{P}}\, a_i/U)$ be arbitrary. Since dom (φ) is small, there is a cone $(\mathrm{dom}\,(\varphi)\xrightarrow{k_i}a_i)_{i\in Z}$ such that $(\dot{\prod\limits_{i\in Z}}k_i)\pi^Z=k$, and we may suppose that $Z\subseteq Y$. Therefore (ii) of Theorem 2 is satisfied. By Theorem 2 this implies (i), i.e,: There is $Z\supseteq W\in U$ such that $\underset{i\in I}{\mathrm{P}}\, a_i/U\models\varphi[(\dot{\prod\limits_{i\in W}}k_i)\pi^W]$. But by the definition of reduced products, $(\dot{\prod\limits_{i\in W}}k_i)\,\pi^W=k$.

2. Let $\underset{i\in I}{\mathrm{P}}\, a_i/U\models\varphi$. Define: $Z\overset{\mathrm{d}}{=}\{i\in I\colon\, a_i\not\models\varphi\}$. By definition, there is a cone $(\mathrm{dom}\,(\varphi)\xrightarrow{k_i}a_i)_{i\in Z}$ such that $(\forall i\in Z)a_i\not\models\varphi[k_i]$. Now $Z\notin U$ because otherwise (i) of Theorem 2 would be satisfied and therefore by Theorem 2 $a_i\models\varphi[k_i]$ would hold for some $i\in Z$, which is not the case. Therefore $(I\setminus Z)\in U$ and by the definition of Z $(\forall i\in(I\setminus Z))a_i\models\varphi$.

QED (Th 1)

Remark. For any category $\mathscr{C}$, the triple

$$\langle \mathrm{STr}\,(\mathscr{C}),\ \mathrm{Ob}\,(\mathscr{C}),\ \models\rangle$$

satisfies the usual abstract model theoretic axioms, i.e., it is a language. Moreover, by Theorem 1 it is a compact language if $\mathscr{C}$ has ultraproducts.

If $\mathscr{C}$ is the category of models (or algebras) of a fixed similarity type, then $\langle \mathrm{STr}\,(\mathscr{C}),\ \mathrm{Ob}\,(\mathscr{C}),\ \models\rangle$ is equivalent to the usual first order language (see [4]).

It would be interesting to get a "logical structure" on Tr $(\mathscr{C})$ by defining two binary operations "$\vee$" and "$\wedge$", and a unary one "$\neg$": $\mathrm{Tr}\,(\mathscr{C})\to\mathrm{Tr}\,(\mathscr{C})$. These operations (or "logical connectives") should satisfy the usual axioms of classical propositional logic. (Quantifiers $\forall$, $\exists$ should also be introduced.)

A possible candidate for the "negation" $\neg$ is the following: For $\varphi\overset{\mathrm{d}}{=}(d,\sigma)\in\mathrm{Tr}\,(\mathscr{C})$ define

$$\neg\varphi\overset{\mathrm{d}}{=}(d,\,(1_d,\,\varphi))\in\mathrm{Tr}\,(\mathscr{C}),$$

where 1_d denotes the identity arrow of d, i.e.,

$$\neg\varphi=(d\xrightarrow{1_d}\varphi).$$

It can be checked that in any category $\mathscr{C}$,

$$(\forall a\in\mathrm{Ob}\,(\mathscr{C}))(\forall\varphi\in\mathrm{Tr}\,(\mathscr{C}))(\forall\,\mathrm{dom}\,(\varphi)\xrightarrow{k}a)\quad a\not\models\varphi[k]\Leftrightarrow a\models\neg\varphi[k].$$

Open problems. 1. Under which conditions can the STr-axiomatizable classes of Ob $(\mathscr{C})$ be characterized by ultraproducts and "ultraroots" analogously to Theorem 3 of [19] or Theorem 1 of [3]?

2. "Horn-trees" can be defined as special elements of STr $(\mathscr{C})$ $(\to\to\to)$. Conjecture: Horn-trees are exactly those small trees which are preserved by arbitrary reduced products.

Acknowledgement: Thanks are due to Professor Peter Burmeister for his help, thorough reading and many useful suggestions.

REFERENCES

[1] Andréka, H.—Makai, E. jr.—Márki, L.—Németi, I., Reduced products in categories, *Contributions to General Algebra* (Proc. Conf. Klagenfurt, 1978), Verlag J. Heyn, Klagenfurt, 1979, pp. 25—45.

[2] Andréka H.—Németi, I., Generalization of variety and quasivariety concept to partial algebras through category theory. Preprint, Math. Inst. Hung. Acad. Sci. No. 5/1976 (to appear in *Dissertationes Mathematicae (Rozprawy Mat.)* No. 204).

[3] Andréka, H.—Németi, I., Formulas and ultraproducts in categories, *Beiträge zur Algebra u. Geom.* **8** (1979), 133—151.

[4] Andréka, H.—Németi, I., Injectivity in categories to represent all first order formulas, *Demonstratio Math.* **12** (1979), 717—732.

[5] Banaschewski, B.—Herrlich, H., Subcategories defined by implications, *Houston J. Math.* **2** (1976), 149—171.

[6] Barwise, J. (ed): *Handbook of Mathematical Logic,* Studies in Logic and the Foundations of Mathematics, Vol 90, North-Holland, Amsterdam, 1977.

[7] Burmeister, P.—John, R.—Pasztor, A., On closed morphisms in the category of partial algebras, *Contributions to General Algebra* (Proc. Conf. Klagenfurt, 1978), Verlag J. Heyn, Klagenfurt, 1979, pp. 69—76.

[8] Chang, C. C.—Keisler, H. J., *Model Theory,* North-Holland, Amsterdam, 1973.

[9] Cohn, P. M., *Universal Algebra,* Harper & Row, London, 1965.

[10] Grätzer, G., *Lattice Theory,* W. H. Freeman and Company, San Francisco, 1971.

[11] Grätzer, G., *Universal Algebras,* Second edition, Springer-Verlag, Berlin, 1979.

[12] Herrlich, H.—Strecker, G. E., *Category Theory,* Allyn and Bacon, London, 1973.

[13] John, R., A note on implicational subcategories, *Contributions to Universal Algebra* (Proc. Conf. Szeged, 1975), (Coll. Math. Soc. J. Bolyai, Vol 17), North-Holland, 1977, pp. 213—223.

[14] Makowsky, J. A., What is the axiomatic theory of modeltheoretic languages? Logical Semester, Banach Centre, Warsaw, 1973 (mimeographed).

[15] Makowsky, J. A., Topological Model Theory, *Model Theory and Applications* (Proc. CIME, Bressanone, 1975), Edizioni Cremonese, Roma, 1975, pp. 123—150.

[16] Matthiessen, G., Regular and strongly finitary structures over strongly algebroidal categories, *Canad. J. Math.* **30** (1978), 250—261.

[17] Németi, I., Axiomatizability of classes of partial algebras via category theory, Dissertation, July, 1976, Budapest.

[18] Németi, I., From hereditary classes to varieties in abstract model theory and partial algebra, *Beiträge zur Algebra u. Geometrie* **7** (1978), 69—78.

[19] Németi, I.—Sain, I., Cone-implicational subcategories and some Birkhoff-type theorems, *Universal Algebra* (Proc. Colloq. Esztergom, 1977), Colloq. Math. Soc. J. Bolyai, Vol. 29, North-Holland, Amsterdam (to appear).

[20] Sain, I., There are general rules for specifying semantics: Considerations on abstract model theory, *Computational Linguistics and Computer Languages* **13** (1979), 195—250.

Mathematical Institute of the Hungarian Academy of Sciences
Reáltanoda u. 13—15, H—1053 Budapest

(Received July 20, 1978; in revised form October 22, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 377—383.

OPTIMAL ESTIMATION OF FUNCTIONS OF PROBABILITY DENSITIES

by

WOLFGANG WERTZ

Introduction. In a series of papers, the author has derived several existence theorems for optimal density estimators ([18]–[21]). From the reasonig in these papers it follows that several results can be generalized to the problem of estimating certain functions of the unknown density. For some parametric cases, optimal solutions of the optimization problem can be derived explicitely from this theory — several examples exhibiting interesting aspects are given, including the estimation of the density itself. KLEBANOW [6] has derived optimal density estimators by an entirely different method; connections to his work are pointed out in the following.

Historically, the problem of estimating functions of the density is as old as density estimation itself: in 1956, GRENANDER estimates the hazard rate in his paper [5] on mortality measurement, in 1964, NADARAJA [8] estimates convolution components, and WATSON and LEADBETTER ([16] and [17]) estimate again the hazard rate by kernel-type estimators. Of the more recent work we only mention: MAJOR [7], RÉVÉSZ [10] and STONE [15] treating regression estimators, RICE and ROSENBLATT [11] and AHMAD and LIN [1] treating hazard functions, YANG [22] estimating the life-expectancy, the thorough study by BOSQ ([2] and [3]) of several of the mentioned problems and SCHÜLER and WOLFF [19], estimating the functional $f \to \int f^2(x)\,dx$. In [13] further references are listed.

Results. Let $(X, \mathfrak{X})$ be a measurable space, λ and μ σ-finite measures on $(X, \mathfrak{X})$, $\mathfrak{F}$ a set of probability densities with respect to μ, $P_f(A) := \int\limits_A f\,d\mu$ for every $A \in \mathfrak{X}$ and $f \in \mathfrak{F}$ and Δ a mapping from $\mathfrak{F}$ to $\mathbf{R}^X$, such that $\Delta(f)\colon X \to \mathbf{R}$ is $\mathfrak{X}$-measurable for every $f \in \mathfrak{F}$. As examples may serve, under appropriate conditions: $\Delta(f) := f$, the distribution function, a convolution component, the density of a certain order statistic, a k-fold convolution $f * \ldots * f$, the density of the arithmetic mean of k independent observations, partial and ordinary derivates, a conditional density, or functionals (from this point of view rather trivial examples) such as p-th moments about a point, the mode, the information $\Delta(f)(x) :\equiv \int (f'/f)^2 f\,d\mu$ and $\Delta(f)(x) :\equiv$ $:\equiv \int f^2\,d\mu$. The problem is to estimate $\Delta(f)$, based on n independent observations.

Let Φ and Ψ be complementary Young-functions and Φ satisfy condition Δ_2, i.e.,

$$(1) \qquad \Phi(2u) \leqq M\Phi(u) \quad \text{for every} \quad u \geqq 0 \quad \text{and a certain} \quad M > 0.$$

In the following, Φ is always assumed to fulfil (1). $(X^n, \mathfrak{X}^n, P_f^n)$ denotes n-fold product spaces, $x^n = (x_1, \ldots, x_n)$ elements of X^n. Let $L_{\Phi,f} = L_{\Phi,f}(P_f^n \otimes \lambda)$ denote the Orlicz-space of the equivalence classes of $\mathfrak{X}^n\otimes\mathfrak{X}$-measurable functions $h\colon X^n\times X\to\mathbf{R}$ with $\int\limits_{X^n\times X}\Phi\circ h dP_f^n\otimes\lambda<\infty$. $L_{\Phi,f}$ is a Banach-space with norm

$$\|h\|_{\Phi,f}=\sup\left\{\int hg\,dP_f^n\otimes\lambda\colon g\in L_{\Psi,f} \text{ and } \int\Psi\circ g\,dP_f^n\otimes\lambda\leqq 1\right\}.$$

We further assume

$$(2)\qquad s:=\sup_{f\in\mathfrak{F}}\|\Delta(f)\|_{\Phi,f}<\infty.$$

For every $h\in\bigcap\limits_{f\in\mathfrak{F}}L_{\Phi,f}$ let $\|h\|_\Phi:=\sup\limits_{f\in\mathfrak{F}}\|h\|_{\Phi,f}$ and $\Lambda_\Phi:=\{h\colon \|h\|_\Phi<\infty\}$. Then Λ_Φ is again a Banach-space. We make no difference between functions and equivalence classes. For further details and references see [19].

DEFINITION. Under the conditions (1) and (2), every $\hat f\in\Lambda_\Phi$ is called an estimator of $\Delta(f)$. $\hat f_0$ is called optimal, if

$$R_\Phi(\hat f_0):=\sup_{f\in\mathfrak{F}}\|\hat f_0-\Delta(f)\|_{\Phi,f}\leqq R_\Phi(\hat f)\quad\text{for every}\quad \hat f\in\Lambda_\Phi.$$

Let M_Ψ denote the set of all $g:=\sum\limits_{i=1}^k\alpha_i g_i$, α_i real, $g_i\in L_{\Psi,f_i}$ $(f_1,\dots,f_k\in\mathfrak{F})$ and $\tilde M_\Psi$ the set of all functionals φ induced by functions $g\in M_\Psi$ by

$$\langle h,\varphi\rangle:=\sum_{i=1}^k\alpha_i\langle h,\varphi_i\rangle_{f_i}:=\sum_{i=1}^k\alpha_i\int hg_i\,dP_{f_i}^n\otimes\lambda.$$

The strongest topology of Λ_Φ, such that every $\varphi\in\tilde M_\Psi$ is continuous, is called the *σ_Φ-topology* (cf. PITCHER's topology [9]). This topology is not stronger than the weak topology of Λ_Φ, and it fulfils Hausdorff's separation axiom. The unit sphere $E:=\{h\in\Lambda_\Phi\colon \|h\|_\Phi\leqq 1\}$ of Λ_Φ is compact in the σ_Φ-topology (see [19]).

THEOREM 1 (Existence of optimal estimators). *Under the condition* (1) *for Ψ and Φ, and* (2), *there is an optimal $\hat f\in\Lambda_\Phi$.*

THEOREM 2 (Reduction by sufficiency). *Let* (1) *and* (2) *be satisfied and let there be a real number u_0 with $u\leqq\Phi(u)$ for every $u\geqq u_0$. Let $\mathfrak{C}\subset\mathfrak{X}^n$ be a σ-algebra, sufficient for $\{P_f^n\colon f\in\mathfrak{F}\}$, $\mathfrak{C}$ or $\mathfrak{X}$ be countably generated and $\hat f\in\Lambda_\Phi$. Then there is an $\hat f_0\in\Lambda_\Phi$ fulfilling: $\hat f_0$ is $\mathfrak{C}\otimes\mathfrak{X}$-measurable,*

$$\|\hat{\boldsymbol{f}}_0-\Delta(f)\|_{\Phi,f}\leqq\|\hat f-\Delta(f)\|_{\Phi,f}\quad \textit{for every}\quad f\in\mathfrak{F},\ \textit{and}\ R_\Phi(\hat f_0)\leqq R_\Phi(\hat f).$$

Further, there is a λ-null-set $N\in\mathfrak{X}$ with

$$\hat f_0(\cdot,x)=E_f[\hat f(\xi^n,\cdot)|\mathfrak{C}](\cdot)\quad P_f^n\textit{-almost}$$

everywhere for every $x\notin N$ and every $f\in\mathfrak{F}$. ($\xi^n=(\xi_1,\dots,\xi_n)$ is a random variable distributed according to P_f^n.)

The proofs of Theorems 1 and 2 are like in [19], with f substituted by $\Delta(f)$.

Reduction by invariance. Let G be a locally compact, σ-compact, amenable group with σ-algebra $\mathfrak{G}$ of its Borel-sets and ν be a right invariant regular Haar-measure. Let D denote the Orlicz-space $L_\Phi(X,\mathfrak{X},\lambda)$. For $Y=X^n$, $=X$, $=\mathfrak{F}$ and

$=D$, let $\gamma\to\gamma_Y$ be homomorphisms from G to the one-to-one mappings of Y onto itself, $(\gamma_1\gamma_2)_Y=(\gamma_1)_Y\circ(\gamma_2)_Y$ and $(\gamma^{-1})_Y=(\gamma_Y)^{-1}$. For short, we write $\gamma:=\gamma_Y$, if no confusion arises.

DEFINITION. The estimation problem is called invariant with respect to G if the following assumptions are fulfilled:

(3.1) $\gamma f\in\mathfrak{F}$ for every $\gamma\in G$ and $f\in\mathfrak{F}$

(3.2) $(\Delta f)(\gamma^{-1}x)=[\Delta(\gamma f)](x)$ for every $\gamma\in G$, $x\in X$, $f\in\mathfrak{F}$

(3.3) $(\gamma d)(x)=d(\gamma^{-1}x)$ for every $\gamma\in G$, $x\in X$, $d\in D$

(3.4) $\gamma A\in\mathfrak{X}^n$ and $P_f^n(A)=P_{\gamma f}^n(\gamma A)$ for every $\gamma\in G$, $A\in\mathfrak{X}^n$, $f\in\mathfrak{F}$

(3.5) For every $h\in\Lambda_\Phi$, the mapping $(\gamma, x^n, x)\to h(\gamma x^n, \gamma x)$ is $\mathfrak{G}\otimes\mathfrak{X}^n\otimes\mathfrak{X}$-measurable

(3.6) $\lambda(\gamma^{-1}A)=\lambda(A)$ for every $\gamma\in G$, $A\in\mathfrak{X}$.

An estimator $\hat f\in\Lambda_\Phi$ is invariant, if $\hat f(\gamma x^n, \cdot)=\gamma[\hat f(x^n, \cdot)]$ for every $\gamma\in G$, $x^n\in X^n$. (We tacitly assume $\hat f(x^n, \cdot)$ to belong to $L_\Phi(X, \mathfrak{X}, \lambda)$, which is no restriction of the generality.)

The amenability of the group G implies the existence of a net $\{g_\alpha\}_{\alpha\in A}$ of functions $g_\alpha\in L_1(G, \mathfrak{G}, \nu)$ satisfying $g_\alpha\geqq 0$, $\int g_\alpha\,d\nu=1$ and

$$\lim_{\alpha\in A}\int|g_\alpha(\delta\gamma^{-1})-g_\alpha(\delta)|\,d\nu(\delta)=0 \quad \text{for every} \quad \gamma\in G$$

(see GREENLEAF [4], theorems 2.4.2 and 2.4.3). Defining $\nu_\alpha(B):=\int_B g_\alpha d\nu$ $(B\in\mathfrak{G})$, (ν_α) is a net of asymptotically right invariant probability measures:

$$\lim_{\alpha\in A}|\nu_\alpha(B\gamma)-\nu_\alpha(B)|=0 \quad \text{for every} \quad \gamma\in G \text{ and } B\in\mathfrak{G}. \tag{4}$$

For a fixed $\hat f\in\Lambda_\Phi$ we set $\hat f_\alpha(x^n, x):=\int\hat f(\gamma x^n, \gamma x)d\nu_\alpha(\gamma)$. Then the risk of $\hat f_\alpha$ is not greater than the risk of $\hat f$. Further, there is an $\hat f_0\in\Lambda_\Phi$ and a subnet of $(\hat f_\alpha)$ converging to $\hat f_0$, $R_\Phi(\hat f_0)\leqq R_\Phi(\hat f)$ and $\hat f_0$ is equivalent to an invariant estimator. This reasoning (for details, the proofs in [21] and [19] are to be slightly modified) leads to the following

THEOREM 3. *Under the conditions of Theorem* 1 *and* (3.1)—(3.6), *there is an invariant optimal estimator.*

Taking $p>1$, $\Phi(t):=|t|^p/p$ and $q:=p/(p-1)$, then for every $h\in L_p(X^n\times X, \mathfrak{X}^n\otimes\mathfrak{X}, P_f^n\otimes\lambda)$ the usual L_p-norm $\|h\|_{p,f}$ equals $q^{-1/q}\cdot\|h\|_{\Phi,f}$. Set $\delta_p(\hat f):=$ $:=\sup_{f\in\mathfrak{F}}\|\hat f-\Delta(f)\|_{p,f}^p$, the usual L_p-risk; then $\delta_p(\hat f)=q^{1-p}[R_\Phi(\hat f)]^p$, hence the concepts of optimality with respect to L_p- and Orlicz-risk are equivalent.

In the case of *translation classes* $\mathfrak{F}$ of densities on the real line actually optimal estimators can be constructed. Theorem 4 makes use of a sufficient statistic, whereas Theorem 5 gives an explicit solution, analogous to the Pitman estimator for translation parameters. A corresponding formula for density estimators is derived by a similar method by Klebanow [6].

We assume for the following quadratic loss, $(X, \mathfrak{X}) := (\mathbf{R}, \mathfrak{B})$, $\mu = \lambda =$ Lebesgue measure, $\mathfrak{F} = \{f_\gamma \colon \gamma \in \mathbf{R}\}$ a one-dimensional translation class of densities $(f_\gamma(x) = f(x-\gamma)$ with a fixed $f)$, $\Delta(f_\gamma)(x) = (\Delta f)(x-\gamma)$, and f and $\Delta(f)$ square integrable. Then, apparently, the problem is invariant with respect to the group of translations.

THEOREM 4. *Let T be a one-dimensional sufficient transformation for γ and let the problem in terms of T remain invariant. Denote $g_\gamma := dP^T_{f_\gamma}/d\lambda$ and $g := g_0$, and let $g_\gamma(x) = g(x-\gamma)$. Then*

$$\hat{f}(x^n, x) = \psi_0(x - T(x^n)) \tag{5}$$

with

$$\psi_0(y) := \int (\Delta(f))(y+t) \cdot g(t)\, dt = \int (\Delta(f))(y-t) \cdot g(-t)\, dt$$

defines an optimal estimator of $\Delta(f)$.

PROOF. With regard to Theorems 2 and 3, one can restrict attention to estimators of the form

$$(x^n, x) \to \psi(x - T(x^n)),$$

hence the problem is to minimize

$$\delta_2(\hat{f}) := \sup_{\gamma \in \mathbf{R}} \iint [\psi(x-t) - (\Delta f_\gamma)(x)]^2 g_\gamma(t)\, dt\, dx =$$

$$= \iint [\psi(x) - (\Delta f)(x+t)]^2 g(t)\, dt\, dx.$$

ψ minimizes this integral, if and only if for every $\varphi \in L_2(\mathbf{R}, \mathfrak{B}, \lambda)$

$$\iint \varphi(x) [\psi(x) - (\Delta f)(x+t)]\, g(t)\, dt\, dx = 0$$

holds. This yields the solution. The details are similar to those in [20].

THEOREM 5. *Let the conditions listed above Theorem 4 be satisfied. Then*

$$\hat{f}(x^n, x) = \int (\Delta(f))(x-t) \prod_{i=1}^{n} f(x_i - t)\, dt \Big/ \int \prod_{i=1}^{n} f(x_i - t)\, dt \tag{6}$$

defines an optimal estimator of $\Delta(f)$.

PROOF. By Theorem 3 it is sufficient to consider translation-invariant estimators, i.e., those satisfying

$$\hat{f}(x_1 - a, \ldots, x_n - a, x - a) = \hat{f}(x_1, \ldots, x_n, x) \quad \text{for every} \quad a \in \mathbf{R},\ x^n \in \mathbf{R}^n,\ x \in \mathbf{R}.$$

Then

$$\delta_2(\hat{f}) = \sup_{\gamma \in \mathbf{R}} \int_{\mathbf{R}^n} \int_{\mathbf{R}} [\hat{f}(x^n, x) - (\Delta(f_\gamma))(x)]^2\, dx \prod_{i=1}^{n} f_\gamma(x_i)\, dx^n =$$

$$= \int_{\mathbf{R}^n} \int_{\mathbf{R}} [\hat{f}(x^n, x) - (\Delta(f))(x)]^2 \prod_{i=1}^{n} f(x_i)\, dx\, dx^n =$$

$$= \int_{\mathbf{R}^n} \int_{\mathbf{R}} [\hat{f}(x_1 - x, \ldots, x_n - x, 0) - (\Delta(f))(x)]^2 \prod_{i=1}^{n} f(x_i)\, dx\, dx^n.$$

We write $h(y_1, \dots, y_n) := \hat{f}(y_1, \dots, y_n, 0)$. Let $\xi_1, \dots, \xi_n$ be independent random variables, distributed according to P_f, and $\eta_1 := \xi_1, \eta_i := \xi_i - \xi_1$ for $i = 2, \dots, n$. It follows that

$$\delta_2(\hat{f}) = E_f\left\{\int \left[h(y, \xi_2 - \xi_1 + y, \dots, \xi_n - \xi_1 + y) - (\Delta(f)(\xi_1 - y)\right]^2 dy\right\} =$$
$$= \int E_f\left[h(y, \eta_2 + y, \dots, \eta_n + y) - (\Delta(f))(\eta_1 - y)\right]^2 dy.$$

Using the minimum property of the regression function (see e.g. SCHMETTERER [12]), the integrand is minimized for every $y \in \mathbf{R}$, if

$$h(y, y_2 + y, \dots, y_n + y) = E_f[\Delta(f)(\eta_1 - y) | \eta_2 = y_2, \dots, \eta_n = y_n].$$

Using the corresponding conditional density, this equals

$$\frac{\int (\Delta(f))(y_1 - y) f(y_1) \prod_{i=2}^{n} f(y_i + y_1)\, dy_1}{\int f(y_1) \prod_{i=2}^{n} f(y_i + y_1)\, dy_1}.$$

By a simple substitution, formula (6) is obtained.

COROLLARY. *Under the conditions of Theorem* 4 *or* 5, *let* $\Delta(f)$ *be a probability density for every* f *and* $\hat{f}$ *be an optimal estimator obtained by* (5) *or* (6). *Then* $\hat{f}(x^n, \cdot)$ *is a probability density for every* $x^n \in \mathbf{R}^n$.

Some examples for the translation case

1) Let f be the density of the normal distribution $N(0, \sigma^2)$ ($\sigma^2 > 0$, known) and $\Delta(f) := f$. Then the optimal estimator is the density of $N\left(\bar{x}_n, \frac{n+1}{n}\sigma^2\right)$ (cf. [20]).

2) Let f be the uniform density: $f := I_{[0,1]}$. Then

$$\hat{f}(x^n, x) := \begin{cases} \dfrac{1 + x - x_{(n)}}{1 + x_{(1)} - x_{(n)}} & \text{if } x_{(n)} - 1 \leqq x < x_{(1)} \\ 1 & \text{if } x_{(1)} \leqq x < x_{(n)} \\ \dfrac{1 + x_{(1)} - x}{1 + x_{(1)} - x_{(n)}} & \text{if } x_{(n)} \leqq x < x_{(1)} + 1 \\ 0 & \text{otherwise} \end{cases}$$

is the optimal estimator for f ($x_{(1)}$ and $x_{(n)}$ denote the smallest and largest observation, respectively). The continuity of $x \to \hat{f}(x^n, x)$ is remarkable; a tendency of smoothing by optimal density estimation occurs in all examples given here.

3) Let f be the exponential density $f(x)=e^{-x}I_{(0,\infty)}(x)$. Then

$$\hat{f}(x^n, x) := \begin{cases} \dfrac{n}{n+1}\, e^{n(x-x_{(1)})} & \text{if} \quad x \leqq x_{(1)} \\ \dfrac{n}{n+1}\, e^{-(x-x_{(1)})} & \text{if} \quad x > x_{(1)} \end{cases}$$

is optimal for f.

Examples 1 to 3 can be easily derived by Theorem 5, 1 and 3 also by use of Theorem 4 using the sufficient statistics $\bar{x}_n$ and $x_{(1)}$, respectively.

4) Let f be the normal density $N(0, \sigma^2)$ and $\Delta(f):=f'$, the derivate of f:

$$\Delta(f_\gamma)(x) = -(2\pi\sigma^6)^{-1/2}(x-\gamma)\exp[-(x-\gamma)^2/2\sigma^2].$$

$$\hat{f}(x^n, x) = -\left[2\pi\sigma^6\left(\frac{n+1}{n}\right)^3\right]^{-1/2}(x-\bar{x}_n)\exp\left[-\frac{n(x-\bar{x}_n)^2}{2(n+1)\sigma^2}\right]$$

is the optimal estimator of $\Delta(f)$, and this is the derivative of the optimal estimator of f (cf. Example 1).

5) Let f be as in Example 4 and $\Delta(f)$ be the density of the empirical mean of k independent $N(0, \sigma^2)$-variables, i.e., $\Delta(f)$ equals the density of $N(0, \sigma^2/k)$. The optimal estimator is the density of $N\left(\bar{x}_n, \sigma^2\left(\frac{1}{k}+\frac{1}{n}\right)\right)$.

6) Estimation of a normal convolution component: let $\xi=\eta+\zeta$, where ζ is $N(0, \tau^2)$ distributed, η has unknown density g, ξ has normal density $N(\gamma, \sigma^2)$ and is observable ($\sigma^2>0$, $\tau^2>0$ are known, γ is unknown), ζ and η are independent. Hence f is normal $N(0, \sigma^2)$. $\Delta(f)$ is the density of η, namely the density of $N(0, \sigma^2+\tau^2)$. (In applications, ζ can represent errors by measurement.) The optimal estimator of $\Delta(f)$ is the density of $N\left(\bar{x}_n, \sigma^2+\tau^2+\frac{\sigma^2}{n}\right)$.

7) Let f be as in Example 3, $\Delta(f)$ the density of the l-th order statistic $\xi^k_{(l)}$ of k independent observations. The optimal estimator is

$$\hat{f}(x^n, x) = nl\binom{k}{l}e^{n(x-x_{(1)})}\int\limits_{\max(0,\, x-x_{(1)})}^{\infty}(1-e^{-t})^{l-1}e^{-(n+k-l+1)t}\,dt.$$

In the case $l=1$ (estimation of the density of the smallest of k observations) this turns out to be

$$\hat{f}(x^n, x) = \begin{cases} \dfrac{nk}{n+k}\, e^{n(x-x_{(1)})} & \text{if} \quad x < x_{(1)} \\ \dfrac{nk}{n+k}\, e^{-k(x-x_{(1)})} & \text{if} \quad x \geqq x_{(1)}, \end{cases}$$

whereas $\Delta(f_\gamma)(x)=ke^{-k(x-\gamma)}I_{(\gamma,\infty)}(x)$.

REFERENCES

[1] AHMAD, I. A.—LIN, P. E., Nonparametric estimation of a vector-valued bivariate failure rate, *Ann. Statist.* **5** (1977), 1027—1038.

[2] BOSQ, D., Contribution à la théorie de l'estimation fonctionnelle, *Publ. Inst. Statist. Univ. Paris* **19** (1970), fasc. 2, 1—96.

[3] BOSQ, D., Contributions à la théorie de l'estimation fonctionnelle, II, *Publ. Inst. Statist. Univ. Paris* **19** (1970), fasc. 3, 97—177.

[4] GREENLEAF, F. P., *Invariant means on topological groups and their applications,* Van Nostrand-Reinhold Co., New York, 1969.

[5] GRENANDER, U., On the theory of mortality measurement, II, *Skand. Aktuarietidskr.* **39** (1956), 125—153.

[6] Клебанов, Л. Б., Параметрические оценки плотности и характеризация семейств распределений с достаточными статистиками для параметра сдвига, *Зап. научн. сем. ЛОМИ* **79** (1978), 11—16, 101.

[7] MAJOR, P., On a non-parametric estimation of the regression function, *Studia Sci. Math. Hungar.* **8** (1973), 347—361.

[8] Надарая, Э. А., Оценка компонента свертки, *Сообщ. АН Груз. ССР* **34**, 1 (1964), 19—24.

[9] PITCHER, T. S., A more general property than domination for sets of probability measures, *Pacific J. Math.* **15** (1965), 597—611.

[10] RÉVÉSZ, P., How to apply the method of stochastic approximation in the nonparametric estimation of a regression function, *Math. Operationsforsch. Statist.* **8** (1977), 119—126.

[11] RICE, J.—ROSENBLATT, M., Estimation of the log survivor function and hazard function, *Sankhyā Ser. A* **38** (1976), 60—78.

[12] SCHMETTERER, L., *Mathematische Statistik,* 2. Auflage, Springer-Verlag, Wien, 1966.

[13] SCHNEIDER, B.—WERTZ, W., Statistical density estimation— a bibliography, *Internat. Statist. Rev.* **47** (1979), 155—175.

[14] SCHÜLER, L.—WOLFF, H., Zur Schätzung eines Dichtefunktionals, *Metrika* **23** (1976), 149—154.

[15] STONE, C. J., Consistent nonparametric regression (with discussion), *Ann. Statist.* **5** (1977), 595—645.

[16] WATSON, G.—LEADBETTER, M., Hazard analysis I, *Biometrika* **51** (1964), 175—184.

[17] WATSON, G.—LEADBETTER, M., Hazard analysis II, *Sankhyā Ser. A* **26** (1964), 101—116.

[18] WERTZ, W., On the existence of density estimators, *Studia Sci. Math. Hungar.* **9** (1974), 45—50.

[19] WERTZ, W., Invariante und optimale Dichteschätzungen, *Math. Balkanica* **4** (1974), 707—722.

[20] WERTZ, W., On unbiased density estimation, *An. Acad. Brasil. Ci.* **47** (1975), 65—72.

[21] WERTZ, W., Invariant density estimation, *Monatsh. Math.* **81** (1976), 315—324.

[22] YANG G. L., Estimation of a biometric function, *Ann. Statist.* **6** (1978), 112—116.

Institut für Statistik der Technischen Universität
A—1040 Wien, Argentinierstrasse 8/7

(Received October 3, 1978)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 385—391.

ON WILKINS' INEQUALITY

by

Á. ELBERT

Let $[a, b]$ be a finite interval, $f(t)$ a positive continuous function on $[a, b]$, finally $\alpha = \min_{a \leqq t \leqq b} f(t)$ and $\beta = \max_{a \leqq t \leqq b} f(t)$. We are interested in values of the integrals

$$I(f) = I(f, x;\, a, b) = \frac{1}{(b-a)^2} \int_a^b [f(t)]^x\, dt \int_a^b [f(t)]^{-x}\, dt.$$

We may assume $x \geqq 0$ since $I(f, -x;\, a, b) = I(f, x;\, a, b)$. By the Cauchy—Schwarz inequality it is clear that $I(f) \geqq 1$.

In the case $x=1$ the lowest upper bound is $(\alpha+\beta)^2/4\alpha\beta$ (see PÓLYA and SZEGŐ, [1] p. 57), and later WILKINS [2] has found — supposing in addition the concavity of the function $f(t)$ — that

$$I(f) \leqq \frac{\beta \left(\dfrac{\beta \log \dfrac{\beta}{\alpha}}{\beta - \alpha} - \dfrac{\beta + \alpha}{2\beta} \right)^2}{2(\beta - \alpha) \left(\dfrac{\beta \log \dfrac{\beta}{\alpha}}{\beta - \alpha} - 1 \right)}. \tag{1}$$

Here we shall deal with the lowest upper bound of $I(f, x;\, a, b)$ for all $x > 0$ (the case $x=0$ is trivial) and not only for the concave functions but for the convex ones, too. Moreover, our proof will be different from that of Wilkins and we shall not assume the monotonicity of $f(t)$ as he did.

We start with a simple lemma.

LEMMA. *Let $z_1(\xi), z_2(\xi)$ be continuous functions defined on $[\alpha, \beta]$ such that*

$$z_1(\xi) > z_0 > z_2(\xi) \quad \textit{for} \quad \alpha \leqq \xi \leqq \beta$$

and let the function $\gamma(\xi)$ be nondecreasing, $-\infty < \gamma(\alpha) < \gamma(\beta) < \infty$. Let $g_0(\xi)$ be a continuous and positive function on $[\alpha, \beta]$. Then the quadratic polynomial

$$Q(z) = \int_\alpha^\beta g_0(\xi)[z - z_1(\xi)][z - z_2(\xi)]\, d\gamma(\xi) = Az^2 + Bz + C$$

has real zeros and $B^2 > 4AC$. The integral here is taken in the Lebesgue—Stieltjes sense.

Proof. Under the conditions of the Lemma $A>0$ and $Q(z_0)<0$, thus the statement of Lemma is trivial.

Let us introduce the functions $h^+(s), h^-(s)$ and $K(s,x)$ for $s>0, x>0$ by

$$h^+(s)=\frac{1}{1+x}\frac{s^{1+x}-1}{s-1}$$

(2)

$$h^-(s)=\begin{cases}\dfrac{\log s}{s-1} & \text{for} \quad x=1\\[2mm] \dfrac{1}{1-x}\dfrac{s^{1-x}-1}{s-1} & \text{for} \quad x\neq 1,\end{cases}$$

and

$$K(s,x)=\frac{[h^+(s)-h^-(s)]^2}{4\big(1-h^+(s)\big)\big(h^-(s)-1\big)}. \tag{3}$$

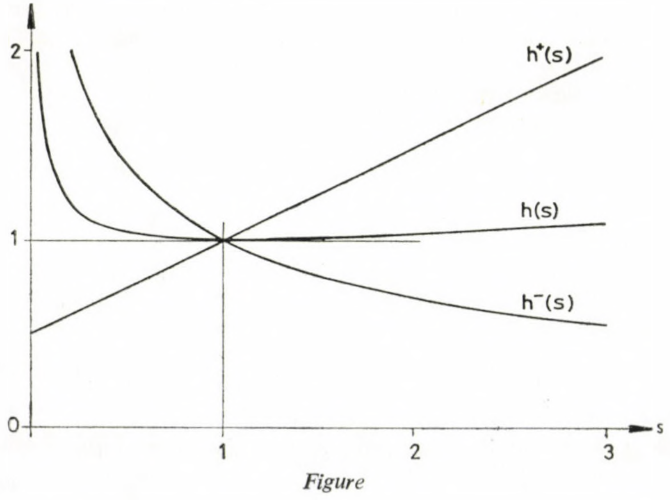
Let $h(s)=h^+(s)h^-(s)$ and $H(s)=(s-1)^2[h^+(s)+h^-(s)-2h(s)]/s$. Then $H(1)=H'(1)=0$ and $H''(s)=x(s-1)(s^x-s^{-x})/s^2>0$ for $s>0, s\neq 1$ therefore $H(s)>0$ for these values of s. On the other hand

$$h'(s)=\frac{H(s)}{(s-1)^3},$$

hence $h(s)$ takes on its minimum at $s=1$ and thus we get the relations

$$\frac{h^+(s)+h^-(s)}{2}>h(s)>1 \qquad s>0,\ s\neq 1. \tag{4}$$

It is not difficult to show that $h^+(s)$ strictly increases $h^-(s)$ strictly decreases on $(0,\infty)$ and $h(s)$ decreases on $[0, 1]$ and increases on $(1,\infty)$ (see Figure, where these functions are displayed for $x=1$).



Figure

In our study a crucial role is played by the fact that the function $K(s, x)$ is strictly decreasing on $(0, 1]$ and strictly increasing on $[1, \infty)$. Indeed, by (3) we have

$$\frac{d}{ds}\log K(s, x) = \frac{x}{(s-1)^2}\,\frac{h^+(s)+h^-(s)-2}{[h^-(s)-h^+(s)][1-h^+(s)][h^-(s)-1]}\,H(s).$$

Taking into consideration (4) and the monotonicity properties of the functions $h^-(s), h^+(s)$ we find that $K(s, x)$ has the monotonicity property stated above. Concerning the lowest upper bound of $I(f, x;\ a, b)$ we shall prove the next theorem.

THEOREM. *Let $f(t)$ be a continuous function on $[a, b]$ with $0<\alpha\leqq f(t)\leqq\beta$ then*

$$I(f, x;\ a, b) \leqq \begin{cases} K\left(\dfrac{\alpha}{\beta}, x\right) & \textit{if } f \textit{ is concave} \\[2ex] K\left(\dfrac{\beta}{\alpha}, x\right) & \textit{if } f \textit{ is convex.} \end{cases}$$

The inequality here is sharp.

REMARK. By (2) and (3) it is clear that our Theorem contains Wilkins' inequality (1) in the case $x=1$ for concave functions $f(t)$.

PROOF. Let us introduce the function $\varrho(u)$ for $0\leqq u\leqq\beta$ by

$$\varrho(u) = \lambda\{t;\ f(t)\geqq u,\ a\leqq t\leqq b\}, \tag{5}$$

where $\lambda\{\cdot\}$ denotes the Lebesgue measure of the point set indicated in $\{\cdot\}$. It is clear that the function $\varrho(u)$ is nonincreasing and $\varrho(u)=b-a$ for $0\leqq u\leqq\alpha$. On the other hand $\varrho(u)$ is concave (convex) on $[\alpha, \beta]$ if the function $f(t)$ is concave (convex). Moreover, in the convex case $\varrho(\beta)=0$. Hence $\varrho(u)$ can be written in the form

$$\varrho(u) = \int_{\alpha}^{\beta} c(u, \xi)\, d\gamma(\xi), \tag{6}$$

where $\gamma(\xi)$ is a suitable nondecreasing function of bounded variation, further in the concave case

$$c(u, \xi) = \begin{cases} 1 & 0\leqq u\leqq\xi \\[1ex] \dfrac{\beta-u}{\beta-\xi} & \xi<u\leqq\beta \end{cases} \tag{7}$$

and in the convex case

$$c(u, \xi) = \begin{cases} 1 & 0\leqq u\leqq\alpha \\[1ex] \dfrac{\xi-u}{\xi-\alpha} & \alpha\leqq u<\xi \\[1ex] 0 & \xi<u\leqq\beta. \end{cases} \tag{8}$$

Since $\varrho(u)=b-a$ for $0\leqq u\leqq\alpha$ we have from (6) in both cases

$$b-a = \int_{\alpha}^{\beta} d\gamma(\xi). \tag{9}$$

The actual value of $\gamma(\xi)$ can be easily determined if $\varrho(u)$ is twice continuously differentiable. Otherwise $\varrho(u)$ is a limit of twice continuously differentiable functions, say, $\varrho_1(u), \varrho_2(u), \dots$ such that

$$\varrho_n(u) = \int\limits_\alpha^\beta c(u, \xi)\, d\gamma_n(\xi), \qquad n = 1, 2, \dots.$$

Then by (9) the sequence $\{\gamma_n(\xi)-\gamma_n(\alpha)\}_{n=1}^\infty$ is bounded, hence we can select a subsequence of it which has the limit $\gamma(\xi)$ almost everywhere in $[\alpha, \beta]$. From now on the quantity δ denotes the quantity β in the concave case and the quantity α in the convex case. Making use of (5) we have for $x>0$

$$\lambda\{t;\, [f(t)]^x \geqq u,\, a \leqq t \leqq b\} = \varrho\left(u^{\frac{1}{x}}\right) \quad 0 \leqq u \leqq \beta^x,$$

hence by (6)

$$\int\limits_a^b [f(t)]^x\, dt = \int\limits_0^{\beta^x} \varrho\left(u^{\frac{1}{x}}\right) du = x \int\limits_0^\beta \varrho(v) v^{x-1}\, dv =$$

$$= x \int\limits_0^\beta \Big[\int\limits_\alpha^\beta c(v, \xi)\, d\gamma(\xi)\Big] v^{x-1}\, dv = \int\limits_\alpha^\beta x \int\limits_0^\beta c(v, \xi) v^{x-1}\, dv\, d\gamma(\xi),$$

and by (7), (8)

$$\int\limits_a^b [f(t)]^x\, dt = \delta^x \int\limits_\alpha^\beta h^+\left(\frac{\xi}{\delta}\right) d\gamma(\xi). \tag{10}$$

From (5) it follows that

$$\lambda\{t;\, [f(t)]^{-x} \geqq u,\, a \leqq t \leqq b\} = \begin{cases} b-a & \text{for} \quad 0 \leqq u \leqq \beta^{-x} \\ b-a-\varrho\left(u^{-\frac{1}{x}}\right) + \lambda\left\{t;\, f(t) = u^{-\frac{1}{x}}\right\} \end{cases}$$

for $\beta^{-x} \leqq u \leqq \alpha^{-x}$.

Owing to the concavity (convexity) of the function $f(t)$ the set $\left\{t;\, f(t) = u^{-\frac{1}{x}}\right\}$ consists of at most two points if $\alpha < u^{-\frac{1}{x}} < \beta$ hence its measure is zero, and this measure can be positive only if either $u^{-\frac{1}{x}} = \alpha$ or $u^{-\frac{1}{x}} = \beta$. Hence by (6), (9) and (7), (8) we have

$$\begin{aligned} \int\limits_a^b [f(t)]^{-x}\, dt &= (b-a)\beta^{-x} + \int\limits_{\beta^{-x}}^{\alpha^{-x}} \left[b-a-\varrho\left(u^{-\frac{1}{x}}\right)\right] du = \\ &= \delta^{-x} \int\limits_\alpha^\beta h^-\left(\frac{\xi}{\delta}\right) d\gamma(\xi). \end{aligned} \tag{11}$$

By (10), (11) the value of $I(f)$ can be expressed as

$$I(f) = \frac{1}{(b-a)^2} \int\limits_\alpha^\beta h^-\left(\frac{\xi}{\delta}\right) d\gamma(\xi) \int\limits_\alpha^\beta h^+\left(\frac{\xi}{\delta}\right) d\gamma(\xi). \tag{12}$$

Let Ψ be some positive number with

$$\Psi^2 > K\left(\frac{\alpha\beta}{\delta^2}, x\right). \tag{13}$$

It is not difficult to show that $K(s, x) > h(s)$ for $s > 0, s \neq 1$. Hence

$$\Psi^2 > h\left(\frac{\alpha\beta}{\delta^2}\right).$$

Let us consider the quadratic polynomial

$$Q(z) = z^2 \int_\alpha^\beta h^+\left(\frac{\xi}{\delta}\right) d\gamma(\xi) - 2(b-a)\Psi z + \int_\alpha^\beta h^-\left(\frac{\xi}{\delta}\right) d\gamma(\xi). \tag{14}$$

By (9) this can be written in the form

$$Q(z) = \int_\alpha^\beta \left[z^2 h^+\left(\frac{\xi}{\delta}\right) - 2\Psi z + h^-\left(\frac{\xi}{\delta}\right)\right] d\gamma(\xi).$$

The quantity in the brackets is a quadratic polynomial of the variable z and by (13) it has two real zeros $z_1\left(\frac{\xi}{\delta}\right)$ and $z_2\left(\frac{\xi}{\delta}\right)$, where

$$\begin{aligned} z_1(s) &= \frac{\Psi + \sqrt{\Psi^2 - h(s)}}{h^+(s)}, \\ z_2(s) &= \frac{\Psi - \sqrt{\Psi^2 - h(s)}}{h^+(s)} = \frac{h^-(s)}{\Psi + \sqrt{\Psi^2 - h(s)}}. \end{aligned} \tag{15}$$

In order to apply the Lemma we need to determine the values $\bar{z}_1 = \min\limits_{\alpha \leqq \xi \leqq \beta} z_1\left(\frac{\xi}{\delta}\right)$, $\bar{z}_2 = \max\limits_{\alpha \leqq \xi \leqq \beta} z_2\left(\frac{\xi}{\delta}\right)$.

By (15) and (3) it is not difficult to check the relations

$$\begin{aligned} G(s) &= h^+(s)^2[z_1(s) - z_1(1)][z_1(s) - z_2(1)][z_2(s) - z_1(1)][z_2(s) - z_2(1)] = \\ &= [h^-(s) - h^+(s)]^2 - 4\Psi^2[1 - h^+(s)][h^-(s) - 1] = \\ &= \begin{cases} 0 & \text{for } \; s = 1 \\ 4[1 - h^+(s)][h^-(s) - 1]\{K(s, x) - \Psi^2\} & \text{for } \; s > 0, \; s \neq 1. \end{cases} \end{aligned} \tag{16}$$

Consider first the concave case. By virtue of (13) the function $G(s)$ does not vanish on $\left[\frac{\alpha}{\beta}, 1\right)$, hence the factors on the right side of (16) have the same property. On the one hand we have from $z_1(1) > z_2(1)$ that

$$z_2(s) < z_1(1) \quad \text{for} \quad \frac{\alpha}{\beta} \leqq s \leqq 1.$$

On the other hand by (15)

$$\frac{d}{ds} z_1(s)|_{s=1} = -\frac{x\left(\Psi+\sqrt{\Psi^2-1}\right)}{2} < 0,$$

therefore

$$z_1(s) > z_1(1) \quad \text{for} \quad \frac{\alpha}{\beta} \leqq s < 1.$$

Taking into consideration the monotonicity of $h(s)$ and $h^-(s)$ (see Figure) we have from (15) that $z_2(s)$ strictly decreases on $\left[\frac{\alpha}{\beta}, 1\right]$, hence

$$z_2\left(\frac{\alpha}{\beta}\right) > z_2(s) > z_2(1) \quad \text{for} \quad \frac{\alpha}{\beta} < s < 1.$$

Thus we conclude that $\bar{z}_1 = z_1(1) > \bar{z}_2 = z_2\left(\frac{\alpha}{\beta}\right)$.

Choosing any value $z_0 \in (\bar{z}_2, \bar{z}_1)$ and applying the Lemma for $Q(z)$ in (14) we obtain

$$\int_\alpha^\beta h^+\left(\frac{\xi}{\beta}\right) d\gamma(\xi) \int_\alpha^\beta h^-\left(\frac{\xi}{\beta}\right) d\gamma(\xi) < \Psi^2(b-a)^2,$$

hence by (12) $I(f) < \Psi^2$. Combining this with (13) we have $I(f) \leqq K\left(\frac{\alpha}{\beta}, x\right)$, in accordance with our Theorem.

A similar argumentation on the interval $\left[1, \frac{\beta}{\alpha}\right]$ will give the desired inequality $I(f) \leqq K\left(\frac{\beta}{\alpha}, x\right)$ in the convex case, and the details will be omitted.

Finally, we show the sharpness of the inequalities in our Theorem.

Let the real number p be given by

$$p = \frac{h^+\left(\frac{\alpha}{\beta}\right) + h^-\left(\frac{\alpha}{\beta}\right) - 2}{2\left(1-h^+\left(\frac{\alpha}{\beta}\right)\right)\left(h^-\left(\frac{\alpha}{\beta}\right)-1\right)}.$$

Then according to the relations in (4) we have $0 < p < 1$. Let the concave function $f^-(t)$ be defined on [0, 1] by

$$f^-(t) = \begin{cases} \alpha + \frac{\beta-\alpha}{p} t & 0 \leqq t \leqq p \\ \beta & p \leqq t \leqq 1. \end{cases}$$

Then

$$\int_0^1 f^-(t)^x \, dt = \beta^x \frac{h^-\left(\frac{\alpha}{\beta}\right) - h^+\left(\frac{\alpha}{\beta}\right)}{2\left(h^-\left(\frac{\alpha}{\beta}\right)-1\right)},$$

and

$$\int_0^1 f^-(t)^{-x}\,dt = \beta^{-x}\,\frac{h^-\left(\frac{\alpha}{\beta}\right)-h^+\left(\frac{\alpha}{\beta}\right)}{2\left(1-h^+\left(\frac{\alpha}{\beta}\right)\right)},$$

hence $I(f^-, x;\ 0, 1)=K\left(\frac{\alpha}{\beta}, x\right)$, as we wanted.

In the convex case the value of p should be chosen as

$$p=\frac{h^+\left(\frac{\beta}{\alpha}\right)+h^-\left(\frac{\beta}{\alpha}\right)-2}{2\left(1-h^+\left(\frac{\beta}{\alpha}\right)\right)\left(h^-\left(\frac{\beta}{\alpha}\right)-1\right)},$$

and the convex function $f^+(t)$ as

$$f^+(t)=\begin{cases}\beta-\dfrac{\beta-\alpha}{p}t & 0\leqq t\leqq p\\ \alpha & p\leqq t\leqq 1.\end{cases}$$

For this function we have $I(f^+, x;\ 0, 1)=K\left(\frac{\beta}{\alpha}, 1\right)$, which completes the proof of our Theorem.

REFERENCES

[1] PÓLYA, G.—SZEGŐ, G., *Aufgaben und Lehrsätze aus der Analysis,* Vol. I, Springer-Verlag, Berlin, 1954.

[2] WILKINS, J. E. JR., The average of the reciprocal of a function, *Proc. Amer. Math. Soc.* **6** (1955), 806—815.

Mathematical Institute of the Hungarian Academy of Sciences,
Reáltanoda u. 13—15, H—1053 Budapest

(Received November 20, 1978)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 393—409.

ON THE GENERALIZATIONS OF TOTAL PARACOMPACTNESS

by

J. DEÁK

In this paper, we deal with some generalizations of the total paracompactness and with their role in dimension theory. In § 1, we define these generalizations (only one of them — the weak total paracompactness — is new, see Definition 1T) and give examples showing that all these classes of spaces are different (in several cases, examples satisfying better separation axioms ought to be found). In § 2, we prove that the inductive dimensions coincide in weakly totally paracompact pointwise totally normal (see Definition 1CC) spaces (Theorem 2A); this is a generalization of a theorem by FRENCH [Fr1]. In § 3, we prove the coincidence of the inductive dimensions for some spaces having an ind-nice subbase (Theorem 3D) — nice in the sense of E. DEÁK [D4, D5], see Definition 3A. The proof of this theorem is based on an inequality valid for all spaces, namely that $\mathrm{ind}^{\triangledown} X \leqq \mathrm{sbd}\, X$ (Theorem 3H), where sbd (see Definition 3F) is the subbase dimension introduced by the author [D*2], while $\mathrm{ind}^{\triangledown}$ is defined in 3G. In § 4, the same inequality yields Theorem 4J for the directional dimension (Definition 4C) of E. DEÁK [D1, D2, D6].

TERMINOLOGY. A *space* is a topological space. A *normal* space need not be a T_1-space. Let X be a space, $A \subset X$ and $\mathcal{S}$ a system of subsets of X; we say that $\mathcal{S}$ is a *covering* of A if $A = \cup\mathcal{S}$; we say that $\mathcal{S}$ *covers* A if $A \subset \cup\mathcal{S}$.

The *inductive dimensions* ind and Ind are used as in [N2] or [P]. Let d_1 and d_2 be dimension functions; $\mathrm{d}_1 X \leqq \mathrm{d}_2 X$ means: "if $\mathrm{d}_2 X$ is finite, then $\mathrm{d}_1 X \leqq \mathrm{d}_2 X$"; $\mathrm{d}_1 X = \mathrm{d}_2 X$ means: "if $\mathrm{d}_1 X$ or $\mathrm{d}_2 X$ is finite, then $\mathrm{d}_1 X = \mathrm{d}_2 X$".

NOTATIONS. For a set A in a space, $\bar{A}$ is the closure, int A is the interior, Fr A is the boundary, $\mathrm{Fr}_B\, A$ is the boundary in the subspace B and $|A|$ is the cardinality of A. Let $\mathcal{S}$ be a system of subsets of a space and n a natural number, then

$$\bar{\mathcal{S}} = \{\bar{S} \colon S \in \mathcal{S}\}, \quad \mathrm{Fr}\,\mathcal{S} = \{\mathrm{Fr}\, S \colon S \in \mathcal{S}\},$$

$$[\mathcal{S}]^n = \{\mathcal{A} \colon \mathcal{A} \subset \mathcal{S}, |\mathcal{A}| = n\}, \quad \mathbf{B}\mathcal{S} = \{\cap A \colon A \subset \mathcal{S}, |A| < \omega\}.$$

AMS (MOS) subject classifications. Primary 54F45; Secondary 54D99, 54E99, 54G20.

Key words and phrases. Totally paracompact (=TPC), almost TPC, σ-TPC, order TPC, closure TPC, dominatedly TPC, weakly TPC, totally normal, pointwise totally normal, S_1-space, strongly paracompact, completely paracompact, nice collection of subspaces, direction, directional structure, ind, Ind, dim, $\mathrm{ind}^{\triangledown}$, sbd, Dim, Dind, indc, DIM.

§ 1. Definitions and examples

From the point of view of the dimension theory, there is a wide gap between compactness and paracompactness: while a lot of fairly strong theorems hold for compact spaces, little is known about the behaviour of the classical dimension functions in paracompact spaces. Looking for an intermediate notion, FORD [F] introduced total paracompactness in 1963; the idea dates back to 1959 when totally paracompact metric spaces were considered (without having been given a name) in an announcement by CORSON, MCMINN, MICHAEL and NAGATA [CMMN]. Total paracompactness and its generalizations proved to be useful in investigating equalities and inequalities between dimension functions.

1A DEFINITION (FORD [F]). A space is *totally paracompact* (TPC)[1] if each of its bases has a locally finite subsystem covering the space.

1B DEFINITION. Let $\mathscr{Y}$ and $\mathscr{Z}$ be systems of sets in a space. $\mathscr{Y}$ is a *tight refinement of* $\mathscr{Z}$ if for each $Y\in\mathscr{Y}$ there is a $Z\in\mathscr{Z}$ with $Y\subset Z$ and $\operatorname{Fr} Y\subset\operatorname{Fr} Z$.

1C DEFINITION [D*3]. A space is *almost totally paracompact* (A-TPC) if each of its bases admits a locally finite open tight refinement covering the space.

1D DEFINITION (NAGAMI [N1]). A space is *σ-totally paracompact* (σ-TPC) if each of its bases has a σ-locally finite open tight refinement covering the space.

1E DEFINITION (FITZPATRICK and FORD [FF]). A space is *order totally paracompact* (O-TPC) if each of its bases has an ordered (not necessarily well-ordered) open tight refinement $\mathscr{V}$ covering the space such that for each $V\in\mathscr{V}$, the system of the elements of $\mathscr{V}$ preceding V is locally finite at the points of V.[2]

1F DEFINITION. Let $\mathscr{Y}$ and $\mathscr{Z}$ be systems of sets in a space. $\mathscr{Y}$ is a *boundary-refinement*[3] of $\mathscr{Z}$ if for each $Y\in\mathscr{Y}$, $\operatorname{Fr}\mathscr{Z}$ has a locally finite closed refinement covering $\operatorname{Fr} Y$.

1G DEFINITION (FRENCH [Fr1]). A space X is *closure totally paracompact* (C-TPC) if every base $\mathscr{B}$ of X has a locally finite boundary-refinement covering X and refining $\overline{\mathscr{B}}$.

1H DEFINITION. A space X is *dominated*[4] by the closed covering $\mathscr{F}$ if for every subset A of X, A is closed iff there is a subsystem $\mathscr{F}_0$ of $\mathscr{F}$ such that $A\subset\bigcup\mathscr{F}_0$ and $A\cap F$ is closed for each $F\in\mathscr{F}_0$.

1I DEFINITION (FRENCH [Fr2]). A space X is *dominatedly totally paracompact*[5]

[1] According to the context, TPC means either "totally paracompact" or "total paracompactness".

[2] In the original definition, local finiteness at the points of $\overline{V}$ is required. FRENCH [Fr1, Fr2] uses the definition given here; cf. the remark at the end of [FF].

[3] A boundary-refinement need not be a refinement.

[4] Cf. PEARS [P]. FRENCH [Fr2] uses the expression "X has the weak topology with respect to $\mathscr{F}$", but PEARS [P] means something else by "weak topology with respect to a closed covering".

[5] FRENCH [Fr2] calls this class of spaces C-TPC, too. We have chosen another name to avoid confusion.

(D-TPC) if for each base $\mathscr{B}$ of X, $\overline{\mathscr{B}}$ admits a closed refinement $\mathscr{F}$ dominating X such that for each $F\in\mathscr{F}$, Fr F is dominated by a closed refinement of Fr $\mathscr{B}$.

1J The following diagram shows the relations between the notions defined above:

$$\text{TPC} \overset{(1)}{\Rightarrow} \text{A-TPC} \overset{(2)}{\Rightarrow} \sigma\text{-TPC} \overset{(3)}{\Rightarrow} \text{O-TPC} \overset{(4)}{\Rightarrow} \text{C-TPC} \overset{(5)}{\Rightarrow} \text{D-TPC}$$

((1), (2) and (5) are obvious; (3) and (4) are proved in French [Fr1]). None of these implications can be reversed. (When French wrote [Fr2], he did not know if (3), (4) or (5) can be reversed.)

1K Example. *An* A-TPC *but not* TPC *space*. Let X be the set of the irrational numbers with the usual topology. X is not TPC [CMMN]. On the other hand, let $\mathscr{B}$ be a base of X. As X is paracompact, $\mathscr{B}$ has a locally finite open refinement $\mathscr{U}$ covering X. The covering dimension of X is 0, so by a theorem of Dowker ([AP] Chapter 4, Theorem 24), $\mathscr{U}$ admits an open refinement $\mathscr{V}$ of order 1 covering X. The elements of $\mathscr{V}$ are open and closed, so $\mathscr{V}$ is obviously a tight refinement of $\mathscr{B}$. Thus X is A-TPC.

1L Example. *A* σ-TPC *but not* A-TPC *space*. Let X be the subspace of the Euclidean plane $\mathbf{R}^2$ defined by

$$\mathbf{R}^2 - X = \{\langle a, b\rangle\colon\ a\in\mathbf{Q},\ b = 0\}$$

where $\mathbf{Q}$ is the set of the rational numbers. X is a Lindelöf space, so it is evidently σ-TPC. We shall show that X is not A-TPC.

a) Take the closed subspace

$$Z = \{\langle a, b\rangle\colon\ a\notin\mathbf{Q},\ b = 0\}$$

of X. Z is homeomorphic to the space of the irrational numbers, so it is not TPC, i.e. it has a base $\mathscr{B}_Z$ not containing any locally finite covering of Z.

b) Put

$$\mathscr{B} = \{(\tilde{B}\times A_y)\cup B\colon\ \emptyset \neq B\in\mathscr{B}_Z,\ 0 < y\in\mathbf{R},$$

$$\tilde{B} = (\inf B,\ \sup B),\ A_y = (-y, 0)\cup(0, y)\}\cup$$

$$\cup\{(a, b)\times(c, d)\colon\ a, b, c, d\in\mathbf{R},\ a < b,\ c < d,\ 0\notin(c, d)\}.$$

$\mathscr{B}$ is a base of X and $\mathscr{B}|Z = \mathscr{B}_Z$. It is easy to see that if G is a non-empty open subset of X and $B\in\mathscr{B}$, then

$$(1)\qquad G\subset B,\ \operatorname{Fr} G\subset \operatorname{Fr} B \Rightarrow G = B.$$

c) Suppose now that X is A-TPC. Then $\mathscr{B}$ has a locally finite open tight refinement $\mathscr{U}$ covering X. According to (1), $\mathscr{U}\subset\mathscr{B}$, so $\mathscr{U}|Z$ is a locally finite open covering of Z with $\mathscr{U}|Z\subset\mathscr{B}_Z$, in contradiction with a). Thus X is not A-TPC.

1M Example. *An* O-TPC *but not* σ-TPC *space*. Take the set[6]

$$X=(\omega_1+1)\times 2-\{\langle\omega_1,1\rangle\}$$

and let

$$\mathcal{B}=\{(\omega_1+1)\times 1\}\cup\{\{\alpha\}\times 2\colon\ \alpha\in\omega_1\}$$

be a base for the topology we introduce on X.

a) Note that every base of X has to contain $\mathcal{B}$, so $\mathcal{B}$ is an open tight refinement of an arbitrary base of X. Moreover, if $\mathcal{B}$ is ordered in such a way that $(\omega_1+1)\times 1$ is the first element, then for an arbitrary $B\in\mathcal{B}$, the elements of $\mathcal{B}$ preceding B form a locally finite system at the points of B, thus X is O-TPC.

b) Take the base $\mathcal{B}$ of X. $\mathcal{B}$ is not σ-locally finite at the point $\langle\omega_1,0\rangle$. Observe that an arbitrary open refinement of $\mathcal{B}$ covering X is equal to $\mathcal{B}$. Thus X is not σ-TPC.

1N Example. *A* C-TPC *but not* O-TPC *space*. Let X be the set ω_1. Consider X with the topology induced by the base $\mathcal{B}_0=\omega_1$.

a) For an arbitrary base $\mathcal{B}$ of $X, \{X\}$ is a covering of X as required in Definition 1G, so X is C-TPC.

b) Take the base $\mathcal{B}_0$ of X. An arbitrary open refinement $\mathcal{V}$ of $\mathcal{B}_0$ covering X is a subsystem of $\mathcal{B}_0$ of cardinality ω_1. In whatever way you order $\mathcal{V}$, it has to contain an element V and an infinite subsystem $\mathcal{V}_0$ of $\mathcal{V}$ such that each element of $\mathcal{V}_0$ precedes V. Now $\mathcal{V}_0$ is not locally finite at $0\in V$, thus X is not O-TPC.

1O Example. *A* D-TPC *but not* C-TPC *space*. Let ξ be a limit ordinal and take the set

$$X=\omega_1\times(\xi+1) \tag{1}$$

with the topology induced by the base

$$\mathcal{B}=\mathcal{B}_1\cup\mathcal{B}_2 \tag{2}$$

where

$$\mathcal{B}_1=\{\{\langle\alpha,\delta\rangle\}\colon\ \alpha\in\omega_1,\ \delta\in\xi\}, \tag{3}$$

$$\mathcal{B}_2=\{B_\alpha^\delta\colon\ \alpha\in\omega_1,\ \delta\in\xi\} \tag{4}$$

and

$$B_\alpha^\delta=\{\langle\alpha,\xi\rangle\}\cup\{\langle\beta,\gamma\rangle\colon\ \beta\in\alpha+1,\ \delta\in\gamma\in\xi\}\quad(\alpha\in\omega_1,\ \delta\in\xi). \tag{5}$$

Put

$$L=\{\langle\alpha,\xi\rangle\colon\ \alpha\in\omega_1\}. \tag{6}$$

Clearly, $\cup\mathcal{B}_1=X-L$, and the points of $X-L$ are isolated in X.

$$\mathcal{U}_\alpha=\{B_\alpha^\delta\colon\ \delta\in\xi\} \tag{7}$$

[6] An ordinal number α is always regarded as the set of ordinal numbers smaller than α. In this example (and in Examples 1N, 1O and 1V as well), the ordinal numbers are considered without the order topology.

is a neighbourhood base of the point $\langle\alpha, \xi\rangle\in L$. Observe that for a set $U\in\mathscr{U}_\alpha$ $(\alpha\in\omega_1)$, we have

$$\operatorname{Fr} U = L-\{\langle\alpha, \xi\rangle\} \quad\text{and}\quad \overline{U} = U\cup L. \tag{8}$$

a) Consider now this space X with[7] $\xi=\omega_1$. It will be shown that X is D-TPC. Let $\mathscr{G}$ be a base of X. For each $\alpha\in\omega_1$, there is a set $G_\alpha\in\mathscr{G}$ such that

$$B_\alpha^{\delta(\alpha)} \subset G_\alpha \subset B_\alpha^{\varepsilon(\alpha)} \tag{9}$$

with some $\delta(\alpha)\in\omega_1$ and $\varepsilon(\alpha)\in\omega_1$. (8) implies

$$\operatorname{Fr} G_\alpha = L-\{\langle\alpha, \omega_1\rangle\},\quad \overline{G}_\alpha = G_\alpha\cup L. \tag{10}$$

Put

$$\mathscr{H} = \{G_\alpha\colon\ \alpha\in\omega_1\} \tag{11}$$

and take the closed covering

$$\mathscr{F} = \overline{\mathscr{H}}\cup\{\{x\}\colon\ x\in X-\cup\overline{\mathscr{H}}\}. \tag{12}$$

Obviously, $\mathscr{F}$ is a refinement of $\overline{\mathscr{G}}$. For each $F\in\mathscr{F}$, $\operatorname{Fr} F$ is a (perhaps empty) subset of L (cf. (10)). Thus $\operatorname{Fr} F$ is dominated by the closed covering

$$\{\operatorname{Fr} G_0\cap\operatorname{Fr} F,\ \operatorname{Fr} G_1\cap\operatorname{Fr} F\},$$

and this covering of $\operatorname{Fr} F$ refines $\operatorname{Fr}\mathscr{G}$. To prove that X is D-TPC, it is enough to show that X is dominated by $\mathscr{F}$.

b) Let $\mathscr{F}_0$ be a subsystem of $\mathscr{F}$ and $A\subset\cup\mathscr{F}_0$. Suppose that A is not closed in X. Then there is a point

$$x = \langle\alpha, \omega_1\rangle\in\overline{A}-A.$$

Since x is a limit point of A, there is a $\beta\in\alpha+1$ such that

$$A\cap(\{\beta\}\times\omega_1)$$

is uncountable (because (7) is a neighbourhood base of x and α is countable). So, by (9)—(12), $A\cap G_\gamma$ is uncountable for each $\beta\in\gamma+1\in\omega_1$ as well. At least one of the sets $\overline{G}_\gamma$ $(\beta\in\gamma+1\in\omega_1)$ has to belong to $\mathscr{F}_0$, as the other elements of $\mathscr{F}$ cannot cover A (see (12)). For such a $\overline{G}_\gamma$, $\overline{G}_\gamma\cap A$ is not closed. Hence X is dominated by $\mathscr{F}$.

c) It will be proved now that X is not C-TPC. Let $\mathscr{B}$ be the base defined in (2)—(5). Assume $\mathscr{D}$ is a refinement of $\overline{\mathscr{B}}$ covering X. It is enough to show that $\mathscr{D}$ is not locally finite.

d) Suppose that $\mathscr{D}$ is locally finite at $\langle 0, \omega_1\rangle$. Then there is a set $D_0\in\mathscr{D}$ with $D_0\cap(\{0\}\times\omega_1)$ uncountable and an $\alpha_0\in\omega_1$ such that $D_0\subset B_{\alpha_0}^0\in\mathscr{B}_2$. If $\mathscr{D}$ is locally finite at $\langle\alpha_0+1, \omega_1\rangle$, then there is a set $D_1\in\mathscr{D}$ with $D_1\cap(\{\alpha_0+1\}\times\omega_1)$ uncountable and an $\alpha_1\in\omega_1$ such that $D_1\subset B_{\alpha_1}^0\in\mathscr{B}_2$. Clearly, $D_1\neq D_0$. In this way, we define by induction a sequence $\{\alpha_n\}_{n\in\omega}$ of countable ordinal numbers and a sequence $\{D_n\}_{n\in\omega}$ of different elements of $\mathscr{D}$. Let $\beta\in\omega_1$ be greater than each α_n $(n\in\omega)$. Now $\{D_n\colon\ n\in\omega\}$ is evidently not locally finite at $\langle\beta, \omega_1\rangle$. Thus X is not C-TPC.

[7] The same construction with $\xi=\omega$ will be used in Example IV.

1P Now we shall give a characterization of C-TPC spaces. This characterization will enable us to define another generalization of C-TPC, different from D-TPC.

1Q Proposition. *A space X is* C-TPC *iff for each base $\mathscr{B}$ of X there is a locally finite covering $\mathscr{G}\cup\mathscr{F}$ of X such that*

(i) *$\mathscr{G}$ is a disjoint open refinement of $\mathscr{B}$;*

(ii) *$\mathscr{F}$ is a closed refinement of* $\operatorname{Fr}\mathscr{B}$;

(iii) $\bigcup\mathscr{G}\cap\bigcup\mathscr{F}=\emptyset$.

1R Remarks. a) This characterization of C-TPC seems to be simpler than the original definition (i.e. 1G combined with 1F).

b) Condition (iii) can be dropped, since if $\mathscr{G}\cup\mathscr{F}$ satisfies (i) and (ii), then

$$\mathscr{G}\cup[\mathscr{F}|(X-\bigcup\mathscr{G})]$$

satisfies (iii) as well. (iii) has been included in order to simplify the structure of $\mathscr{G}\cup\mathscr{F}$.

1S Proof of Proposition 1Q. a) Assume that for a base $\mathscr{B}$ of X, there is a covering $\mathscr{G}\cup\mathscr{F}$ satisfying (i)—(iii). Then $\mathscr{H}=\overline{\mathscr{G}}\cup\mathscr{F}$ is a locally finite closed covering of X refining $\overline{\mathscr{B}}$. If $H\in\mathscr{F}$, then $\{\operatorname{Fr} H\}$ is, according to (ii), a locally finite closed covering of $\operatorname{Fr} H$ refining $\operatorname{Fr}\mathscr{B}$. On the other hand, if $H\in\overline{\mathscr{G}}$, then $\operatorname{Fr} H\subset\bigcup\mathscr{F}$ (because of (i) and since $\mathscr{G}\cup\mathscr{F}$ is a covering of X), thus $\mathscr{F}|\operatorname{Fr} H$ is a locally finite closed covering of $\operatorname{Fr} H$ refining $\operatorname{Fr}\mathscr{B}$ (cf. (ii)). Thus X is C-TPC.

b) Suppose now that X is C-TPC and let $\mathscr{B}$ be a base of X. Then X has a well-ordered locally finite closed covering

$$\mathscr{C}=\{C_\alpha\colon\ \alpha<\gamma\}$$

such that

(1) $$\forall\alpha<\gamma\ \ \exists B_\alpha\in\mathscr{B},\quad C_\alpha\subset\overline{B}_\alpha$$

and $\mathscr{C}$ is a boundary-refinement of $\mathscr{B}$, i.e. for each $\alpha<\gamma$, $\operatorname{Fr} C_\alpha$ has a locally finite closed covering $\mathscr{K}_\alpha$ refining $\operatorname{Fr}\mathscr{B}$. Take now

$$\mathscr{G}=\{G_\alpha\colon\ \alpha<\gamma\}$$

where

(2) $$G_\alpha=(B_\alpha\cap\operatorname{int} C_\alpha)-\bigcup_{\beta<\alpha}\overline{G}_\beta\quad(\beta<\gamma)$$

and

(3) $$\begin{cases}\mathscr{F}=\mathscr{F}_1\cup\mathscr{F}_2,\quad \mathscr{F}_1=\bigcup\limits_{\alpha<\gamma}\mathscr{K}_\alpha,\\ \mathscr{F}_2=\{C_\alpha\cap\operatorname{Fr} B_\alpha\colon\ \alpha<\gamma\}.\end{cases}$$

It will be proved that $\mathscr{G}\cup\mathscr{F}$ is a locally finite covering of X satisfying (i) and (ii). Thus, by 1Rb), the proof of Proposition 1Q will be complete.

c) Assume

(4) $$x\in X-\bigcup(\mathscr{G}\cup\mathscr{F}_2).$$

As $\mathscr{C}$ covers X, there is a smallest α such that $x\in C_\alpha$. If $\beta<\alpha$, we have $x\notin\overline{G}_\beta$, since (2) implies

$$\overline{G}_\beta\subset\overline{\operatorname{int} C_\beta}\subset C_\beta.$$

Further, (4) gives $x \notin G_\alpha$, so

$$x \notin B_\alpha \cap \operatorname{int} C_\alpha. \tag{5}$$

On the other hand, by (3) and (4):

$$x \notin \operatorname{Fr} B_\alpha \cap \operatorname{int} C_\alpha. \tag{6}$$

According to (5), (6) and (1):

$$x \notin \bar{B}_\alpha \cap \operatorname{int} C_\alpha \supset C_\alpha \cap \operatorname{int} C_\alpha = \operatorname{int} C_\alpha,$$

thus $x \in \operatorname{Fr} C_\alpha$. $\mathscr{K}_\alpha$ covers $\operatorname{Fr} C_\alpha$, so $x \in \bigcup \mathscr{F}_1$, i.e. $\mathscr{G} \cup \mathscr{F}$ covers X.

d) $\mathscr{G}$ is locally finite, since $G_\alpha \subset C_\alpha$ $(\alpha < \gamma)$ and $\mathscr{C}$ is locally finite. For every $\alpha < \gamma$, $\mathscr{K}_\alpha$ is locally finite, and

$$\bigcup \mathscr{K}_\alpha \subset C_\alpha \quad (\alpha < \gamma),$$

thus

$$\{\bigcup \mathscr{K}_\alpha \colon \alpha < \gamma\}$$

is locally finite as well, and so is $\mathscr{F}_1$, too. Moreover, $\mathscr{F}_2$ is evidently locally finite.

e) It is clear from (2) that $\mathscr{G}$ satisfies (i) (observe that each G_α is open for $\mathscr{G}$ is locally finite).

f) (3) guarantees that $\mathscr{F}$ satisfies (ii).

1T Definition. A space X is *weakly totally paracompact* (W-TPC) if for each base $\mathscr{B}$ of X there is a covering $\mathscr{G} \cup \mathscr{F}$ of X such that

(i) $\mathscr{G}$ is a disjoint open refinement of $\mathscr{B}$;

(ii) $\mathscr{F}$ is a locally finite closed refinement of $\operatorname{Fr} \mathscr{B}$;

(iii) $\bigcup \mathscr{G} \cap \bigcup \mathscr{F} = \emptyset$.

1U Remarks. a) The only difference between the definition of W-TPC and the characterization 1Q of C-TPC is that $\mathscr{G}$ is now not required to be locally finite.

b) An observation similar to 1Rb) holds here, too.

c) It is evident that each C-TPC space is W-TPC. On the other hand, there exists a W-TPC space which is not C-TPC (moreover, not even D-TPC).

1V Example. *A* W-TPC *but not* D-TPC *space.* Take the space X from Example 1O, but now with $\xi = \omega$. We shall use the notations introduced in 1O(1)—1O(8).

a) Let $\mathscr{E}$ be a base of X. Put

$$\mathscr{G} = \{\{x\} \colon x \notin L\}$$

and

$$\mathscr{F} = \{L - \{\langle 0, \omega\rangle\},\ L - \{\langle 1, \omega\rangle\}\}.$$

$\mathscr{G} \cup \mathscr{F}$ covers X. $\mathscr{G}$ evidently satisfies 1T(i) (with $\mathscr{E}$ instead of $\mathscr{B}$). Similarly to 1O(9)—1O(10), one can take sets E_0 and E_1 from $\mathscr{E}$ such that

$$\operatorname{Fr} E_i = L - \{\langle i, \omega\rangle\} \quad (i = 0, 1),$$

thus $\mathscr{F}$ is a locally finite closed system refining $\operatorname{Fr} \mathscr{B}$, and so X is W-TPC.

b) Take the base $\mathscr{B}$ of X defined by 1O(2) and let $\mathscr{F}$ be an arbitrary closed covering of X refining $\bar{\mathscr{B}}$. To prove that X is not D-TPC, it suffices to show that X cannot be dominated by $\mathscr{F}$.

c) We define series $x_i \in X-L$ and $F_i \in \mathscr{F}$ $(i \in \omega)$ by induction. Let x_0 be an arbitrary point of $\omega_1 \times \{0\}$ and $F_0 \in \mathscr{F}$ such that $x_0 \in F_0$. For $0 \neq n \in \omega$, the sets F_i $(i \in n)$ do not cover $\omega_1 \times \{n\}$ (since no countable refinement of $\overline{\mathscr{B}}$ covers $\omega_1 \times \{n\}$), so there is a point

$$x_n \in (\omega_1 \times \{n\}) - \bigcup_{i \in n} F_i$$

and a set $F_n \in \mathscr{F}$ such that $x_n \in F_n$. Now each F_k $(k \in \omega)$ contains only a finite number of the points x_n $(n \in \omega)$, thus (as X is a T_1-space) $A \cap F_k$ is closed for each $k \in \omega$, where

$$A = \{x_n \colon n \in \omega\}.$$

Further, the subsystem

$$\mathscr{F}_0 = \{F_k \colon k \in \omega\}$$

of $\mathscr{F}$ covers A. On the other hand, A is not closed in X (each point x_n can be written in the form $\langle \alpha(n), n \rangle$ where $\alpha(n) \in \omega_1$; take a $\beta \in \omega_1$ greater than each $\alpha(n)$; $\langle \beta, \omega \rangle$ is now a limit point of A). Thus $\mathscr{F}$ does not dominate X and X is not D-TPC.

1W Remarks. a) We do not know if every D-TPC space is W-TPC.

b) While Examples 1K and 1L are as good as possible (separable metric), Examples 1M, 1N, 1O and 1V are not Hausdorff spaces (1M and 1N are not even T_1-spaces). It seems to be interesting to look for examples satisfying better separation axioms.

1X Proposition. *A closed subspace of a regular hereditarily normal* W-TPC *space is* W-TPC *as well.*

1Y Remark. Similar statements hold for the other generalizations of TPC, see French [Fr1, Fr2].

1Z Proof of Proposition 1X. Let A be a closed subspace of the hereditarily normal space X and let $\mathscr{B}_A$ be a base of A. Then X has a base $\mathscr{B}$ such that $\mathscr{B}_A = \mathscr{B} | A$ and for each $B \in \mathscr{B}$,

$$(1) \qquad A \cap \operatorname{Fr} B = \operatorname{Fr}_A (A \cap B)$$

(this is quite obvious from the hereditary normality and regularity of X, but see also French [Fr2], Lemma 5). Take now a covering $\mathscr{G} \cup \mathscr{F}$ as in definition 1T. $(\mathscr{G} \cup \mathscr{F}) | A$ is a covering of A. $\mathscr{G} | A$ clearly satisfies 1T(i) (with $\mathscr{B}_A$ instead of $\mathscr{B}$). (1) guarantees that $\mathscr{F}$ satisfies 1T(ii). Thus A is W-TPC.

1AA Definition (Dowker [Do], see also [N1] and [P]). A space X is *totally normal* if it is normal and for every open subset G of X, there is a covering $\mathscr{G}$ of G such that $\mathscr{G}$ is locally finite in G and cozero in X.

1BB Every totally normal space is a hereditarily normal S_1-space[8] (for the proof, see [P]).

[8] A space X is an S_1-*space* if its T_0-reflexion is a T_1-space or, equivalently, if each open subset of X is the union of some closed subsets of X.

1CC Definition (Lifanov and Pasynkov [LP]). A space X is *pointwise totally normal*[9] if it is hereditarily normal and for every open subset G of X, there exists a point-finite covering $\mathcal{G}$ of G such that $\mathcal{G}$ is cozero in X.

1DD Clearly, every totally normal space is pointwise totally normal. Several theorems on totally normal spaces hold for pointwise totally normal spaces as well[10], see Lifanov and Pasynkov [LP] and French [Fr2]. In particular, every pointwise totally normal space is an S_1-space and a "locally finite closed sum theorem" holds for Ind in pointwise totally normal spaces.

1EE Definition. A space X is *strongly paracompact*[11] if an arbitrary open covering of X has a star-finite open refinement covering X (cf. [P] 2.2.8).

1FF Definition (Zarelua [Z2], see also [P] 2.2.11). A space X is *completely paracompact* if for each open covering $\mathcal{U}$ of X there is a countable collection $\{\mathcal{V}_i : i=1, 2, \ldots\}$ of star-finite[12] open coverings of X such that $\bigcup_{i=1}^{\infty} \mathcal{V}_i$ has a subcovering refining $\mathcal{U}$.

1GG Proposition. *Every completely paracompact space is* σ-TPC.[13]

1HH Remark. Fitzpatrick and Ford [FF] proved the weaker statement that each strongly paracompact space is O-TPC. The proof of 1GG is essentially the same.

1II Example. *A* TPC *but not completely paracompact space.* Let X be the hedgehog space (star-space) with ω_1 spines. X is metrizable and not completely paracompact[14] (see [P] 2.3.16).

Let now $\mathcal{B}$ be a base of X and $B_0 \in \mathcal{B}$ a set containing the centre point of X. Each spine is a copy of the interval $(0, 1]$. There is an $\varepsilon > 0$ such that B_0 contains $(0, \varepsilon)$ on each spine. The elements of $\mathcal{B}$ lying in $I = (\varepsilon/2, 1]$ (on a fixed spine) form a base for I; I is TPC, so there is a locally finite subsystem of $\mathcal{B}$ which is a covering of I. Take now such a collection for each spine and add B_0; in this way we get a locally finite subsystem of $\mathcal{B}$ covering X, thus X is TPC.

§ 2. On the coincidence of ind and Ind

The coincidence of ind and Ind was first proved for spaces which are

(i) compact metrizable (Brouwer [B], 1924).

This result was later extended to larger classes of spaces:

(ii) separable metrizable (Hurewicz [H], 1927);

[9] Such a space is called a *Dowker space* in [LP] and French [Fr2]. Our terminology is motivated by the expression "pointwise paracompact" sometimes used instead of "metacompact".

[10] *Added in proof.* Throughout this paper, "pointwise totally normal" could be replaced by the more general notion *strongly hereditarily normal* introduced in Engelking's *Dimension Theory* (PWN—North-Holland, 1978).

[11] Or: X has the *star-finite property*.

[12] It does not change the definition if star-countable is substituted for star-finite.

[13] *Added in proof.* 2.4. C. (b) in Engelking's *Dimension Theory*.

[14] A completely paracompact metrizable space is often called *strongly metrizable*.

(iii) strongly paracompact metrizable (Morita [Mo], 1950);
(iv) completely paracompact metrizable (Zarelua [Z1, Z2], 1961);
(v) completely paracompact, totally normal (essentially contained in Zarelua [Z1, Z2]);[1]
(vi) TPC metrizable (Ford [F], 1963);[2]
(vii) O-TPC metrizable (Fitzpatrick and Ford [FF], 1967);
(viii) σ-TPC, totally normal (Nagami [N1], 1969);
(ix) C-TPC, totally normal (French [Fr1], 1972);
(x) D-TPC, pointwise totally normal (French [Fr2], 1975).

The following diagram shows the relations between these classes of spaces:

$$\begin{tikzcd}
 & & (vii) \arrow[r, Rightarrow] & (ix) \arrow[r, Rightarrow] & (x) \\
(ii) \arrow[d, Rightarrow] & (i) \arrow[l, Rightarrow] \arrow[r, Rightarrow] & (vi) \arrow[u, Rightarrow] \arrow[r, Rightarrow] & (viii) \arrow[u, Rightarrow] & \\
(iii) \arrow[r, Rightarrow] & (iv) \arrow[r, Rightarrow] \arrow[uur, Rightarrow] & (v) \arrow[ur, Rightarrow] & &
\end{tikzcd}$$

Thus (x) is the most general of them. In this section, we prove another generalization of (ix).

2A Theorem. *If the space X is* W-TPC *and pointwise totally normal, then* $\operatorname{ind} X = \operatorname{Ind} X$.

2B Proof. As pointwise totally normal spaces are S_1-spaces, it is enough to prove that

$$\operatorname{Ind} X \leqq \operatorname{ind} X. \tag{1}$$

We prove (1) by induction for $\operatorname{ind} X$. (1) is obvious if $\operatorname{ind} X = -1$. Suppose now that $n \geqq 0$, $\operatorname{ind} X = n$ and

$$\operatorname{ind} Y < n \Rightarrow \operatorname{Ind} Y < n \tag{2}$$

holds for each pointwise totally normal W-TPC space Y. X has now, by the definition of ind, a base $\mathscr{B}^*$ with

$$B \in \mathscr{B}^* \Rightarrow \operatorname{ind} \operatorname{Fr} B < n. \tag{3}$$

For each $B \in \mathscr{B}^*$, $\operatorname{Fr} B$ is a subspace of a pointwise totally normal space, so it is pointwise totally normal; $\operatorname{Fr} B$ is a closed subspace of a W-TPC hereditarily normal regular space (X is regular because it is an S_1-space), so $\operatorname{Fr} B$ is W-TPC, too (1X). Thus (2) and (3) give

$$B \in \mathscr{B}^* \Rightarrow \operatorname{Ind} \operatorname{Fr} B < n. \tag{4}$$

[1] Katuta [K], 1966 rediscovered a weaker form of Zarelua's result.

[2] It is only an inference that Ford's dissertation contains this theorem; we could not get hold of [F].

Take now a closed subset Φ and an open subset Γ of X such that $\Phi\subset\Gamma$. Take an open set V with $\Phi\subset V$ and $\overline{V}\subset\Gamma$ and let $\mathscr{B}\subset\mathscr{B}^*$ be a base of X refining the covering

$$\{\Gamma, X-\overline{V}\}. \tag{5}$$

Let $\mathscr{G}\cup\mathscr{F}$ be a covering of X as described in Definition 1T and put

$$W=\operatorname{int}[(\Gamma\cap\bigcup\mathscr{F})\cup\bigcup\{G\colon\ G\in\mathscr{G},\ G\subset\Gamma\}].$$

It is easy to check that $\Phi\subset W\subset\Gamma$ and

$$\operatorname{Fr} W\subset\bigcup\mathscr{F}. \tag{6}$$

By 1T(ii),

$$\mathscr{F}_0=\mathscr{F}|\operatorname{Fr} W \tag{7}$$

is a locally finite covering of Fr W. For each $F\in\mathscr{F}_0$, F is a closed subspace of some Fr B, where $B\in\mathscr{B}^*\subset\mathscr{B}$, so (4) implies

$$F\in\mathscr{F}_0\Rightarrow\operatorname{Ind} F<n. \tag{8}$$

Now $\mathscr{F}_0$ is a locally finite closed collection in a pointwise totally normal space, thus

$$\operatorname{Ind}\bigcup\mathscr{F}_0=\sup\{\operatorname{Ind} F\colon\ F\in\mathscr{F}_0\}$$

(French [Fr2], Lemma 8), and from (6), (7) and (8) we have $\operatorname{Ind}\operatorname{Fr} W<n$, so (1) has been proved.

2C Remark. One can similarly prove that $\dim X\leqq\operatorname{ind} X$ for an arbitrary normal W-TPC space X (use [P] 3.5.1 and 4.2.9; cf. the proof of [P] 4.5.8).

§ 3. Coincidence of ind and Ind in some spaces with a nice subbase

In order to assure the coincidence of ind and Ind in TPC spaces (or in some similar class of spaces), we need an additional property (total normality, for instance), since ind and Ind may differ even in compact T_2-spaces (Filippov [Fi1, Fi2], see also [P]). E. Deák [D4, D5] observed that another kind of additional property, namely the existence of an "ind-*nice base*" is sufficient to guarantee the coincidence of ind and Ind in TPC spaces. The idea of nice bases originates from Katětov [Kv], 1956 (although he does not use the word "nice", nor any denomination at all).

3A Definition (E. Deák [D4, D5]). Let d be a dimension function. A system $\mathscr{S}$ of subsets of a space X is d-*nice* if $\mathrm{d}X$ is finite and $\mathscr{A}\in[\mathscr{S}]^k$ implies $\mathrm{d}\cap\operatorname{Fr}\mathscr{A}\leqq\leqq\mathrm{d}X-k$ for $1\leqq k\leqq\mathrm{d}X+1$.

3B Remark. Our definition of nice systems is somewhat different from the original one, cf. [D*2] § 1.

3C We shall prove a generalization of a result by E. Deák [D4, D5]:

3D Theorem. a) [D*3] *If X is an* A-TPC S_1-*space with an* ind-*nice subbase, then* $\operatorname{ind} X=\operatorname{Ind} X$.

b) *If X is a normal* S_1*-space, each of its closed subspaces is* W-TPC *and X has an* ind*-nice subbase, then* $\operatorname{ind} X = \operatorname{Ind} X$.

3E Theorem 3Db) will be a corollary to an inequality we are going to prove (which will be applied to another problem in § 4). First of all, we define two dimension functions.

3F Definition [D*2]. For a space X, $\operatorname{sbd} X \leqq n$ if X has a subbase $\mathcal{S}$ with $\mathcal{A} \in [\mathcal{S}]^{n+1}$ implying $\overline{\bigcap \mathcal{A}} \subset \bigcup \mathcal{A}$ $(n = -1, 0, 1, \ldots)$.

3G Definition. For a space X, $\operatorname{ind}^\triangledown X = -1$ iff X is empty; $\operatorname{ind}^\triangledown X \leqq n$ $(n = 0, 1, 2, \ldots)$ iff X has a base $\mathcal{B}$ such that $\operatorname{ind}^\triangledown \bigcup \mathcal{C} < n$ for any locally finite closed refinement $\mathcal{C}$ of $\operatorname{Fr} \mathcal{B}$.

3H Theorem. *For an arbitrary space X,* $\operatorname{ind}^\triangledown X \leqq \operatorname{sbd} X$.

3I Remark. The proof of 3H is similar to that of [D*2] (3.3) or [D*3] (2.1), which are corollaries to the present theorem.

3J Proof of Theorem 3H. In order to prove the theorem, one has to show that

$$(1)\qquad \operatorname{sbd} X \leqq n \Rightarrow \operatorname{ind}^\triangledown X \leqq n \qquad (n = -1, 0, 1, \ldots).$$

Let n be a fixed integer, $n \geqq -1$; suppose $\operatorname{sbd} X \leqq n$ and let $\mathcal{S}$ be a subbase of X with

$$(2)\qquad \mathcal{A} \in [\mathcal{S}]^{n+1} \Rightarrow \overline{\bigcap \mathcal{A}} \subset \bigcup \mathcal{A}.$$

(cf. 3F). We consider the statements $(*k)$ for $0 \leqq k \leqq n+1$:

$$(*k)\quad \begin{cases} \textit{if } \mathcal{F} \textit{ is a locally finite collection of closed sets in } X \textit{ and if for each } F \in \mathcal{F} \\ \textit{there exists a system } \mathcal{A} = \mathcal{A}(\mathcal{F}) \in [\mathcal{S}]^k \textit{ with } F \subset \overline{\bigcap \mathcal{A}} - \bigcup \mathcal{A}, \textit{ then} \\ \operatorname{ind}^\triangledown \bigcup \mathcal{F} \leqq n-k. \end{cases}$$

To prove the theorem, it is enough to show that

$$(3)\qquad (*k+1) \Rightarrow (*k) \qquad (0 \leqq k \leqq n).$$

Indeed, $(*n+1)$ is an immediate consequence of (2), and, on the other hand, $(*0)$ implies $\operatorname{ind}^\triangledown X \leqq n$ (take the collection $\mathcal{F} = \{X\}$ and set $\mathcal{A}(X) = \emptyset$), thus (1) follows from (3).

To prove (3), suppose $(*k+1)$ holds (with $0 \leqq k \leqq n$ fixed) and let $\mathcal{F}$ be a collection satisfying the premises of the implication in $(*k)$. To verify (3), we have to prove

$$(4)\qquad \operatorname{ind}^\triangledown \bigcup \mathcal{F} \leqq n-k.$$

Further, (4) is true if

$$(5)\qquad \operatorname{ind}^\triangledown \bigcup \mathcal{L} \leqq n-k-1$$

is valid for an arbitrary locally finite closed refinement $\mathcal{L}$ of $\operatorname{Fr}_{\bigcup \mathcal{F}} \mathcal{B}$, where

$$(6)\qquad \mathcal{B} = \mathbf{B}\mathcal{S} \,|\, \bigcup \mathcal{F}$$

(cf. 3G; "locally finite" and "closed" are to be understood in $\bigcup\mathscr{F}$, but — as $\bigcup\mathscr{F}$ is closed — we may just as well say "locally finite closed refinement in X").

Now we shall prove (5). For each $L\in\mathscr{L}$, there is a set $B_L\in\mathscr{B}$ with

$$(7)\qquad L\subset \operatorname{Fr}_{\bigcup\mathscr{F}} B_L \quad (L\in\mathscr{L}).$$

Further, by (6), there is a finite subcollection $\mathscr{S}_L$ of $\mathscr{S}$ with

$$(8)\qquad B_L=\bigcap\mathscr{S}_L\cap\bigcup\mathscr{F}.$$

Take now the collection

$$(9)\qquad \mathscr{K}=\{L\cap\operatorname{Fr}_F(S\cap F)\colon F\in\mathscr{F}, L\in\mathscr{L}, S\in\mathscr{S}_L\}.$$

The proof of the theorem will be complete if we show

(a) $\bigcup\mathscr{L}=\bigcup\mathscr{K}$,

(b) *$\mathscr{K}$ is a locally finite collection of closed sets in X*

and

(c) *for each $K\in\mathscr{K}$, there is a collection $\mathscr{E}=\mathscr{E}(\mathscr{K})\in[\mathscr{S}]^{k+1}$ with $K\subset\overline{\bigcap\mathscr{E}}-\bigcup\mathscr{E}$.*

Indeed, (b), (c), $(*k+1)$ and (a) imply (5). The statements (a), (b) and (c) will be proved in the points a), b) and c) below.

a) By (9), every $K\in\mathscr{K}$ is a subset of some $L\in\mathscr{L}$, thus it is enough to show that an arbitrary point $x\in L\in\mathscr{L}$ belongs to $\bigcup\mathscr{K}$. $x\in L$, (7) and (8) imply

$$x\in\operatorname{Fr}_{\bigcup\mathscr{F}}\left(\bigcap\mathscr{S}_L\cap\bigcup\mathscr{F}\right),$$

so there is a set $S\in\mathscr{S}_L$ with

$$(\text{a}1)\qquad x\in\operatorname{Fr}_{\bigcup\mathscr{F}}\left(S\cap\bigcup\mathscr{F}\right).$$

Now (a1) guarantees that there is a set $F\in\mathscr{F}$ with

$$(\text{a}2)\qquad x\in\operatorname{Fr}_F(S\cap F)$$

(it is left to the reader to prove (a2). Thus

$$x\in L\cap\operatorname{Fr}_F(S\cap F)$$

with some $F\in\mathscr{F}$, $S\in\mathscr{S}_L$, and (a) has been proved.

b) One can verify (b) by a straightforward argument.

c) Take a set $K\in\mathscr{K}$ such that

$$(\text{c}1)\qquad K=L\cap\operatorname{Fr}_F(S\cap F)\neq\emptyset,\quad F\in\mathscr{F},\quad L\in\mathscr{L},\quad S\in\mathscr{S}_L.$$

We have supposed that $\mathscr{F}$ satisfies the premises of the implication in $(*k)$, thus there is a system $\mathscr{A}\in[\mathscr{S}]^k$ with

$$(\text{c}2)\qquad F\subset\overline{\bigcap\mathscr{A}}-\bigcup\mathscr{A}.$$

Put

$$\mathscr{E}=\mathscr{A}\cup\{S\}.$$

By (c1), $S\cap F\neq\emptyset$. Furthermore, (c2) gives $F\cap\bigcup\mathcal{A}=\emptyset$, thus $S\notin\mathcal{A}$ and $\mathcal{E}\in[\mathcal{S}]^{k+1}$. We are going to show that

(c3) $$K\subset\overline{\bigcap\mathcal{E}}-\bigcup\mathcal{E}$$

and this will prove (c). As $K\subset F$, (c2) implies $K\cap\bigcup\mathcal{A}=\emptyset$; further, $K\cap S=\emptyset$ (because S is open), so

(c4) $$K\cap\bigcup\mathcal{E}=\emptyset.$$

Take now a point $x\in K$ and let G be an arbitrary open neighbourhood of x in X. Then, according to (c1), there is a point

(c5) $$y\in S\cap F\cap G$$

and an open neighbourhood G_1 of y in X with

(c6) $$G_1\subset S\cap G.$$

From (c2) and (c5), we have a point

(c7) $$z\in\bigcap\mathcal{A}\cap G_1.$$

Now (c6) and (c7) give

$$z\in\bigcap\mathcal{A}\cap S\cap G=\bigcap\mathcal{E}\cap G.$$

As x was an arbitrary point of K and G was an arbitrary open neighbourhood of x in X, we have $K\subset\overline{\bigcap\mathcal{E}}$; this and (c4) give (c3), thus the proof of the theorem is complete.

3K Proof of Theorem 3Db). As X is an S_1-space, it is enough to prove that $\operatorname{Ind} X\leqq\operatorname{ind} X$. X has an ind-nice subbase, so by [D*2], Proposition (4.5), $\operatorname{sbd} X\leqq\leqq\operatorname{ind} X$. From the theorem above, we have $\operatorname{ind}^\triangledown X\leqq\operatorname{sbd} X$. Furthermore, $\operatorname{Ind} X\leqq\leqq\operatorname{ind}^\triangledown X$ follows from 4F and Proposition 4G.

3L Remark. In the proof of Theorem 3Db) (more precisely: in the proof of [D*2], Proposition (4.5)) we make use of somewhat less than the existence of an ind-nice subbase: it is enough to suppose that there is a subbase satisfying the condition in Definition 3A for $k=\operatorname{ind} X+1$ only. The same remark holds for Theorem 3Da).

§ 4. On the directional dimension

In this section, we prove another consequence of Theorem 3H. For the convenience of the reader, we begin with some definitions instead of just referring to their sources. All the definitions concerning directions and directional dimension are due to E. Deák [D1, D2].

4A Definition. A *direction* on a space X is a linearly ordered family $\mathcal{R}=(\mathcal{R},<)$ of pairs (G, F) with G an open and F a closed subset of X such that

(i) $G\subset F$ for each $(G, F)\in\mathcal{R}$;

(ii) $(G_1, F_1)<(G_2, F_2)$ implies $F_1\subset G_2$;

(iii) $\mathcal{G}(\mathcal{R})=\{G\colon \exists F, (G, F)\in\mathcal{R}\}$ contains the union of any subfamily of it;

(iv) $\mathcal{H}(\mathcal{R})=\{H\colon \exists G, (G, X-H)\in\mathcal{R}\}$ contains the union of any subfamily of it.

4B Definition. A *directional structure* on a space X is a system $\mathfrak{R}$ of directions on X. $\mathfrak{R}$ is *compatible* if

$$\bigcup\{\mathcal{G}(\mathcal{R})\cup\mathcal{H}(\mathcal{R})\colon\ \mathcal{R}\in\mathfrak{R}\}$$

is a subbase of X.

4C Definition. The *directional dimension* of a space X, denoted by $\operatorname{Dim} X$, is the minimum of the cardinalities of the compatible directional structures on X.

4D Remark. For further details on directional structures, see E. Deák [D2, D6]. Our terminology is somewhat different from that of E. Deák, but Definition **4C** is equivalent to the original one, cf. [D*1] (0.5).

4E Definition (Egorov and Podstavkin [EP]). For a space X, $\operatorname{Dind} X=-1$ iff X is empty; $\operatorname{Dind} X\leqq n$ $(n=0, 1, \ldots)$ iff each finite open covering of X has a (finite) disjoint open refinement $\mathcal{V}$ such that

$$\operatorname{Dind}(X-\bigcup\mathcal{V})<n.$$

4F If X is normal, then $\operatorname{Ind} X\leqq\operatorname{Dind} X$ [EP]; if X is perfectly normal, then $\operatorname{Ind} X=\operatorname{Dind} X$ [EP]. Kulpa [Ku] has proved that $\operatorname{Dind} X=\operatorname{Dind}\beta X$ for an arbitrary normal space X.

4G Proposition. *If each closed subspace of the space X is* W-TPC, *then* $\operatorname{Dind} X\leqq\operatorname{ind}^{\triangledown} X$.

4H Proof. (Induction for $\operatorname{ind}^{\triangledown}$.) 4G is evidently true for $\operatorname{ind}^{\triangledown} X=-1$. Assume $n\geqq 0$. Induction hypothesis:

$$\operatorname{ind}^{\triangledown} Y<n\Rightarrow\operatorname{Dind} Y<n \tag{1}$$

for any space Y with its closed subspaces W-TPC. Suppose now $\operatorname{ind}^{\triangledown} X=n$ and let $\mathcal{U}$ be a finite open covering of X and $\mathcal{B}$ a base of X such that

$$\operatorname{ind}^{\triangledown}\bigcup\mathcal{C}<n \tag{2}$$

for an arbitrary locally finite closed refinement $\mathcal{C}$ of $\operatorname{Fr}\mathcal{B}$. We may suppose without loss of generality that $\mathcal{B}$ refines $\mathcal{U}$. As X is W-TPC, there is a covering $\mathcal{G}\cup\mathcal{F}$ of X such that $\mathcal{G}$ is a disjoint open refinement of $\mathcal{B}$, $\mathcal{F}$ is a locally finite closed refinement of $\operatorname{Fr}\mathcal{B}$ and

$$\bigcup\mathcal{F}=X-\bigcup\mathcal{G}. \tag{3}$$

Now $\mathcal{G}$ refines $\mathcal{U}$ and, according to (2), $\operatorname{ind}^{\triangledown}\bigcup\mathcal{F}<n$. $\bigcup\mathcal{F}$ is a closed subspace of X, thus each closed subset of $\bigcup\mathcal{F}$ is W-TPC and (1) implies

$$\operatorname{Dind}\bigcup\mathcal{F}<n. \tag{4}$$

Proposition 4G follows from (3) and (4).

4I Definition (Mancuso [M]). For a space X, $\operatorname{indc} X=-1$ iff X is empty; $\operatorname{indc} X=n$ $(n=0, 1, \ldots)$ iff for each compact subset C of X and open subset G of X with $C\subset G$ there is an open set H such that $C\subset H\subset G$ and

$$\operatorname{indc}\operatorname{Fr} H<n.$$

4J Theorem. a) [D*2] *For an arbitrary space* X, indc $X \leqq \text{Dim } X$;
b) [D*3] *if* X *is* A-TPC, *then* Ind $X \leqq \text{Dim } X$;
c) *if each closed subspace of* X *is* W-TPC, *then* Dind $X \leqq \text{Dim } X$.

4K Remark. 4Ja) and b) generalize a theorem by E. Deák [D3]. 4Jc) is a generalization of the first part of Theorem (3.4) in [D*3].

4L Proof of Theorem 4Jc). By [D*2], Proposition (4.3), sbd $X \leqq \text{Dim } X$. According to Theorem 3H, $\text{ind}^{\triangledown} X \leqq \text{sbd } X$. Proposition 4G gives us Dind $X \leqq \text{ind}^{\triangledown} X$, thus Dind $X \leqq \text{Dim } X$.

4M Remark. A theorem similar to Theorem 4J holds for the half-directional dimension Dim introduced by the author [D*1], but only for S_1-spaces.

REFERENCES

[AP] Aleksandrov, P. S. and Pasynkov, B. A., *Introduction to dimension theory,* Nauka, Moscow, 1973 (in Russian).

[B] Brouwer, L. E. J., Bemerkungen zum natürlichen Dimensionsbegriff, *Proc. Acad. Amsterdam* **27** (1924), 635—638.

[CMMN] Corson, H. H., McMinn, T. J., Michael, E. A. and Nagata, J., Bases and local-finiteness, Preliminary report, *Notices AMS* **6** (1959), 814.

[D1] Deák, E., Eine vollständige Charakterisierung der Teilräume eines euklidischen Raumes mittels der Richtungsdimension, *Publ. Math. Inst. Hung. Acad. Sci.* **9** Ser A (1964), 437—464.

[D2] Deák, E., Theory and application of directional structures, *Topics in topology,* ed. by Á. Császár, North-Holland, 1974, 187—211.

[D3] Deák, E., Untersuchungen über Richtungsstrukturen, I. Weitere Beziehungen der Richtungsdimension zu den klassischen Dimensionen für gewisse Klassen topologischer Räume, *Studia Sci. Math. Hung.* **10** (1975), 435—458.

[D4] Deák, E., Über gewisse Verschärfungen der beiden klassischen induktiven topologischen Dimensionsbegriffe, I. Räume mit „hübschen" Basen, *Studia Sci. Math. Hung.* **11** (1976), 229—246.

[D5] Deák, E., Über gewisse Verschärfungen der beiden klassischen induktiven topologischen Dimensionsbegriffe, II. Offen-erblich total-parakompakte Räume, *Studia Sci. Math. Hung.* **11** (1976), 341—356.

[D6] Deák, E., *Dimension und Konvexität,* Akadémiai Kiadó, Budapest (to appear).

[D*1] Deák, J., A new characterization of the class of subspaces of a Euclidean space, *Studia Sci. Math. Hungar.* **11** (1976), 253—258.

[D*2] Deák, J., Subbase and dimension I, *Studia Sci. Math. Hungar.* **11** (1976), 389—397.

[D*3] Deák, J., Subbase and dimension II, *Studia Sci. Math. Hungar.* **14** (1979).

[Do] Dowker, C. H., Inductive dimension of completely normal spaces, *Quart. J. Math.* **4** (1953), 267—281.

[EP] Egorov, V. and Podstavkin, Ju., On a definition of dimension, *Soviet. Math.* **9** (1968), 188—191.

[F] Ford, R. M., *Basis properties in dimension theory,* Doctoral dissertation, Auburn Univ., Auburn, Ala., 1963.

[FF] Fitzpatrick, B. Jr. and Ford, R. M., On the equivalence of small and large inductive dimension in certain metric spaces, *Duke Math. J.* **34** (1967), 33—38.

[Fi1] Filippov, V. V., A bicompactum with non-coinciding dimensionalities, *Soviet. Math.* **10** (1969), 208—211.

[Fi2] Filippov, V. V., Solution of a problem of P. S. Aleksandrov, *Math. Sb. USSR* **12** (1970), 41—47.

[Fr1] French, J. A., Coincidence of small and large inductive dimension, *Topo 72, Lecture Notes* **378** (1974), 132—139.

[Fr2] French, J. A., Some completely normal spaces in which small and large inductive dimension coincide, *Houston J. Math.* **2** (1976), 181—193.

[H] HUREWICZ, W., Normalbereiche und Dimensionstheorie, *Math. Ann.* **96** (1927), 736—764.

[K] KATUTA, Y., A note on the inductive dimension of product spaces, *Proc. Japan Acad.* **42** (1966), 1011—1015.

[Ku] KULPA, W., A note on the dimension Dind, *Coll. Math.* **24** (1972), 181—183.

[Kv] KATĚTOV, M., On the dimension of non-separable spaces, *Czechoslovak. Math. J.* **6** (1956), 485—516 (in Russian).

[LP] LIFANOV, I. K. and PASYNKOV, B. A., On two classes of spaces and dimension, *Moscow Univ. Vestnik.* Ser II **25** (1970) No 3, 33—37 (in Russian).

[M] MANCUSO, V. J., Another inductive dimension, *Topo 72, Lecture Notes* **378** (1974), 267—270.

[Mo] MORITA, K., On the dimension of normal spaces II, *J. Math. Soc. Japan* **2** (1950), 16—33.

[N1] NAGAMI, K., A note on the large inductive dimension of totally normal spaces, *J. Math. Soc. Japan* **21** (1969), 282—290.

[N2] NAGAMI, K., *Dimension theory,* Academic Press, 1970.

[P] PEARS, A. R., *Dimension theory of general spaces,* Cambridge Univ. Press, 1975.

[Z1] ZARELUA, A. V., On a theorem of Hurewicz, *Dokl. AN SSSR* **141** (1961), 777—780 (in Russian).

[Z2] ZARELUA, A. V., On a theorem of Hurewicz, *Mat. Sb.* **60** (1963), 17—28 (in Russian).

Mathematical Institute of the Hungarian Academy of Sciences,
H—1053 Budapest, Reáltanoda u. 13—15

(Received December 12, 1978)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 411—422.

ON THE ORDER OF CONVERGENCE OF FINITE ELEMENT METHODS FOR THE NEUMANN PROBLEM

by
L. VEIDINGER

Finite element methods for Neumann type problems have been investigated recently by many authors (see, for example, [1]—[5]). It is well-known that for such problems the functions admissible in the associated variational problems are not required to satisfy any boundary conditions and hence regions in two or more dimensions of general shape may be treated without any difficulty. In the present paper we shall obtain error bounds for some finite element methods in two dimensions under weak regularity assumptions on the boundary of the region.

1. Let R be a bounded open plane region whose boundary C consists of a finite number of piecewise analytic simple closed curves. For the sake of simplicity we shall assume that the boundary C consists of two analytic arcs which meet at the corner $A=(0, 0)$ and form an interior angle $\pi\alpha$ $(0<\alpha<2)$ there. The general case can be treated in the same way.

We consider the Neumann problem

$$Lu(x, y) = g(x, y), \quad (x, y)\in R,$$

$$(1) \qquad \frac{\partial u(x, y)}{\partial N} = 0, \quad (x, y)\in C,$$

where

$$Lu \equiv \frac{\partial}{\partial x}\left[a(x, y)\frac{\partial u}{\partial x}\right]+\frac{\partial}{\partial x}\left[b(x, y)\frac{\partial u}{\partial y}\right]+\frac{\partial}{\partial y}\left[b(x, y)\frac{\partial u}{\partial x}\right]+$$

$$+\frac{\partial}{\partial y}\left[c(x, y)\frac{\partial u}{\partial y}\right]-f(x, y)u$$

and N is the conormal with direction cosines

$$\cos(N, x) = \frac{1}{E}[a\cos(n, x)+b\cos(n, y)], \quad \cos(N, y) = \frac{1}{E}[b\cos(n, x)+c\cos(n, y)].$$

Here n is the outward normal to C and

$$E = \{[a\cos(n, x)+b\cos(n, y)]^2+[b\cos(n, x)+c\cos(n, y)]^2\}^{1/2}.$$

At the corner A we require that

$$(2) \qquad \frac{\partial u}{\partial N_1} = \frac{\partial u}{\partial N_2} = 0,$$

where N_1 is the "left-hand conormal" and N_2 is the "right-hand conormal" to the boundary C.

Let the coefficients $a(x, y), b(x, y), c(x, y), f(x, y)$ and the right-hand side $g(x, y)$ be infinitely differentiable in R. Suppose that at all points of R

$$a\xi^2+2b\xi\eta+c\eta^2 \geqq \nu(\xi^2+\eta^2) \quad (\nu = \text{const.} > 0) \tag{3}$$

for all real ξ, η. Moreover, we assume that either $f(x, y) \geqq c_1 > 0$ or $f(x, y) \equiv 0$ (c_1 is a positive constant).

If $f(x, y) \geqq c_1 > 0$, then the Neumann problem has a unique solution $u(x, y)$. If $f(x, y) \equiv 0$ and

$$\iint_R g(x, y)\, dx\, dy = 0, \tag{4}$$

then the Neumann problem is solvable[1] and the solution is unique to within an additive constant. In this case we make the solution unique by requiring that

$$\iint_R u(x, y)\, dx\, dy = 0. \tag{5}$$

Let Ω be an open region in the plane of R. We denote by $W_2^{(s)}(\Omega)$ the Hilbert space of all functions which, together with their generalized partial derivatives up to the sth order, belong to $L_2(\Omega)$. The norm is given by

$$\|v\|^2_{s,\Omega} = \sum_{j=0}^{s} |v|^2_{j,\Omega}, \quad \text{where} \quad |v|^2_{j,\Omega} = \sum_{|i|=j} \|D^i v\|^2_{L_2(\Omega)}.$$

Here we use the notation $i=(i_1, i_2)$, $|i|=i_1+i_2$, $D^i=\dfrac{\partial^{|i|} v}{\partial x^{i_1} \partial y^{i_2}}$.

It is well-known that under the above assumptions the solution $u(x, y)$ of the Neumann problem (1) minimizes the functional

$$F(v) = \iint_R \left[a\left(\frac{\partial v}{\partial x}\right)^2 + 2b \frac{\partial v}{\partial x} \frac{\partial v}{\partial y} + c\left(\frac{\partial v}{\partial y}\right)^2 + fv^2 + 2gv \right] dx\, dy \tag{6}$$

in the space $W_2^{(1)}(R)$.

LEMMA 1. *Let $u(x, y)$ be the solution of the Neumann problem* (1) *and let*

$$x^* = k_A x + l_A y, \quad y^* = m_A x + n_A y \tag{7}$$

be a linear transformation which transforms the operator L into the normal form at the point $A=(0, 0)$. Let $r=\sqrt{x^2+y^2}$. If the transformation (7) *transforms the angle $\pi\alpha$ $(0<\alpha<2)$ into an angle $\pi\alpha^*$ $(0<\alpha^*<2)$*[2], *then for $n=0, 1, \dots$ we have*

$$u(x, y) = \sum_{0<\frac{m}{\alpha^*}+p\leqq n+1} a_{m,p}(\varphi)\, r^{\frac{m}{\alpha^*}+p} (\log r)^{q_{m,p}} + w(x, y), \tag{8}$$

[1] If $f(x, y) \equiv 0$ and $\iint_R g(x, y)\, dx\, dy \neq 0$, then the Neumann problem has no solutions.

[2] It is easy to show that α^* depends only on α, $a(A)$, $b(A)$, $c(A)$ and does not depend on the choice of the transformation (7).

where m is a positive integer, p and $q_{m,p}$ are nonnegative integers, the coefficients $a_{m,p}(\varphi)$ are infinitely differentiable functions of the polar coordinate $\varphi=\operatorname{arc\,tg}\frac{y}{x}$, $q_{1,0}=0$, if $\alpha^*\neq\frac{1}{s}$ (s an integer), $q_{1,0}=1$, if $\alpha^*=1, \frac{1}{2}, \dots, w(x,y)\in W_2^{(n+2)}(R)$.

This lemma follows from the results of KONDRAT'EV and WIGLEY (see [6]—[8]).

2. We cover R by a finite number of arbitrary real (non-curved) triangles T such that any two triangles are either disjoint or have a common vertex or a common side. We retain only those triangles T for which

$$\iint\limits_{T\cap R} dx\,dy > 0,$$

i.e. for which T and R have some common area. Denote by M_h the set of triangles covering R. To every set of triangles covering R we associate two parameters: h, ϑ. h is the largest side and ϑ is the smallest angle of all triangles T of the given set M_h (see Fig. 1). In the sequel we assume that

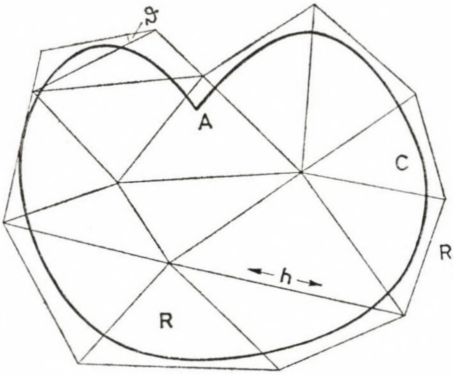


Fig. 1

$$h < c_2\bar{h}, \quad \vartheta \geqq \vartheta_0 > 0,$$

where $\bar{h}$ is the smallest side of all triangles $T\in M_h$, c_2 and ϑ_0 do not depend on h. Denote by R_h the union of all triangles $T\in M_h$. We emphasize that neither curved triangles nor special triangles near the boundary are necessary in the Neumann case; the triangles T are real and arbitrary. For example, we may choose a regular mesh consisting of right isosceles triangles (see Fig. 2); this mesh may be especially advantageous in practical computations.

Let $k\geqq 2$, $T\in M_h$ and let P_1, P_2, P_3 be the vertices of T. We choose the following nodes for T:

(a) the vertices of T,

(b) the $k-2$ points on each side of T that divide the side into $k-1$ equal segments, and

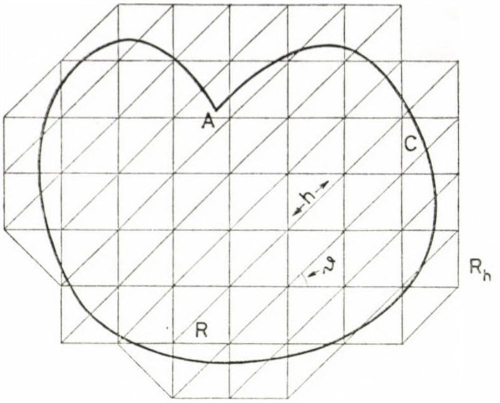


Fig. 2

(c) $\frac{1}{2}(k-3)(k-2)$ distinct points in the interior of T chosen so that if a polynomial of degree $k-4$ vanishes at all of them, then it vanishes identically.

Here (c) applies only to $k \geqq 4$ and (b) applies only to $k \geqq 3$. Denote by $P_1, P_2, \ldots, P_{3k-3+\frac{1}{2}(k-2)(k-3)}$ the nodes of T. Let

$$p_1 = p(P_1),\ p_2 = p(P_2),\ \ldots,\ p_{3k-3+\frac{1}{2}(k-2)(k-3)} = p\left(P_{3k-3+\frac{1}{2}(k-2)(k-3)}\right),$$

where $p(P_\mu)$ is any real-valued function defined at the nodes P_μ. Let $p(x, y)$ be the Lagrange interpolation polynomial of degree $\leqq k-1$ determined by its values p_μ at the points P_μ.

Denote by $H_{k-1}(R)$ the set of all functions (defined and continuous on R) which are equal on each triangle $T \in M_h$ to the corresponding polynomial $p(x, y)$. It is clear that $H_{k-1}(R) \subset W_2^{(1)}(R)$. The solution $u(x, y)$ of the Neumann problem (1) is approximated by the function $u_h(x, y)$ which minimizes the functional (6) in the space $H_{k-1}(R)$. If $f(x, y) \geqq c_1 > 0$, then the existence and uniqueness of $u_h(x, y)$ follows from the inequality (3). If $f(x, y) \equiv 0$, then $u_h(x, y)$ (exists and) is unique to within an additive constant. To make $u_h(x, y)$ unique we impose the condition

$$\iint_R u_h(x, y)\, dx\, dy = 0. \tag{10}$$

LEMMA 2. *Let $T \in M_h$, $v(x, y) \in W_2^{(k)}(T)$ and define the interpolate $p(x, y)$ of $v(x, y)$ by requiring that $v(x, y) - p(x, y)$ vanish at all the nodes. Then*

$$\|v-p\|_{1,T} \leqq c_3 h^{k-1} |v|_{k,T} \tag{11}$$

where c_3 is a positive constant which does not depend on the triangle T and the function $v(x, y)$.

For a proof, see [3], p. 146.

Lemma 3. *Let $T\in M_h$ and let $p(x, y)$ be the Lagrange interpolation polynomial of degree $\leqq k-1$ determined by the conditions $p(P_\mu)=p_\mu$ $(\mu=1, 2, \dots, 3k-3+ +0{,}5(k-2)(k-3))$. Then*

(12) $$|p|_{1,T} \leqq c_4 p_{\max}$$

and

(13) $$\|p\|_{L_2(T)} \leqq c_5 h p_{\max},$$

where $p \max\limits_{\mu=1,2,\dots,3k-3+\frac{1}{2}(k-2)(k-3)} |p_\mu|$, *$c_4$ and c_5 are positive constants which do not depend on the triangle T and the parameters p_μ.*

For a proof, see [9], p. 404.

Theorem 1. *Let $u(x, y)$ be the solution of the Neumann problem* (1) *(if $f(x, y)\equiv 0$, then we assume that $u(x, y)$ satisfies* (5)*) and let $u_h(x, y)$ be the function which minimizes the functional* (6) *in the space $H_{k-1}(R)$ (if $f(x, y)\equiv 0$, then we assume that $u_h(x, y)$ satisfies* (10)*). Then*

(14) $$\|u-u_h\|_{1,R} < c_6 h^{\beta^*}$$

and

(15) $$\max_{(x,y)\in R} |u(x, y)-u_h(x, y)| < c_7 h^{\beta^*} |\log h|^{1/2},$$

where

$$\beta^* = \begin{cases} \dfrac{1}{\alpha^*}, & \text{if } \dfrac{1}{2} < \alpha^* < 2 \quad \text{and} \quad \alpha^* \neq 1, \\ \dfrac{1}{\alpha^*} - \varepsilon, & \text{if } \alpha^* = 1, \dfrac{1}{2}, \\ 2, & \text{if } \dfrac{1}{k-1} \leqq \alpha^* < \dfrac{1}{2}, \\ k-1, & \text{if } 0 < \alpha^* < \dfrac{1}{k-1} \end{cases}$$

or if there are no corners, c_6 and c_7 are positive constants which depend only on k, the region R, the coefficients of the operator L and the right-hand side $g(x, y)$, ε is any positive real number.

Proof. Let the functional $D(v)$ be defined by

$$D(v) = \iint\limits_R \left[a\left(\frac{\partial v}{\partial x}\right)^2 + 2b\frac{\partial v}{\partial x}\frac{\partial v}{\partial y} + c\left(\frac{\partial v}{\partial y}\right)^2 + fv^2\right] dx\,dy$$

for all $v(x, y)\in W_2^{(1)}(R)$ and let $z(x, y)\in H_{k-1}(R)$. Then, using a standard technique (see, for example, [1], p. 6), we can easily prove that

(16) $$D(u-u_h) \leqq D(u-z).$$

If $f(x, y)\geqq c_1>0$, then from (16) it follows immediately that

(17) $$\|u-u_h\|_{1,R} \leqq c_8\|u-z\|_{1,R},$$

where c_8 is a positive constant which depends only on the coefficients of the operator L. If $f(x, y)\equiv 0$, then (17) follows from (16) by using the conditions (5), (10) and the Poincaré inequality (see, for example, [10], p. 488).

By Lemma 1 we have for $n=0, 1, \dots$

$$u(x, y) = \sum_{0<\frac{m}{\alpha^*}+p\leqq n+1} a_{m,p}(\varphi)\, r^{\frac{m}{\alpha^*}+p} (\log r)^{q_{m,p}} + w(x, y).$$

By the Calderón extension theorem (see, for example, [11], p. 171) there exists a function $w_{\text{ext}}(x, y)\in W_2^{(n+2)}(R_h)$ such that $w_{\text{ext}}(x, y)=w(x, y)$ for all $(x, y)\in R$ and

$$\|w_{\text{ext}}\|_{n+2, R_h} \leqq c_9 \|w\|_{n+2, R}, \tag{18}$$

where c_9 is a positive constant which depends only on the region R. Let the function $u_{\text{ext}}(x, y)$ be defined by

$$u_{\text{ext}}(x, y) = \sum_{0<\frac{m}{\alpha^*}+p\leqq n+1} a_{m,p}(\varphi)\, r^{\frac{m}{\alpha^*}+p} (\log r)^{q_{m,p}} + w_{\text{ext}}(x, y) \tag{19}$$

for all $(x, y)\in R_h$. We define the interpolate $\varrho_h u_{\text{ext}}(x, y)$ of $u_{\text{ext}}(x, y)$ by requiring that $u_{\text{ext}}(x, y)-\varrho_h u_{\text{ext}}(x, y)$ vanish at all the nodes. Then from (17) it follows that

$$\|u-u_h\|_{1, R} \leqq c_8 \|u_{\text{ext}}-\varrho_h u_{\text{ext}}\|_{1, R_h}. \tag{20}$$

Denote by $r(A, T)$ the distance from the corner A to the triangle $T\in M_h$. Let us first consider a triangle $T\in M_h$ for which $r(A, T)>r_1$, where r_1 is a fixed small positive real number. Setting $n=k$ in (19) we obtain from the Sobolev embedding theorem (see, for example, [11], p. 32) that if $T\in M_h$, $r(A, T)>r_1$, then

$$\max_{(x,y)\in T} |D^i u_{\text{ext}}(x, y)| < c_{10},$$

where $i=(i_1, i_2)$, $|i|=i_1+i_2=k$ and c_{10} is a positive constant which depends only on the data of the Neumann problem. Substituting this into (11) we have

$$\|u_{\text{ext}}-\varrho_h u_{\text{ext}}\|_{1, T}^2 < c_{11} h^{2k-2} \operatorname{mes} T, \tag{21}$$

where c_{11} is a positive constant which depends only on the data of the Neumann problem. Summing (21) over all triangles $T\in M_h$ for which $r(A, T)>r_1$ we obtain that

$$\sum_{T\in M_h, r(A,T)>r_1} \|u_{\text{ext}}-\varrho_h u_{\text{ext}}\|_{1, T}^2 = O(h^{2k-2}). \tag{22}$$

If $\frac{1}{k-1}<\alpha^*<2$ and $\alpha^*\neq 1, \frac{1}{2}, \dots, \frac{1}{k-2}$, then by Lemma 2 we have

$$\sum_{T\in M_h, h\leqq r(A,T)\leqq r_1} \|u_{\text{ext}}-\varrho_h u_{\text{ext}}\|_{1, T}^2 =$$

$$= O\Big(h^{2k-2} \sum_{T\in M_h, h\leqq r(A,T)\leqq r_1} (\operatorname{mes} T)[r(A, T)]^{\frac{2}{\alpha^*}-2k}\Big) = \tag{23}$$

$$= O\Big(h^{2k-2} \iint_{h\leqq r\leqq r_1} r^{\frac{2}{\alpha^*}-2k}\, dx\, dy\Big) = O\Big(h^{2k-2}\int_h^{r_1} r^{\frac{2}{\alpha^*}-2k+1}\, dr\Big) = O\big(h^{\frac{2}{\alpha^*}}\big),$$

where $r=\sqrt{x^2+y^2}$. If $\alpha^*=1, \frac{1}{2}, \dots, \frac{1}{k-1}$, then in a similar way we get

$$\sum_{T\in M_h, h\leqq r(A,T)\leqq r_1} \|u_{\text{ext}}-\varrho_h u_{\text{ext}}\|^2_{1,T} = O\big(h^{\frac{2}{\alpha^*}-\varepsilon}\big). \tag{24}$$

If $0<\alpha^*<\frac{1}{k-1}$, then setting $n=k$ in (19) we find that $u_{\text{ext}}(x,y)\in W_2^{(k)}(G_A)$, where $G_A=\{P\in R_h, 0\leqq r\leqq r_1\}$ and by Lemma 2 we have

$$\sum_{T\in M_h, h\leqq r(A,T)\leqq r_1} \|u_{\text{ext}}-\varrho_h u_{\text{ext}}\|^2_{1,T} = O(h^{2k-2}). \tag{25}$$

(23), (24) and (25) imply that

$$\sum_{T\in M_h, h\leqq r(A,T)\leqq r_1} \|u_{\text{ext}}-\varrho_h u_{\text{ext}}\|^2_{1,T} = O(h^{2\beta^*}). \tag{26}$$

Let us now consider a triangle $T\in M_h$ such that $r(A,T)\leqq h$. If $1<\alpha^*<2$, then from (19) it follows that

$$\begin{aligned} |u_{\text{ext}}-u(A)|^2_{1,T} &= O\Big(\iint\limits_{0\leqq r\leqq h} r^{\frac{2}{\alpha^*}-2}\,dx\,dy\Big) = \\ &= O\Big(\int_0^h r^{\frac{2}{\alpha^*}-1}\,dr\Big) = O\big(h^{\frac{2}{\alpha^*}}\big) \end{aligned} \tag{27}$$

and

$$\|u_{\text{ext}}-u(A)\|^2_{L_2(T)} = O\big(h^{\frac{2}{\alpha^*}+2}\big). \tag{28}$$

On the other hand, if we define the interpolate $p(x,y)$ of $u(x,y)$ by requiring that $u(x,y)-p(x,y)$ vanish at all the nodes, then by (12) and (19) we have

$$|u(A)-p|_{1,T} = O\big(h^{\frac{1}{\alpha^*}}\big). \tag{29}$$

Finally, from (13) and (19) we obtain that

$$\|u(A)-p\|_{L_2(T)} = O\big(h^{\frac{1}{\alpha^*}+1}\big). \tag{30}$$

(27), (28), (29) and (30) imply that if $T\in M_h$, $r(A,T)\leqq h$ and $1<\alpha^*<2$, then

$$\|u_{\text{ext}}-p\|_{1,T} = O\big(h^{\frac{1}{\alpha^*}}\big). \tag{31}$$

Let us now consider the case when $\frac{1}{k-1}\leqq\alpha^*<1$ and $\alpha^*\neq\frac{1}{2}$[3]. Then from Lemma 1 it follows that the partial derivatives $\frac{\partial u(A)}{\partial x}$ and $\frac{\partial u(A)}{\partial y}$ exist and, consequently,

[3] If $k=2$, then this possibility is precluded.

the conditions (2) can be written in the form

$$\begin{aligned} a(A)\frac{\partial u(A)}{\partial x}\sin\omega_1+b(A)\frac{\partial u(A)}{\partial x}\cos\omega_1+\\ +b(A)\frac{\partial u(A)}{\partial y}\sin\omega_1+c(A)\frac{\partial u(A)}{\partial y}\cos\omega_1=0 \end{aligned}\tag{32}$$

and

$$\begin{aligned} a(A)\frac{\partial u(A)}{\partial x}\sin\omega_2+b(A)\frac{\partial u(A)}{\partial x}\cos\omega_2+\\ +b(A)\frac{\partial u(A)}{\partial y}\sin\omega_2+c(A)\frac{\partial u(A)}{\partial y}\cos\omega_2=0, \end{aligned}\tag{33}$$

where ω_1 and ω_2 are the angles which are formed between the x-axis and the tangents to the boundary at the corner A. From (32) and (33) we obtain that $\frac{\partial u(A)}{\partial x}=$ $=\frac{\partial u(A)}{\partial y}=0$. Hence, using (19), we find that

$$\frac{\partial w_{\text{ext}}(A)}{\partial x}=\frac{\partial w_{\text{ext}}(A)}{\partial y}=0.\tag{34}$$

From (34), using Taylor's formula, it follows that

$$\max_{(x,y)\in T}\left|\frac{\partial w_{\text{ext}}(x,y)}{\partial x}\right|=O(h),\quad \max_{(x,y)\in T}\left|\frac{\partial w_{\text{ext}}(x,y)}{\partial y}\right|=O(h).\tag*{(35}$$

(19) and (35) imply that

$$\begin{aligned} |u_{\text{ext}}-u(A)|^2_{1,T}&=O\Big(\iint\limits_{0\leqq r\leqq h}\big[r^{\frac{2}{\alpha^*}-2}+h^2\big]\,dx\,dy\Big)=\\ &=O\Big(\int_0^h r^{\frac{2}{\alpha^*}-1}\,dr+R^4\Big)=O\big(h^{\frac{2}{\alpha^*}}+h^4\big). \end{aligned}\tag{36}$$

Further, since $u(A)=w(A)$, we obtain from (19) and (35) that

$$\|u_{\text{ext}}-u(A)\|^2_{L_2(T)}=O\big(h^{\frac{2}{\alpha^*}+2}+h^6\big).\tag{37}$$

On the other hand, using the relation $u(A)=w(A)$, Taylor's formula, (19), (35) and (12) we have

$$|u(A)-p|^2_{1,T}=O\big(h^{\frac{2}{\alpha^*}}+h^4\big),\tag{38}$$

where $p(x,y)$ is defined as in (29). Using (13) in place of (12) we obtain

$$\|u(A)-p\|^2_{L_2(T)}=O\big(h^{\frac{2}{\alpha^*}+1}+h^5\big).\tag{39}$$

(36), (37), (38) and (39) imply that

$$\|u_{\mathrm{ext}}-p\|_{1,T} = \begin{cases} O\left(h^{\frac{1}{\alpha^*}}\right), & \text{if } \dfrac{1}{2}<\alpha^*<1, \\ O(h^2), & \text{if } \dfrac{1}{k-1}\leqq\alpha^*<\dfrac{1}{2}. \end{cases} \tag{40}$$

For $\alpha^*=1, \dfrac{1}{2}$ we have

$$\|u_{\mathrm{ext}}-p\|_{1,T} = O\left(h^{\frac{1}{\alpha^*}-\varepsilon}\right), \tag{41}$$

where ε is any positive real number. Finally, if $0<\alpha^*<\dfrac{1}{k-1}$, then $u_{\mathrm{ext}}(x,y)\in W_2^{(k)}(G_A)$ and by Lemma 2 we obtain that

$$\|u_{\mathrm{ext}}-p\|_{1,T} = O(h^{k-1}). \tag{42}$$

It follows from (9), (31), (40), (41) and (42) that

$$\sum_{T\in M_h,\, 0<r(A,T)<h} \|u_{\mathrm{ext}}-\varrho_h u_{\mathrm{ext}}\|_{1,T}^2 = O(h^{2\beta^*}). \tag{43}$$

(22), (26) and (43) imply that

$$\|u_{\mathrm{ext}}-\varrho_h u_{\mathrm{ext}}\|_{1,R_h} = O(h^{\beta^*}).$$

Substituting this into (20) we get the inequality (14). The inequality (15) follows directly from (14) and a theorem of V. P. IL'IN (see [12], p. 101). This completes the proof of Theorem 1.

In the special case when $Lu\equiv\Delta u-u$ and the boundary is sufficiently smooth the estimate (15) can be significantly improved. Denote by $W_\infty^{(s)}(R)$ the Banach space of all functions which, together with their generalized partial derivatives up to the s-th order, belong to $L_\infty(R)$. The norm is given by

$$\|v\|_{s,\infty,R} = \max_{|i|\leqq s}\{\|D^i v\|_{L_\infty(R)}\},$$

where $i=(i_1, i_2)$, $|i|=i_1+i_2$. SCOTT has proved that if $Lu\equiv\Delta u-u$ and $u(x,y)\in W_\infty^{(k)}(R)$, then

$$\|u-u_h\|_{L_\infty(R)} < c_{12}\left\{\begin{matrix} h^2\,|\log h|, & k=2 \\ h^k, & k\geqq 3 \end{matrix}\right\}\|u\|_{k,\infty,R}, \tag{44}$$

where c_{12} is a positive constant which depends only on k and the region R (see [5]). The estimate (44) is probably valid for more general elliptic operators (see [13]) but the smoothness assumption on C cannot be relaxed. The experimental results due to STRANG and FIX (see [3], p. 268) indicate that the estimate (15) may be optimal in the case when the boundary C has corners.

3. If we want to compute also the partial derivatives of $u(x,y)$, then we have to use the Hermite interpolation polynomials introduced by KOUKAL (see [14]). As in the previous section, we cover R by a finite number of arbitrary real triangles. We assume now that the corner A is a vertex of a triangle T. Let P_1, P_2, P_3 be the vertices of the triangle $T\in M_h$ and let P_0 be the centre of gravity of T. Let $l\geqq 2$,

$p^i_\mu=p^i(P_\mu)$ $(\mu=0, 1, 2, 3)$, where $i=(i_1, i_2)$, p^i is any real-valued function defined at the points P_μ and let $p(x, y)$ be the Hermite interpolation polynomial of degree $\leqq 2l-1$ determined by the conditions

$$D^i p(P_j) = p^i_j, \quad |i| \leqq l-1, \quad j = 1, 2, 3,$$
$$D^i p(P_0) = p^i_0, \quad |i| \leqq l-2,$$

where $|i|=i_1+i_2$. Denote by $I_{2l-1}(R)$ the set of all functions defined on R which are equal on each triangle $T\in M_h$ to the corresponding polynomial $p(x, y)$. It is easy to show that $I_{2l-1}(R)\subset W_2^{(1)}(R)$. The solution $u(x, y)$ of the Neumann problem (1) is approximated by the function $v_h(x, y)$ which minimizes the functional (6) in the space $I_{2l-1}(R)$. If $f(x, y)\geqq c_1>0$ then the existence and uniqueness of $v_h(x, y)$ follows immediately from the inequality (3). If $f(x, y)\equiv 0$, then $v_h(x, y)$ (exists and) is unique to within an additive constant. To make $v_h(x, y)$ unique we impose the condition

$$\iint\limits_R v_h(x, y)\,dx\,dy = 0. \tag{45}$$

In the sequel we shall restrict ourselves to the case $l=2$ but our results can be probably generalized to arbitrary l (see [14]). We shall need the following two lemmas.

LEMMA 4. *Let $T\in M_h$, $v(x, y)\in W_2^{(s)}(T)$ $(s=3, 4)$ and let $p(x, y)$ be the cubic polynomial determined by the conditions*

$$p(P_\mu) = v(P_\mu) \quad (\mu = 0, 1, 2, 3),$$

$$\frac{\partial p(P_\mu)}{\partial x} = \frac{\partial v(P_\mu)}{\partial x}, \quad \frac{\partial p(P_\mu)}{\partial y} = \frac{\partial v(P_\mu)}{\partial y} \quad (\mu = 1, 2, 3).$$

Then

$$\|v-p\|_{1,T} < c_{13} h^{s-1} |v|_{s,T}$$

where c_{13} is a positive constant which does not depend on the triangle T and the function $v(x, y)$.

For a proof, see [4], p. 146.

LEMMA 5. *Let $T\in M_h$ and let $p^i_\mu=p^i(P_\mu)$, where $i=(i_1, i_2)$, $|i|=i_1+i_2\leqq 1$ and p^i is any real-valued function defined at the points P_μ $(\mu=0, 1, 2, 3)$.* Let $p(x, y)$ *be the cubic polynomial determined by the conditions*

$$p(P_\mu) = p^{(0,0)}_\mu \quad (\mu = 0, 1, 2, 3)$$

$$\frac{\partial p(P_\mu)}{\partial x} = p^{(1,0)}_\mu, \ \frac{\partial p(P_\mu)}{\partial y} = p^{(0,1)}_\mu \quad (\mu = 1, 2, 3).$$

Then

$$|p|_{1,T} \leqq c_{14}(p^{(0,0)}_{\max} + h p^{(1,0)}_{\max} + h p^{(0,1)}_{\max})$$

and

$$\|p\|_{L_2(T)} \leqq c_{15} h (p^{(0,0)}_{\max} + h p^{(1,0)}_{\max} + h p^{(0,1)}_{\max}),$$

where

$$p_{\max}^{(0,0)} = \max_{\mu=0,1,2,3} |p_\mu^{(0,0)}|, \quad p_{\max}^{(1,0)} = \max_{\mu=1,2,3} |p_\mu^{(1,0)}|, \quad p_{\max}^{(0,1)} = \max_{\mu=1,2,3} |p_\mu^{(0,1)}|,$$

c_{14} and c_{15} are positive constants which do not depend on the triangle T and the parameters p.

For a proof, see [9], p. 403.

Using Lemma 4 and Lemma 5 in place of Lemma 2 and Lemma 3, respectively, and repeating the proof of Theorem 1 (see also [9]) we obtain the following theorem.

THEOREM 2. *Let $u(x, y)$ be the solution of the Neumann problem* (1) *(if $f(x, y)\equiv 0$, then we assume that $u(x, y)$ satisfies* (5)*) and let $v_h(x, y)$ be the function which minimizes the functional* (6) *in the space $I_3(R)$ (if $f(x, y)\equiv 0$, then we assume that $v_h(x, y)$ satisfies* (45)*).*

Then

$$\|u-v_h\|_{1,R} < c_{16} h^{\gamma^*}$$

and

$$\max_{(x,y)\in R} |u(x, y)-v_h(x, y)| < c_{17} h^{\gamma^*} |\log h|^{\frac{1}{2}},$$

where

$$\gamma^* = \begin{cases} \dfrac{1}{\alpha^*}, & \textit{if} \quad \dfrac{1}{3} < \alpha^* < 2 \quad \textit{and} \quad \alpha^* \neq 1, \dfrac{1}{2}, \\ \dfrac{1}{\alpha^*}-\varepsilon, & \textit{if} \quad \alpha^* = 1, \dfrac{1}{2}, \dfrac{1}{3}, \\ 3, & \textit{if} \quad \alpha^* < \dfrac{1}{3} \textit{ or if there are no corners.} \end{cases}$$

c_{16} and c_{17} are positive constants which depend only on the region R, the coefficients of the operator L and the right-hand side $g(x, y)$, ε is any positive real number.

REFERENCES

[1] FRIEDRICHS, K. O. and KELLER, H. B., A finite difference scheme for generalized Neumann problems. *Numerical solution of partial differential equations, Proceedings of a Symposium held at the University of Maryland, ed. by J. H. Bramble,* Academic Press, New York, 1966.

[2] HILBERT, S., A mollifier useful for approximations in Sobolev spaces and some applications to approximating solutions of differential equations, *Math. Comp.* **27** (1973), 81—89.

[3] AUBIN, J. P., *Approximation of elliptic boundary-value problems,* Wiley—Interscience, New York, 1972.

[4] STRANG, G. and FIX, G., *An analysis of the finite element method,* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973.

[5] SCOTT, R., Optimal L^∞ estimates for the finite element method on irregular meshes, *Math. Comp.* **30** (1976), 681—697.

[6] KONDRAT'EV, V. A., Boundary value problems for elliptic equations in domains with conical or angular points (in Russian), *Trudy Moskov. Mat. Obšč.* **16** (1967), 209—292.

[7] WIGLEY, N. M., Asymptotic expansions at a corner of solutions of mixed boundary value problems, *J. Math. Mech.* **13** (1964), 549—576.

[8] WIGLEY, N. M., Mixed boundary value problems in plane domains with corners, *Math. Z.* **115** (1970), 33—52.

[9] VEIDINGER, L., On the order of convergence of a finite element scheme, *Acta Math. Acad. Sci. Hungar.* **25** (1974), 401—412.

[10] COURANT, R. and HILBERT, D., *Methoden der Mathematischen Physik, Vol.* II, Springer, Berlin, 1937.

[11] AGMON, S., *Lectures on elliptic boundary value problems,* D. Van Nostrand Company, Inc., Princeton, 1965.

[12] IL'IN, V. P., Some inequalities in function spaces and their application to the investigation of the convergence of variational processes (in Russian), *Trudy Mat. Inst. Steklov.* **53** (1959), 64—127.

[13] RANNACHER, R., Zur L^∞-Konvergenz finiter elemente beim Dirichlet problem, *Mat. Z.* **149** (1976), 69—77.

[14] KOUKAL, S., Piecewise polynomial interpolations in the finite element method, *Apl. Mat.* **18** (1973), 146—160.

Mathematical Institute of the Hungarian Academy of Sciences,
H—1053 Budapest, Reáltanoda u. 13—15

(Received March 14, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 423—427.

ON A TRANSFORMATION PROBLEM OF AUTONOMOUS DIFFERENTIAL SYSTEMS FROM STABILITY THEORETICAL VIEWPOINT

by

G. TÓTH

1. Introduction

In this note we study the triangularization problem of autonomous differential system

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, \dots, x_n) \\ &\dots \\ \dot{x}_n &= f_n(x_1, \dots, x_n) \end{aligned} \tag{1}$$

given on a domain $T \subseteqq \mathbf{R}^n$ where, from now on, we shall suppose that $f_1, \dots, f_n \in C^\alpha(T)$ for some fixed $\alpha \in N \cup \{\infty\}$. The word *triangularization* means that type of C^α-transformation for which the transformed system has a triangular form in the new variables $u_1, \dots, u_n$:

$$\begin{aligned} \dot{u}_1 &= g_1(u_1) \\ \dot{u}_2 &= g_2(u_1, u_2) \\ &\dots \\ \dot{u}_n &= g_n(u_1, \dots, u_n). \end{aligned} \tag{2}$$

The special feature of the transformed system is that its component equations are independent from each other, i.e. it can be solved successively by solving an autonomous and $n-1$ nonautonomous differential equations.

Our present aim is to point out some connections of this kind of reduction with stability properties of the induced dynamical system φ of (1). In this way we obtain some necessary conditions and negative results for the triangularizability. In the planar case, however, there are sufficient conditions, too (see [4]).

2. Preliminary notions and definitions

On the domain T the differential system induces a C^α-dynamical system $\varphi\colon \mathbf{R} \times T \to T$ for which we shall suppose that it is globally defined. The mapping φ has the following properties:

(i) $\varphi \in C^\alpha(\mathbf{R} \times T)$,

(ii) $\varphi(t, \varphi(s, p)) = \varphi(t+s, p)$ whatever $p \in T$ and $t, s \in R$,

(iii) $\varphi(0, \cdot) = \mathrm{id}_T$.

For an arbitrary set $\mathscr{I} \subseteqq R$ we shall employ the notation

$$\varphi(\mathscr{I}; p) = \{\varphi(t, p)\colon t \in \mathscr{I}\}.$$

The sets $\varphi(\mathbf{R}_+; p)$ or $\varphi(\mathbf{R}_-; p)$ or $\varphi(\mathbf{R}; p)$ are called positive or negative half-orbit or orbit starting from the point $p\in T$, respectively.

In the theory of dynamical systems certain special orbits play an important part. The one-point orbits are called critical and if the point $p\in T$ is not critical and $\varphi(\tau, p)=p$ for some $\tau>0$ the orbit $\varphi(\mathbf{R}: p)$ is called periodic. Critical points and periodic orbits are the constant and periodic solutions of the differential system (1), respectively.

With each $p\in T$ we associate the positive and negative limit sets $L_\varphi^+(p)$ and $L_\varphi^-(p)$, respectively, as follows:

$$L_\varphi^\eta(p) = \bigcap \{\overline{\varphi(\mathbf{R}_\eta; q)}\colon q\in\varphi(\mathbf{R}_\eta; p)\},\quad \eta\in\{+,-\}.$$

The common property of critical points and periodic orbits is that they satisfy the relation

$$\varphi(\mathbf{R}; p) \subseteqq L_\varphi^+(p). \tag{3}$$

An orbit $\varphi(\mathbf{R}, p)$ is called *Poisson stable* if (3) holds. Especially, in the planar case the critical points and periodic orbits are the only Poisson stable orbits (see [1]).

A set $M\subseteqq T$ is called positively or negatively invariant if $\varphi(\mathbf{R}_+; M)\subseteqq M$ or $\varphi(\mathbf{R}_-; M)\subseteqq M$ are valid, resp. The (positively and negatively) invariant subdomains of T permit to define restrictions of dynamical systems in the usual way.

Now let φ be a dynamical system on the domain $T\subseteqq\mathbf{R}^n$. We say that φ *has the property* A^s, $s\in N$, if there exist a compact positively invariant subset K of T and a finite sequence $p_1, \dots, p_{s+1}\in K$ of noncritical points such that

$$p_{i+1}\in L_\varphi^+(p_i)$$

for every $i=1, \dots, s$.

Finally, we give the definition of equivalence. Let φ and ψ be C^α-dynamical systems on the domains T and S, respectively. A C^α-diffeomorphism

$$U\colon T\to S$$

is called C^α-equivalence between φ and ψ if

$$U\big(\varphi(t, p)\big) = \psi\big(t, U(p)\big)$$

for every $t\in R$ and $p\in T$.

The mapping U is an ordinary transformation between the corresponding differential systems of φ and ψ, respectively.

A differential system (1) is called triangularizable if φ is C^α-equivalent to the C^α-dynamical system ψ induced by a triangularized differential system (2).

3. Necessary conditions for the triangularizability

Our further investigations are based on the following

LEMMA. *The differential system* (1) *can be transformed into the system* (2) *with some transformation*

$$U = (u_1, \dots, u_n)\colon T\to U(T)$$

if and only if

$$g_i(u_1, \dots, u_i) = \sum_{j=1}^{n} \frac{\partial u_i}{\partial x_j} f_j(x_1, \dots, x_n), \qquad i = 1, \dots, n \tag{4}$$

and

$$\det \frac{\partial(u_1, \dots, u_n)}{\partial(x_1, \dots, x_n)} \neq 0$$

are valid for the component functions $u_1, \dots, u_n$ *on* T.

PROOF. Simple calculation.

The main result of our paper can be stated as follows:

THEOREM. *Let us suppose that the induced dynamical system of the differential system* (1) *has the property* A^s. *Then a necessary condition of the triangularizability of* (1) *is*

$$s < n.$$

PROOF. Let us suppose that the differential system (1) can be transformed to a triangular form (2) by a C^α-transformation

$$U = (u_1, \dots, u_n) \colon T \to U(T)$$

and let $s \geqq n$, i.e. we can choose noncritical points $p_1, \dots, p_{n+1} \in T$ with

$$p_{i+1} \in L^+(p_i) \tag{5}$$

for every $i=1, \dots, n$.

Let us consider the first equation of the partial differential system (4) along $\varphi(R, p_1)$. Then

$$\dot{\tilde{u}}_1(t) = g_1(\tilde{u}_1(t))$$

where $\tilde{u}_1(t) = u_1(\varphi(t, p_1))$ for every $t \in \mathbf{R}$. So the following two cases can occur:

I. There exists a number $t^* \in \mathbf{R}$ for which $g_1(\tilde{u}_1(t^*)) = 0$. In this case the unicity of solutions implies that $\tilde{u}_1(t) = \tilde{u}_1(t^*) = c_1 \in \mathbf{R}$ for every $t \in \mathbf{R}$. Assumption (5) implies that $u_1(\varphi(t, p_i)) = c_1$ for every $t \in \mathbf{R}$ and $i = 1, \dots, n$.

II. $g_1(\tilde{u}_1(t)) \neq 0$ for every $t \in \mathbf{R}$. Then the first equation of (4) can be rewritten in the integral form

$$\int_{\tilde{u}_1(0)}^{\tilde{u}_1(t)} \frac{d\xi}{g_1(\xi)} = t. \tag{6}$$

Condition (5) implies that we can choose a sequence $(t_n)_{n \in \mathbf{N}} \subseteqq \mathbf{R}$ with $\lim_{n \to \infty} t_n = \infty$ so that $\lim_{n \to \infty} \varphi(t_n, p_1) = p_2$.

After performing the substitution $t_n = t$ in (6) we can take the corresponding limits which leads to the relation $\lim_{n \to \infty} g_1(\tilde{u}_1(t_n)) = 0$ and so $g_1(u_1(p_2)) = 0$. Therefore $u_1(\varphi(t, p_2)) = c_1 \in \mathbf{R}$ for each $t \in R$.

Summarizing the results of the two cases we can state that

$$u_1(\varphi(t, p_i)) = c_1$$

for every $t \in \mathbf{R}$ and $i \geqq 2$.

Repeating the previous reasonings for the second equation of (4) along the orbit $\varphi(\mathbf{R};\, p_2)$:

$$\dot{\tilde{u}}_2(t) = g_2(\tilde{u}_1(t), \tilde{u}_2(t)) = g_2(c_1, \tilde{u}_2(t))$$

where $\tilde{u}_2(t)=u_2(\varphi(t, p_2))$, $t\in R$, we obtain that

$$u_2(\varphi(t, p_i)) = c_2$$

for every $t\in\mathbf{R}$ and $i\geqq 3$.

After the n-th step we obtain that the transformation U maps the orbit $\varphi(\mathbf{R};\, p_{n+1})$ to a point $(c_1, \dots, c_n)$ which is contradiction.

Since every Poisson stable trajectory has the property A^s for any $s\in\mathbf{N}$ we obtain:

COROLLARY 1. *The differential system* (1) *is not triangularizable if its induced dynamical system has a noncritical Poisson stable orbit.*

The Poisson stability of periodic orbits implies a simple consequence as follows:

COROLLARY 2. *If the differential system* (1) *has a periodic solution then it is not triangularizable.*

There are many physical examples for oscillating systems which can be described by second-order autonomous differential equations. Perhaps the most important is the van der Pol equation

$$\ddot{x}+\mu(x^2-1)\dot{x}+x = 0, \quad \mu > 0$$

which has a unique limit circle (see [3]) and therefore it is not triangularizable.

There is another application of Theorem 1 which enlights the connection between the critical point set and the notion of triangularizability as follows:

THEOREM 2. *Let φ be the induced dynamical system of* (1) *on the domain T. Further let $K\subseteqq T$ be a compact positively invariant subset of T such that for every non critical point $p\in K$*

$$\operatorname{card}(L_\varphi^+(p)) \geqq 2.$$

Then one of the following statements is valid:

(α) *K consists of critical points or K contains a critical point set with cardinality $2^{\aleph_0}$;*

(β) *The differential system* (1) *is not triangularizable.*

PROOF. Let us suppose that there exists a noncritical point p_1 in K. From the compactness of K it follows that $L_\varphi^+(p_1)\subseteqq K$ and so $\operatorname{card}(L_\varphi^+(p_1))\geqq 2$. Since $L_\varphi^+(p_1)$ is connected (see [1]) either every point of $L_\varphi^+(p_1)$ is critical and so $\operatorname{card}(L_\varphi^+(p_1))=2^{\aleph_0}$ or there is a noncritical point $p_2\in L_\varphi^+(p_1)$.

The verification can be accomplished by a successive application of the previous reasonings. If, after the n-th step the second alternative is still valid then φ has the property A^n and hence the differential system (1) is not triangularizable.

EXAMPLE 1. Let us consider on the (canonically parametrized) two-dimensional torus T the differential system

$$\dot{x}_1 = 1, \quad \dot{x}_2 = \alpha$$

where $\alpha \in \mathbf{R}$ is an irrational number. Then every orbit is dense on T and hence the system is not triangularizable.

EXAMPLE 2. Let us consider the second-order half-linear equation

$$\ddot{x}+h(x, \dot{x}) = 0 \tag{7}$$

for which $h \in C^{\alpha}(\mathbf{R}^2)$ and

$$h(\lambda x_1, \lambda x_2) = \lambda h(x_1, x_2)$$

for every $\lambda \in \mathbf{R}$ and $(x_1, x_2) \in \mathbf{R}^2$.

If we perform the transformation

$$u_1(x_1, x_2) = \frac{x_1}{x_2},$$

$$u_2(x_1, x_2) = x_2,$$

we obtain the triangularized form

$$\dot{u}_1 = 1+u_1 h(u_1, 1),$$

$$\dot{u}_2 = -u_2 h(u_1, 1).$$

In spite of this fact the equation (7) may have periodic solution (if $h(x_1, x_2)=x_1$ for instance) but our transformation is not a diffeomorphism.

REFERENCES

[1] NIEMITZKIĬ, V. V., STEPANOV, V. V., *Qualitative Theory of Differential Equations,* Moscow, 1949.

[2] SAMOVOL, V. S., On the Reduction of Dynamical Systems into Triangular Form, *Diff. Ur.* **5** (1969), 1076—1083.

[3] STOUBLE RAIMOND, A., *Nonlinear Differential Equations,* New York, Toronto, London 1962.

[4] TÓTH, G., On the Triangularizability of Planar Orthogonal Differential Equations, *Period. Math. Hungar.* **8** (1977), 243—251.

Mathematical Institute of the Hungarian
Academy of Sciences, Budapest
Reáltanoda u. 13—15, Hungary 1053

(Received March 26, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 429—432.

OPERATORS WITH THE SPECTRAL DECOMPOSITION PROPERTY ARE DECOMPOSABLE

by

B. NAGY

In a recent paper [4] ERDÉLYI and LANGE have studied operators with the spectral decomposition property (abbreviated SDP), an apparent generalization of the decomposable operators in the sense of FOIAŞ [5]. In this note we prove that the two classes of operators actually coincide. This will answer Problems 4, 5 and 7 on p. 115 in [3], improve on a result of FOIAŞ [6, p. 1545], and provide, apparently, the most natural definition of a decomposable operator.

In what follows operator will mean a bounded linear operator on a Banach space X over the complex field $\mathbf{C}$. For any subset H of $\mathbf{C}$, $\bar{H}$ is its closure and H^c is its complement. For any operator T, $\sigma(T)$ and $\varrho(T)$ denote its spectrum and resolvent set, respectively, and $T|Y$ is its restriction to the T-invariant subspace (=closed linear manifold) Y. T is said to have the single-valued extension property (SVEP) if for any holomorphic function f the relation $(z-T)f(z)\equiv 0$ implies $f(z)\equiv 0$ on its domain. If T has the SVEP and x belongs to X, then $\varrho(x)$ denotes the unique maximal open set in C on which a (necessarily unique) holomorphic function $\bar{x}$ exists that satisfies

$$(z-T)\bar{x}(z) = x \quad \text{for} \quad z\in\varrho(x).$$

We set $\sigma(x)=\varrho(x)^c$ and $X(F)=\{x\in X\colon \sigma(x)\subset F\}$ for any $F\subset C$. A T-invariant subspace Y of X is said to be spectral maximal for T [2, p. 18] if for any T-invariant subspace Z of X the relation $\sigma(T|Z)\subset\sigma(T|Y)$ implies $Z\subset Y$.

Let n be a positive integer. The operator T is called n-decomposable if for any open covering $(G_1, \ldots, G_n)$ of $\sigma(T)$ there are spectral maximal subspaces $Y_1, \ldots, Y_n$ for T such that

$$X = Y_1+\ldots+Y_n,$$

$$\sigma(T|Y_i) \subset G_i \quad \text{for} \quad i = 1, \ldots, n.$$

T is called decomposable if T is n-decomposable for every positive integer n [2, p. 30]. Following Erdélyi and Lange [4], T is said to have the SDP if T satisfies the definition of 2-decomposability with "spectral maximal" replaced by "invariant".

THEOREM. *If the operator T has the SDP, then T is decomposable.*

PROOF. Erdélyi and Lange have shown that every operator with the SDP has the single-valued extension property [4, Theorem 8]. Now we prove that for any compact subset F of $\mathbf{C}$ the manifold $X(F)$ is closed in X and therefore, by [2, Proposition 1.3.8], $X(F)$ is a spectral maximal subspace for T.

Let $v \in F^c$ and let $3d$ denote the distance between v and F. Let $D(a, r)$ denote the open disk in C with center a and radius r, and set

$$G_1 = G_1(v) = \overline{D(v, d)}^c \quad \text{and} \quad G_2 = G_2(v) = D(v, 2d).$$

Since T has the SDP and $G_1 \cup G_2 = \mathbf{C}$, there are T-invariant subspaces $Y_i = Y_i(v)$ such that

$$\sigma(T|Y_i) \subset G_i \quad \text{for} \quad i = 1, 2, \tag{1}$$
$$X = Y_1 + Y_2.$$

Let $x \in X(F)$, then

$$x = y_1 + y_2 \quad \text{with} \quad y_i \in Y_\iota \quad (i = 1, 2).$$

Since $\sigma(x) \subset F$ and $\sigma(y_2) \subset \sigma(T|Y_2) \subset G_2$, we obtain that $\sigma(y_1) \subset F \cup G_2$. Hence

$$\bar{x}(z) = \bar{y}_1(z) + \bar{y}_2(z) \quad \text{for} \quad z \in F^c \cap G_2^c.$$

Let H be a bounded Cauchy domain [9, p. 288] such that $F \subset H$ and $\overline{H} \subset (\overline{G_2})^c$. Let B denote the positively oriented boundary of H and put $k = (2\pi i)^{-1}$. Then $\sigma(y_2) \subset G_2$ implies $\int\limits_B \overline{y_2}(z)\, dz = 0$, hence

$$x = k \int\limits_{|z| = |T|+1} (z-T)^{-1} x\, dz = k \int\limits_B \bar{x}(z)\, dz = k \int\limits_B \bar{y}_1(z)\, dz.$$

We shall show that $\overline{y_1}(z) \in Y_1$ for every $z \in B$. By (1) there exist mappings $g_i \colon B \to Y_i$ $(i=1, 2)$ such that

$$\bar{y}_1(z) = g_1(z) + g_2(z).$$

Applying $z-T$ we obtain

$$y_1 - (z-T)\, g_1(z) = (z-T) g_2(z) \in Y_1 \cap Y_2,$$

for the left side is in Y_1 and the right side is in Y_2. Since $z \in B$, z belongs to the principal component of the resolvent set $\varrho(T|Y_2)$. By a result of SCROGGS [8, Corollary 4.1], z also belongs to $\varrho(T|Y_1 \cap Y_2)$. Hence there exists

$$h(z) = (z - T|Y_1 \cap Y_2)^{-1} (z-T)\, g_2(z)$$

and belongs to $Y_1 \cap Y_2$. We have

$$(z - T|Y_2)\big(g_2(z) - h(z)\big) = 0.$$

Since $z - T|Y_2$ is injective, we obtain that

$$g_2(z) = h(z) \in Y_1.$$

Hence $\overline{y_1}(z) \in Y_1$ for every $z \in B$ and therefore $x \in Y_1 = Y_1(v)$. Consequently, we have

$$X(F) \subset \bigcap \{Y_1(v) \colon v \in F^c\}.$$

Denote the intersection on the right side by Y and let $y \in Y$. Then for every $v \in F^c$

$$\sigma(y) \subset \sigma\big(T|Y_1(v)\big) \subset G_1(v).$$

Hence $\sigma(y)\subset\bigcap\{G_1(v)\colon v\in F^c\}=F$ and $X(F)=Y$. Being closed, $X(F)$ is a spectral maximal subspace for T.

Now let (H_1, H_2) be an open covering of $\sigma(T)$ and let Y_1, Y_2 be T-invariant subspaces of X such that

$$F_i=\sigma(T|Y_i)\subset H_i \quad (i=1,2) \quad \text{and} \quad X=Y_1+Y_2.$$

Then $X(F_i)$ are spectral maximal subspaces such that $\sigma(T|X(F_i))=F_i$ and $Y_i\subset X(F_i)$ for $i=1,2$, hence $X=X(F_1)+X(F_2)$. Thus T is 2-decomposable, and a recent result of RADJABALIPOUR [7] yields that T is decomposable. The proof is complete.

Now let Z denote a family of closed subsets of $\mathbf{C}$, which contains the closures of all open subsets of $\mathbf{C}$. Weakening a concept due to APOSTOL ([1, p. 1495], cf. also [6, p. 1540]), we give the following

DEFINITION. The operator T has the spectral precapacity E with domain Z if E is a mapping of Z into the family of T-invariant subspaces of X such that (i) if (G_1, G_2) is an open covering of $\mathbf{C}$, then

$$X=E(\overline{G_1})+E(\overline{G_2}),$$

(ii) if F belongs to Z, then $\sigma(T|E(F))\subset F$.

The previous proof then yields the following partial extension of a result of FOIAŞ [6, p. 1545].

COROLLARY. *If T has the spectral precapacity E, with domain Z, then T is decomposable, and for any F in Z*

$$E(F)\subset X(F)=\bigcap\{E(\overline{G_1(v)})\colon v\in F^c\}$$

(here $G_1(v)$ means the same as in the proof above).

We note that if T is decomposable then the mapping E_0, defined by

$$E_0(F)=X(F) \quad \text{for } F \text{ closed in } \mathbf{C},$$

is a spectral precapacity. Actually, E_0 is the unique spectral capacity [6] possessed by T. However, even the identity operator I has different spectral precapacities. Indeed, for any closed set F define

$$\begin{aligned} E_1(F) &= \{0\} \quad \text{if 1 is a boundary point of } F,\\ &= X(F) \quad \text{otherwise.}\end{aligned}$$

Then I has the spectral precapacities E_k $(k=0, 1)$, both even satisfying the relation (cf. [6, p. 1545]) $E_k(F_1\cap F_2)=E_k(F_1)\cap E_k(F_2)$ for any closed sets F_1, F_2.

Acknowledgement. After completing the manuscript, the author learned that E. ALBRECHT (On decomposable operators, to appear in *Integral Equations and Operator Theory*) and R. LANGE (On generalization of decomposability, preprint) had also obtained the main result of this paper. The three proofs are different, and the present proof seems to be the simplest.

REFERENCES

[1] APOSTOL, C., Spectral decompositions and functional calculus, *Rev. Roum. Math. Pures Appl.* **13** (1968), 1481—1528.

[2] COLOJOARĂ, I. and FOIAŞ, C., *Theory of generalized spectral operators,* Gordon and Breach, New York, 1968.

[3] ERDÉLYI, I. and LANGE, R., *Spectral decompositions on Banach spaces,* Lecture Notes in Math., Springer-Verlag, Berlin, 1977.

[4] ERDÉLYI, I. and LANGE, R., Operators with spectral decomposition properties, *J. Math. Anal. Appl.* **66** (1978), 1—19.

[5] FOIAŞ, C., Spectral maximal spaces and decomposable operators in Banach spaces, *Arch. Math.* **14** (1963), 341—349.

[6] FOIAŞ, C., Spectral capacities and decomposable operators, *Rev. Roum. Math. Pures Appl.* **13** (1968), 1539—1545.

[7] RADJABALIPOUR, M., Equivalence of decomposable and 2-decomposable operators, *Pacific J. Math.* **77** (1978), 243—247.

[8] SCROGGS, J. E., Invariant subspaces of a normal operator, *Duke Math. J.* **26** (1959), 95—111.

[9] TAYLOR, A. E., *Introduction to functional analysis,* Wiley, New York, 1958.

Department of Mathematics, Faculty of Chemistry
University of Technology, H—1521 Budapest

(Received September 14, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 433—451.

RESIDUATION GROUPOIDS AND LATTICES

by

BRUNO BOSBACH

0. This paper is written in a selfcontained manner. No special knowledge is expected besides some familiarity with basic notions and rules of general algebra. In spite of this, however, since the results presented here have close connections to the theory of complementary semigroups it might be helpful to have a look at [2].

Let us call in this note NR-lattice (normally residuated lattice) every algebra $(L, \cap, \cup, *, :)$ which is a distributive lattice with minimum 1 with respect to $\cap$, $\cup$, a residuation groupoid (see section 1) with respect to $*$, $:$, and which in addition satisfies:

(N11) $a*(b\cup c) = (a*b)\cup(a*c)$ (N21) $(a\cup b):c = (a:c)\cup(b:c)$

(N12) $a*(b\cap c) = (a*b)\cap(a*c)$ (N22) $(a\cap b):c = (a:c)\cap(b:c)$

(N13) $(a\cap b)*c = (a*c)\cup(b*c)$ (N23) $a:(b\cap c) = (a:c)\cup(b:c)$

(N14) $(a\cup b)*c = (a*c)\cap(b*c)$ (N24) $a:(b\cup c) = (a:b)\cap(b:c)$

and (N) $a\cap b = \big(a:(b*a)\big)\cup\big(b:(a*b)\big) = \big((a:b)*a\big)\cup\big((b:a)*b\big).$

Obviously for the expert every normal complementary semigroup satisfies the laws of an NR-lattice but for the sake of selfcontainedness we should give some concrete examples here. Let us consider therefore as

Example 1. $R^{\geqq 0}$ with respect to max, min, and $a*b:=\max(a, b)-a=:b:a$.

Example 2. $R^{\geqq 0}$ with respect to max, min, and $a*b:=b$ if $a<b$ and $=0$ otherwise $=:b:a$.

Example 3. N with respect to LCM, GCD, and $a*b:=\mathrm{LCM}\,(a, b)/a=:b:a$.

Example 4. The set of all positive order automorphisms of $R\big(x<\alpha(x)\big)$ with respect to max, min, and $\alpha*\beta:=\alpha^{-1}\max(\alpha, \beta)$, $\beta:\alpha:=\max(\beta, \alpha)\alpha^{-1}$.

Example 5. The set of all principal ideals (a) of a ring R with $aR=Ra$ and $(b):(a):=\{x\,|\,xa\in(b)\}=(c)$, $(a)*(b):=\{x\,|\,ax\in(b)\}=(d)$ considered with respect to inclusion, $*$, $:$ ([2], page 285).

All these structures provide NR-lattices the order of which can be defined by means of $*$ and $:$ since $a\geqq b\Leftrightarrow a*b=b*b$ and $a\geqq b\Leftrightarrow b:a=b:b$. Thus we may expect a fruitful investigation if we look at the interactions of the fundamental operations. Such an attempt, of course, is by no ways new since already in the thirties and the early forties R. P. DILWORTH and MORGAN WARD treated residuation questions which arose from abstract idealtheory. Thus our paper is closely related to Dilworth/Ward. Nevertheless the topic of this note is new as question and as result.

Starting from an arbitrary residuation groupoid $(R, *, :)$ we study residuation arithmetic as far as necessary to construct a join-closed extension of $(R, *, :)$ which is an NR-lattice if and only if the original residuation groupoid satisfies one of the laws:

$$(N*) \quad x \leqq a*b \wedge x \leqq b*a \Rightarrow x = x*x$$

or

$$(N:) \quad x \leqq a:b \wedge x \leqq b:a \Rightarrow x = x:x.$$

This extension then will turn out to be canonical and we can prove that the congruences of the extension are uniquely determined by the congruences of $(R, *, :)$.

On the other hand it is well-known that a lattice can be embedded in an NR-lattice if and only if it is distributive since in this case it even can be embedded in a Boolean lattice. Thus the natural question arises which residuation groupoids admit a Boolean extension. An answer upon this question will complete our paper.

1. Arithmetic in residuation groupoids

We start by defining the fundamental structure of this note.

(1.1) $(R, *, :)$ *is called a residuation groupoid if it satisfies*

(R01) $(a*a)*b = b$ (R02) $b:(a:a) = b$

(R11) $(a*b)*(a*c) = (b*a)*(b*c)$ (R12) $(c:a):(b:a) = (c:b):(a:b)$

(R21) $a*(b*b) = c*c$ (R22) $(b:b):a = c:c$

(R31) $(a*b)*(a*c) = a*((b:a)*c)$ (R32) $(c:a):(b:a) = (c:(a*b)):a$

(R4) $a*(b:c) = (a*b):c$

(R5) $a*b = c*c = b*a \Rightarrow a = b$

$(R, *)$ *is called a right residuation groupoid if it satisfies* (R01), (R11), (R21), (R5).

Obviously we have by (R21)

(1.2) *Every right residuation groupoid satisfies* $a*a=b*b$. □

Furthermore we obtain

(1.3) *Every right residuation groupoid is partially ordered with respect to* $a \geqq b$ *iff* $a*b=a*a$ *and* $a*a=:1$ *is minimum of this partial order.*

Proof. Step by step we can show:

(i) $a*a=b*b$ by (1.2). Thus $a*a$ is a constant which we may denote by 1.

$$\text{(ii)}\qquad a \leqq a, \text{ since } a*a = 1$$

$$a \leqq b \wedge b \leqq a \Rightarrow a = b \qquad \text{by (R5)}$$

$$\begin{aligned} a \leqq b \wedge b \leqq c &\Rightarrow c*b = 1 = b*a \\ &\Rightarrow c*a = (c*b)*(c*a) \\ &\qquad\;\; = (b*c)*(b*a) \\ &\qquad\;\; = (b*c)*1 = 1 \qquad \text{by (R21)} \\ &\Rightarrow a \leqq c. \end{aligned}$$

$$\text{(iii)}\qquad a*1 = a*(b*b) = 1 \quad \square \qquad \text{by (R21).}$$

Let now $(R, *, :)$ be a residuation groupoid. We obtain

$$(1.4)\qquad a*a = b:b$$

since
$$\begin{aligned} a*a &= (b:b)*(b:b) \\ &= (b:b)*\big((b*b):(b*b)\big) \\ &= \big((b:b)*(b*b)\big):(b*b) \\ &= (b*b):(b*b) \\ &= b:b \\ &= 1. \quad \square \end{aligned}$$

$$(1.5)\qquad a*b = 1 \Leftrightarrow b:a = 1,$$

since
$$\begin{aligned} a*b = 1 \Rightarrow b:a &= \big(b:(a*b)\big):a \\ &= (b:a):(b:a) = 1 \qquad \text{(R32)} \end{aligned}$$

and since $\quad b:a = 1 \Rightarrow a*b = 1\quad$ follows by duality. □

Obviously (1.5) guarantees the important fact that $(R, *, :)$ is dual in the operations $*$, :, if we put $a*b$ for $b:a$ and $b:a$ for $a*b$, respectively.

$$(1.6)\qquad a \geqq b*a \wedge a \geqq (a:b),$$

since
$$\begin{aligned} a*(b*a) = 1 &\Leftrightarrow (b*a):a = 1 \qquad (1.5) \\ &\Leftrightarrow b*(a:a) = 1 \\ &\Leftrightarrow b*1 = 1 \end{aligned}$$

from which follows the rest by duality. □

(1.7) *Every residuation groupoid satisfies the implications:*[1]

$$a \geqq b \Rightarrow a*c \leqq b*c$$

and

$$b \geqq c \Rightarrow a*b \geqq a*c.$$

PROOF.

$$\begin{aligned} a \geqq b &\Rightarrow a*b = 1 \qquad (1.6)\\ &\Rightarrow (b*c)*(a*c) = (b*c)*\big((a*b)*(a*c)\big)\\ &\qquad\qquad\qquad\;\; = (b*c)*\big((b*a)*(b*c)\big)\\ &\qquad\qquad\qquad\;\; = 1 \end{aligned}$$

and

$$\begin{aligned} b \geqq c \Rightarrow (a*b)*(a*c) &= (b*a)*(b*c)\\ &= (b*a)*1 = 1. \quad \square \end{aligned}$$

Next we prove two formulas which we shall need in the later sections.

(1.8) *In every residuation groupoid the element $a*b$ can be expressed by*

$$\begin{aligned} a*b &= \big(b:(a*b)\big)*b\\ &= \big((b:(a*b))*(a:(b*a))\big)*\big((b:(a*b))*b\big). \end{aligned}$$

PROOF. We have

$$x \geqq x:(y*x) \leqq y. \tag{R4}$$

Hence

$$\big(b:(a*b)\big)*b \leqq a*b$$

and

$$\big(b:(a*b)\big)*b \geqq a*b \tag{1.7}$$

which implies:

(i)

$$\big(b:(a*b)\big)*b = a*b.$$

Now we consider

$$\begin{aligned} t &:= \big((b:(a*b))*(a:(b*a))\big)*\big((b:(a*b))*b\big)\\ &= \big((b:(a*b))*(a:(b*a))\big)*(a*b). \qquad (\mathrm{i}) \end{aligned}$$

We obtain

$$t \leqq a*b \quad \text{by definition}$$

and

$$t \geqq a*b \quad \text{by applying the calculation:}$$

$$\begin{aligned} &a \leqq x \wedge b \leqq x\\ \Rightarrow (a*b)*(a*c) &\geqq (a*x)*(a*c) \qquad (1.7)\\ &= (x*a)*(x*c) \qquad (\mathrm{R}11)\\ &= x*c. \end{aligned}$$

This yields

$$\big((b:(a*b))*(a:(b*a))\big)*\big((b:(a*b))*b\big) = a*b. \quad \square$$

[1] Here and further on we omit the formulation of the duals.

Now we can prove (1.9) and (1.10) below:

(1.9) $$(a*b)*(b*a) = b*a.$$

PROOF.

$$\begin{aligned}(a*b)*(b*a) &= \big((b:(a*b))*(a:(b*a))\big)*\big((b:(a*b))*b\big)\\ &\quad *\big((a:(b*a))*(b:(a*b))\big)*\big((a:(b*a))*a\big)\\ &= \big((b:(a*b))*(a:(b*a))\big)*\big((b:(a*b))*b\big)\\ &\quad *\big((b:(a*b))*(a:(b*a))\big)*\big((b:(a*b))*a\big)\\ &= \big((b:(a*b))*b\big)*\big((b:(a*b))*a\big) && \text{(R11)}\\ &= b*a. \quad \square && \text{(R11)}\end{aligned}$$

(1.10) $$(a:b)*(b:a) = b:a.$$

PROOF.

$$\begin{aligned} & (a:b)*b \leqq b\\ \Rightarrow\quad & a:\big((a:b)*b\big) \geqq a:b\\ \Rightarrow\quad & \big(a:((a:b)*b)\big)*b \leqq (a:b)*b\\ \Rightarrow\quad & \big((a:((a:b)*b))*b\big):\big((a:b)*b\big) = 1\\ \Rightarrow\quad & \big(a:((a:b)*b)\big)*\big(b:((a:b)*b)\big) = 1\\ \Rightarrow\quad & \big(b:((a:b)*b)\big):\big(a:((a:b)*b)\big) = 1 && (1.5)\\ \Rightarrow\quad & (b:a):\big(((a:b)*b):a\big) = 1 && \text{(R12)}\\ \Rightarrow\quad & (b:a):\big((a:b)*(b:a)\big) = 1\\ \Rightarrow\quad & (a:b)*(b:a) \geqq b:a\\ \Rightarrow\quad & (a:b)*(b:a) = b:a. \quad \square\end{aligned}$$

In the remainder of this section we shall be concerned with join closed groupoids.

(1.11) *Let $(R, *)$ be a right residuation groupoid. Then the system of equations*

$$\text{(F)}\qquad \begin{aligned} a*x &= a*b\\ x*a &= 1\end{aligned}$$

has at most one solution. Assume c to be a solution. Then c is $\sup(a, b)$ *with respect to* $\leqq$.

PROOF. Let c, d be solutions of (E). Then we obtain

$$\begin{aligned} c*d &= (c*a)*(c*d)\\ &= (a*c)*(a*d) = 1\\ \Rightarrow\quad c &\geqq d.\end{aligned}$$

This by duality yields $c \geqq d \wedge d \geqq c$ which means

$$c = d.$$

Furthermore we have

$$\begin{aligned} c*b &= (c*a)*(c*b) \\ &= (a*c)*(a*b) \\ &= 1. \end{aligned}$$

Hence we obtain $c \geqq a$ by assumption

and $c \geqq b$ as has just been shown.

Let us assume now that v satisfies

$$v \geqq a \wedge v \geqq b.$$

We obtain

$$\begin{aligned} v*c &= (v*a)*(v*c) \\ &= (a*v)*(a*c) \\ &= (a*v)*(a*b) \\ &= (v*a)*(v*b) \\ &= (v*a)*1 = 1 \end{aligned}$$

$$\Rightarrow \quad v \geqq c. \quad \square$$

(1.12) *By a $\cup$-closed residuation groupoid we mean a residuation groupoid in which all joins exist.*

It is easily shown that a $\cup$-closed residuation groupoid need not satisfy $a*(a\cup b)=a*b$. But in a later section we shall see that every residuation groupoid can be embedded into a $\cup$-closed residuation groupoid which satisfies $a*(a\cup b)= =a*b$. Hence the identity $a*(a\cup b)=a*b$ turns out to be natural. Therefore we define:

(1.13) *By a naturally $\cup$-closed right residuation groupoid we mean a $\cup$-closed right residuation groupoid which satisfies $a*(a\cup b)=a*b$.*

It follows:

(1.14) *Let $(R, *, \cup)$ be naturally $\cup$-closed. Then the equation holds*

$$(a\cup b)*c = (a*b)*(a*c),$$

since

$$\begin{aligned} (a\cup b)*c &= ((a\cup b)*a)*((a\cup b)*c) \\ &= (a*(a\cup b))*(a*c) \\ &= (a*b)*(a*c). \quad \square \end{aligned}$$

(1.15) *Let $(R, *, \cup)$ be naturally $\cup$-closed. Then $(R, *, :)$ satisfies the equations* (N11) *and* (N21).

PROOF. By (1.11) and duality it is sufficient to show

$$(a*(bc\,\cup))*(a*b)=1 \text{ which follows by (1.7)}$$

and

$$\begin{aligned}(a*b)*(a*(b\cup c)) &= (b*a)*(b*(b\cup c))\\ &= (b*a)*(b*c)\\ &= (a*b)*(a*c). \quad \square\end{aligned}$$

(1.16) *Let $(R,*,:,\cup)$ be $\cup$-closed. Then the equivalence holds*

$$a*(a\cup b)=a*b \Leftrightarrow (a\cup b):a=b:a.$$

PROOF. $(a\cup b):a\geqq b:a$ by (1.7). Let us assume now $a*(a\cup b)=a*b$. Then we obtain $(a\cup b):a\leqq b:a$ by

$$\begin{aligned}&((a\cup b):a):(b:a)\\ &=((a\cup b):(a*b)):a && \text{(R32)}\\ &=((a\cup b):(a*(a\cup b))):a && \text{(ass)}\\ &=((a\cup b):a):((a\cup b):a). \quad \square && \text{(R32)}\end{aligned}$$

2. Arithmetic in normal residuation groupoids

In this section we consider normal residuation groupoids. We shall see that these residuation groupoids are equationally defined dual structures with a very strong arithmetic.

(2.1) *A residuation groupoid $(R,*,:)$ is called normal if it satisfies the implication:*

$$(\mathrm{N}^*)\quad x\leqq a*b\wedge x\leqq b*a\Rightarrow x=1.$$

This leads to

(2.2) *A residuation groupoid is normal iff it satisfies:*

$$a*c=1=b*c\Rightarrow((b:(a*b))*(a:(b*a)))*((b:(a*b))*c)=1.$$

PROOF. Let (N^*) be satisfied. Then we obtain $c=1$ by $c\leqq a\wedge c\leqq b$ and (1.8)

$$\begin{aligned}((b:(a*b))*(a:(b*a)))*((b:(a*b))*c) &\leqq (b:(a*b))*c\\ &\leqq (b:(a*b))*b\\ &=a*b\end{aligned}$$

and as a consequence of (R11) we have

$$\begin{aligned}((b:(a*b))*(a:(b*a)))*((b:(a*b))*c) &\leqq (a:(b*a))*c\\ &\leqq (a:(b*a))*a\\ &=b*a.\end{aligned}$$

Let now the given implication hold and assume $x\leqq a:b\wedge x\leqq b:a$ to be true. Then putting x for c, $a:b$ for a, and $b:a$ for b we are led to $x=1$. Thus the dual of our implication holds, too, and thereby the axiom (N^*) is satisfied. □

As a second result the proof of the last theorem verifies:

(2.2′) *A residuation groupoid is normal iff it satisfies the axiom*

$$(\mathrm{N}^{:})\quad x \leqq b{:}a \wedge x \leqq a{:}b \Rightarrow x = 1.$$

In what follows we consider naturally $\cup$-closed normal residuation groupoids.

(2.3) *Every naturally $\cup$-closed normal residuation groupoid is $\cap$-closed, too, and moreover $a\cap b$ satisfies the equation* (N), *i.e.*

$$a\cap b = (a{:}(b*a))\cup(b{:}(a*b)).$$

PROOF. We have:

$$a*(a{:}(b*a)) = 1$$

$$a*(b{:}(a*b)) = (a*b){:}(a*b) = 1$$

and so by symmetry we obtain

$$a, b \geqq (a{:}(b*a))\cup(b{:}(a*b)).$$

On the other hand we obtain $a\geqq x\wedge b\geqq x\Rightarrow((a{:}(b*a))\cup(b{:}(a*b)))*x=1$ by (1.14, 2.2). This completes the proof. □

(2.4) *Every naturally $\cup$-closed normal residuation groupoid satisfies the equation:*

$$(a\cap b)*b = a*b,$$

since $$(a\cap b)*b \geqq a*b \tag{1.7}$$

and $$(a\cap b)*b \leqq (b{:}(a*b))*b = a*b. \quad \square \tag{1.8}$$

(2.5) *Every naturally $\cup$-closed normal residuation groupoid satisfies the equation*

$$(a*(b\cap c))*(a*c) = (a*b)*(a*c).$$

PROOF.

$$(a*(b\cap c))*(a*c) \geqq (a*b)*(a*c) \tag{1.7}$$

and

$$\begin{aligned}(a*(b\cap c))*(a*c) &= ((b\cap c)*a)*((b\cap c)*c)\\ &= ((b\cap c)*a)*(b*c) && (2.4)\\ &\leqq (b*a)*(b*c) && (1.7)\\ &= (a*b)*(a*c). \quad \square\end{aligned}$$

Now we are able to prove the main rules of arithmetic in normal residuation groupoids:

(2.6) *Let* $(R, *, :)$ *be a naturally $\cup$-closed normal residuation groupoid. Then* $(R, \cap, \cup)$ *is a distributive lattice.*

PROOF. We have to show $a\cup(b\cap c)\geqq(a\cup b)\cap(a\cup c)$. Therefore, since the right side is equal to

$$((a\cup b){:}((a\cup c)*(a\cup b)))\cup((a\cup c){:}((a\cup b)*(a\cup c)))$$

by symmetry it is sufficient to show:

$$\begin{aligned}
&(a\cup(b\cap c))*((a\cup b):((a\cup c)*(a\cup b)))\\
&=((a\cup(b\cap c))*(a\cup b)):((a\cup c)*(a\cup b))\\
&=((a*(b\cap c))*(a*(a\cup b))):((a*c)*(a*(a\cup b))) && (1.14)\\
&=((a*(b\cap c))*(a*b)):((a*c)*(a*b)) && (1.13)\\
&=((a*c)*(a*b)):((a*c)*(a*b))=1.\quad\square && (2.5)
\end{aligned}$$

(2.7) *Every naturally* $\cup$*-closed normal residuation groupoid satisfies the equations* (N12) *and* (N22), *whence especially we have*

$$a*(b\cap c)=(a*b)\cap(a*c).$$

PROOF.

$$a*(b\cap c)\leqq(a*b)\cap(a*c)$$

and

$$\begin{aligned}
&a*(b\cap c)*((a*b)\cap(a*c))\\
&=(a*(b\cap c))*((a*b):((a*c)*(a*b)))\cup((a*c):((a*b)*(a*c)))\\
&=(a*(b\cap c))*((a*b):((a*c)*(a*b)))\\
&\cup\,(a*(b\cap c))*((a*c):((a*b)*(a*c))) && (1.15)\\
&=((a*(b\cap c))*(a*b)):((a*c)*(a*b))\\
&\cup\,((a*(b\cap c))*(a*c)):((a*b)*(a*c))\\
&=((a*c)*(a*b)):((a*c)*(a*b)) && (2.5)\\
&\cup\,((a*b)*(a*c)):((a*b)*(a*c)) && (2.5)\\
&=1\cup1=1.\quad\square
\end{aligned}$$

(2.8) *Every naturally* $\cup$*-closed normal residuation groupoid satisfies the equations* (N13) *and* (N23), *whence especially we have*

$$(a\cap b)*c=(a*c)\cup(b*c).$$

PROOF. First we prove the formula for x, y which satisfy $x\cap y=1$.

$$\begin{aligned}
x\cap y=1&\Rightarrow x\cap(c:(y*c))=1\\
&\Rightarrow x*(c:(y*c))=(x\cap(c:(y*c)))*(c:(y*c))=c:(y*c) && (2.4)\\
&\Rightarrow(c:(y*c)):(x*(c:(y*c)))=1\\
&\Rightarrow(c:(y*c)):((x*c):(y*c))=1\\
&\Rightarrow\quad c:((x*c)\cup(y*c))=1 && (1.14)\\
&\Rightarrow\quad (x*c)\cup(y*c)=c,
\end{aligned}$$

since $(x*c)\cup(y*c)\leqq c$. Now we obtain the general formula by

$$\begin{aligned}(a\cap b)*c &= \big((a*b)\cap(b*a)\big)*\big((a\cap b)*c\big)\\ &= \big((a*b)*((a\cap b)*c)\big)\cup\big((b*a)*((a\cap b)*c)\big)\\ &= \big(((a\cap b)*b)*((a\cap b)*c)\big)\cup\big((a\cap b)*a\big)*\big((a\cap b)*c\big)\\ &= (b*c)\cup(a*c). \quad \square \qquad \text{(R.11)}\end{aligned}$$

(2.9) *Every naturally* $\cup$*-closed normal residuation groupoid satisfies the equations* (N14) *and* (N24), *whence especially we have*

$$(a\cup b)*c=(a*c)\cap(b*c).$$

PROOF. First we show for arbitrary d

$$\text{(i)} \qquad \begin{aligned} a\cup b\leqq c\Rightarrow \big((a\cup b)*c\big)*\big((a*(a\cup b))*d\big) &= \big((a*b)*(a*c)\big)*\big((a*b)*d\big)\\ &= \big((a*c)*(a*b)\big)*\big((a*c)*d\big)\\ &= (a*c)*d, \quad \text{since} \quad b\leqq c.\end{aligned}$$

From this follows for arbitrary d and $a\cup b\leqq c$ the equation

$$\text{(ii)} \quad \begin{aligned}\big((a\cup b)*c\big)*d &= \big((a\cup b)*c\big)*\big(((a*b)\cap(b*a))*d\big)\\ &= \big((a\cup b)*c\big)*\big(((a*(a\cup b))\cap(b*(a\cup b)))*d\big)\\ &= \big(((a\cup b)*c)*((a*(a\cup b))*d)\big)\cup\big(((a\cup b)*c)*((b*(a\cup b))*d)\big)\\ &= \big((a*c)*d\big)\cup\big((b*c)*d\big) \qquad \text{(i)}\\ &= \big((a*c)\cap(b*c)\big)*d,\end{aligned}$$

from which follows $(a\cup b)*c=(a*c)\cap(b*c)$, since $x*d=y*d$ $(\forall d\in R)$ implies $x=y$.

Let us assume now $a\cup b$ to be arbitrary. Then we can develop the general formula by (i) and (ii) as follows:

$$\text{(iii)} \qquad \begin{aligned}(a\cup b)*c &= \big((a\cup b)\cap c\big)*c\\ &= \big((a\cap c)\cup(b\cap c)\big)*c\\ &= \big((a\cap c)*c\big)\cap\big((b\cap c)*c\big)\\ &= (a*c)\cap(b*c). \quad \square\end{aligned}$$

Combining (1.15), (2.3), (2.6), (2.7), (2.8), (2.9) we obtain as a first main result about normal join closed residuation groupoids the

THEOREM. *Every normal naturally join closed residuation groupoid is an NR-lattice.* □

We finish this section by a remark.

(2.10) *In a naturally $\cup$-closed normal residuation groupoid the equivalence holds:*

$$a\cap b=1 \Leftrightarrow \big((a*x)*(b*x)\big)*\big((a*x)*y\big)=x*y.$$

PROOF.

$$\begin{aligned} a\cap b=1 \Rightarrow \big((a*x)*(b*x)\big)*\big((a*x)*y\big) &\\ &= \big((a*x)\cup(b*x)\big)*y\\ &= \big((a\cap b)*x\big)*y\\ &= x*y \end{aligned}$$

and if the equation on the right side is true we obtain $a\cap b=1$ by putting $x=y=$ $=z=a\cap b$. □

3. An embedding theorem

Let $(R, *)$ be a right residuation groupoid. We define on $R\times R$

(3.1) $$[a|b]*[c|d]:=[(a*b)*(a*c)|(a*b)*(a*d)].$$

It follows

(3.2) *Let* $(R, *)$ *satisfy the axioms* (R01), (R11), (R21). *Then* $(R\times R, *)$ *satisfies these three axioms, too.*

PROOF.

(i) $$([a|b]*[a|b])*[c|d]=[1|1]*[c|d]=[c|d];$$

(ii) $$[a|b]*([c|d]*[c|d])=[a|b]*[1|1]=[1|1];$$

(iii) $$([a|b]*[c|d])*([a|b]*[u|v])$$

$$\begin{aligned} &= [(a*b)*(a*c)|(a*b)*(a*d)]*[(a*b)*(a*u)|(a*b)*(a*v)]\\ &= \big[\big(((a*b)*(a*c))*((a*b)*(a*d))\big)*\big(((a*b)*(a*c))*((a*b)*(a*u))\big)\big| \ldots\\ &= \big[\big(((a*c)*(a*b))*((a*c)*(a*d))\big)*\big(((a*c)*(a*b))*((a*c)*(a*u))\big)\big| \ldots\\ &= \big[\big(((c*a)*(c*b))*((c*a)*(c*d))\big)*\big(((c*a)*(c*b))*((c*a)*(c*u))\big)\big| \ldots\\ &= \big[\big(((c*a)*(c*d))*((c*a)*(c*b))\big)*\big(((c*a)*(c*d))*((c*a)*(c*u))\big)\big| \ldots\\ &= \big[\big(((c*d)*(c*a))*((c*d)*(c*b))\big)*\big(((c*d)*(c*a))*((c*d)*(c*u))\big)\big| \ldots\\ &= ([c|d]*[a|b])*([c|d]*[u|v]). \quad \square \end{aligned}$$

(3.3) *Let* $(R, *, :)$ *be a residuation groupoid (which need not satisfy* (R31), (R32)*). Then* $(R\times R, *, :)$ *satisfies axiom* (R4) *if we define*

$$[d|c]:[b|a]:=[(d:a):(b:a)|(c:a):(b:a)].$$

PROOF.

$$
\begin{aligned}
&([a|b]*[c|d]):[u|v]\\
&=[(a*b)*(a*c)\,|\,(a*b)*(a*d)]:[u|v]\\
&=\big[(((a*b)*(a*c)):v):(u:v)\big|(((a*b)*(a*d)):v):(u:v)\big]\\
&=\big[((a*b)*((a*c):v)):(u:v)\big|((a*b)*((a*d):v)):(u:v)\big] && \text{(R4)}\\
&=\big[(a*b)*(((a*c):v):(u:v))\big|(a*b)*(((a*d):v):(u:v))\big] && \text{(R4)}\\
&=\big[(a*b)*((a*(c:v)):(u:v))\big|(a*b)*((a*(d:v)):(u:v))\big] && \text{(R4)}\\
&=[a|b]*[(c:v):(u:v)\,|\,(d:v):(u:v)] && \text{(R4)}\\
&=[a|b]*([c|d]:[u|v]). \quad \square
\end{aligned}
$$

We remark once more that we did not apply the axioms (R31), (R32) to prove (3.3). Next we show

(3.4) *Let* $(R, *, :)$ *be a residuation groupoid. Then* $(R\times R, *, :)$ *satisfies all axioms of the residuation groupoid excepted* (R5).

PROOF. It is sufficient to prove that (R32) is true since the rest follows by duality. We remark

$$[a|1]*[b|1]=[a*b|1] \quad \text{and} \quad [a|1]:[b|1]=[a:b|1]$$

and develop:

(i)
$$
\begin{aligned}
&([a|1]*[b|1])*([a|1]*[c|d])\\
&=\big[a*((b:a)*c)\big|a*((b:a)*d)\big]\\
&=[a|1]*(([b|1]:[a|1])*[c|d]),
\end{aligned}
$$

(ii)
$$
\begin{aligned}
&([u|v]:[a|1]):([b|c]:[a|1])\\
&=[u:a|v:a]:[b:a|c:a]\\
&=\big[((u:a):(c:a)):((b:a):(c:a))\big|((v:a)\ \dots\\
&=\big[((u:a):((c:a)*(b:a))):(c:a)\big|((v:a):\ \dots && \text{(R32)}\\
&=\big[((u:a):(((c:a)*b):a)):(c:a)\big|((v:a):\ \dots && \text{(R4)}\\
&=\big[((u:(a*((c:a)*b))):a)):(c:a)\big|((v:a)\ \dots && \text{(R32)}\\
&=\big[((u:(a*((c:a)*b))):(a*c)):a\big|((v:(\ \dots && \text{(R32)}\\
&=\big[((u:((a*c)*(a*b))):(a*c)):a\big|((v:((\ \dots && \text{(R31)}\\
&=\big[((u:(a*c)):((a*b):(a*c))):a\big|((v:(a*c)\ \dots && \text{(R32)}\\
&=([u|v]:[a*b|a*c]):[a|1]\\
&=([u|v]:([a|1]*[b|c])):[a|1],
\end{aligned}
$$

$$
\begin{aligned}
\text{(iii)}\qquad & ([u|v]:[a|b]):([c|d]:[a|b]) \\
&= ([u:b|v:b]:[a:b|1]):(([c|d]:[b|1]):[a:b|1]) \\
&= ([u:b|v:b]:([a:b|1]*([c|d]:[b|1]))):[a:b|1] && \text{(ii)} \\
&= (([u|v]:[b|1]):(([a:b|1]*[c|d]):[b|1])):[a:b|1] && \text{(ii)} \\
&= (([u|v]:([b|1]*([a:b|1]*[c|d]))):[b|1]):[a:b|1] && \text{(ii)} \\
&= (([u|v]:([b|1]*(([a|1]:[b|1])*[c|d]))):[b|1]):[a:b|1] \\
&= ([u|v]:(([b|1]*[a|1])*([b|1]*[c|d]))):[a|b] && \text{(i)} \\
&= ([u|v]:(([a|1]*[b|1])*([a|1]*[c|d]))):[a|b] && \text{(R11)} \\
&= ([u|v]:[(a*b)*(a*c)|(a*b)*(a*d)]):[a|b] \\
&= ([u|v]:([a|b]*[c|d])):[a|b]. \quad \square
\end{aligned}
$$

Although by our last theorems nearly all rules of the residuation groupoid remain true in $(R\times R, *, :)$ the last axiom, namely $[a|b]*[c|d]=[1|1]=[c|d]*[a|b]\Rightarrow\Rightarrow[1|1]$ fails since $[a|1]*[1|a]=[1|1]=[1|a]*[a|1]$ but $[a|1]\neq[1|a]$. Therefore we are forced to introduce a congruence relation, if we want to construct an extension satisfying the implication of axiom (R5). Here we succeed by the natural definition

$$[a|b] \equiv [c|d] \Leftrightarrow [a|b]*[c|d] = [1|1] = [c|d]*[a|b]. \tag{3.5}$$

It follows immediately:

$$[a|b] \equiv [c|d] \Leftrightarrow [c|d]:[a|b] = [1|1], \tag{3.6}$$

since

$$
\begin{aligned}
[a|b] \equiv [c|d] &\Rightarrow [a|b]*[c|d] = [1|1] \\
&\Rightarrow [c|d]:[a|b] \\
&= ([c|d]:([a|b]*[c|d])):[a|b] \\
&= ([c|d]:[a|b]):([c|d]:[a|b]) \\
&= [1|1]. \quad \square
\end{aligned}
$$

Now we can prove (3.7) and (3.8) below:

(3.7) *The relation θ defined by $\equiv$ is a congruence relation.*

PROOF.

(i) θ is an equivalence relation, since reflexivity and symmetry are obvious and transitivity results from

$$
\begin{aligned}
& [a|b] \equiv [c|d] \wedge [c|d] \equiv [u|v] \\
\Rightarrow [a|b]*[u|v] &= ([a|b]*[c|d])*([a|b]*[u|v]) \\
&= ([c|d]*[a|b])*([c|d]*[u|v]) \\
&= [1|1]*[1|1] = [1|1].
\end{aligned}
$$

(ii) θ is a congruence relation, since

$$\begin{aligned}[x|y] &\equiv [u|v] \Rightarrow \\ [x|y]*[a|b] &= ([x|y]*[u|v])*([x|y]*[a|b]) \\ &= ([u|v]*[x|y])*([u|v]*[a|b]) \\ &= [u|v]*[a|b]\end{aligned}$$

$$\begin{aligned}&\wedge([a|b]*[x|y])*([a|b]*[u|v]) \\ &\quad = ([x|y]*[a|b])*([x|y]*[u|v]) = [1|1] \\ &\quad \Rightarrow [a|b]*[x|y] \equiv [a|b]*[u|v].\end{aligned}$$

The rest follows by (3.6) and the principle of duality. □

(3.8) *$(R\times R, *, :)/\theta$ is a residuation groupoid.*

PROOF. All we have to verify is axiom (R5). Therefore assume to be given:

$$[a|b]\theta*[c|d]\theta = [1|1]\theta = [c|d]\theta*[a|b]\theta.$$

Since $[a|b]\equiv[1|1]\Rightarrow[a|b]=[1|1]$ it follows

$$\begin{aligned}&\Rightarrow [a|b]*[c|d] \equiv [1|1] \equiv [c|d]*[a|b] \\ &\Rightarrow [a|b]*[c|d] = [1|1] = [c|d]*[a|b] \\ &\qquad [a|b]\theta = [c|d]\theta. \quad \square\end{aligned}$$

We are now going to construct the announced extension. First we show:

(3.9) *Every residuation groupoid $(R, *, :)$ has an extension $(R_1, *, :)$ in which all systems of equations $a*x=a*b\wedge x*a=1$ with coefficients $a, b\in R$ are solvable.*

PROOF. We consider $(R\times R, *, :)/\theta$. This is a residuation groupoid. We consider the set of all $[a|1]\theta$. Since $[a|1]\theta*[b|1]\theta=[a*b|1]\theta$ and $[a|1]\theta:[b|1]\theta=$ $=[a:b|1]\theta$ and since $[a|1]\theta\neq[b|1]\theta\Rightarrow a*b\neq1\vee b*a\neq1$, there is an isomorphism $(R, *, :)\to(\{[a|1]\theta|a\in R\}, *, :)$. So we can embed $(R, *, :)$ into a residuation groupoid $(R_1, *, :)$ by exchanging a for $[a|1]\theta$. And this embedding guarantees the second requirement since $[a|1]\theta*[a|b]\theta=[a*b|1]\theta$ and $[a|b]\theta*[a|1]\theta=$ $=[1|1]\theta$. □

Applying the last fact we obtain as a principal result of this paper the following

EMBEDDING THEOREM. *Let $(R, *, :)$ be a residuation groupoid. Then $(R, *, :)$ admits a canonical naturally $\cup$-closed residuation groupoid extension.*

PROOF. According to (3.8), we construct an extension series

$$(R, *, :) \subseteqq (R_1, *, :) \subseteqq \ldots \subseteqq (R_n, *, :) \ldots$$

and define $S:=\cup R_\nu$ and $a*b:=c$ iff c equals $a*b$ in the first R_ν which satisfies $a, b\in R_\nu$. By analogy we define $a:b$. Then by construction every equation system in question has a solution and if $(T, *, :)$ is another extension in which these equations are solvable we can show that the set of all joins of elements of R is closed under $*$ and $:$ and isomorphic to $(S, *, :)$ with respect to these operations as follows:

(i) Obviously, $T_1 := \{a \cup b \in S \mid a \in R \wedge b \in R\}$ is closed under $*$ and $:$.

(ii) Φ: $[a|b]\theta \to a \cup b$ is a bijective function from R_1 to T_1, since

$$\begin{aligned}[a|b]\theta \geqq [c|d]\theta &\Leftrightarrow [a|b]*[c|d] = [1|1] \\ &\Leftrightarrow [(a*b)*(a*c)|(a*b)*(a*d)] = [1|1] \\ &\Leftrightarrow (a*b)*(a*c) = 1 = (a*b)*(a*d) \\ &\Leftrightarrow (a*b)*(a*c) \cup (a*b)*(a*d) = 1 \\ &\Leftrightarrow (a \cup b)*(c \cup d) = 1 \\ &\Leftrightarrow (a \cup b) \geqq (c \cup d).\end{aligned}$$

(iii) Φ is an isomorphism in consequence of duality and:

$$\begin{aligned}\Phi([a|b]\theta * [c|d]\theta) &= \Phi\big([(a*b)*(a*c)|(a*b)*(a*d)]\theta\big) \\ &= \big((a*b)*(a*c)\big) \cup \big((a*b)*(a*d)\big) \\ &= (a \cup b)*(c \cup d) \\ &= \Phi([a|b]\theta) * \Phi([c|d]\theta).\end{aligned}$$

(iv) Going from n to $n+1$ we can complete the proof by way of induction. □

We finish this section by a theorem concerning the congruences of the above extension.

(3.10) *Let $(S, *, :)$ be the smallest $\cup$-closed extension of the residuation groupoid $(R, *, :)$ and θ be a congruence of $(S, *, :)$. Then θ is uniquely determined by its restriction to $(R, *, :)$ and in addition a congruence of $(S, \cup)$.*

PROOF. Let $\equiv$ be a congruence on $(R, *, :)$ satisfying $a*b \equiv 1 \equiv b*a \Rightarrow a = b$. Define on R_1 the relation:

$$\begin{aligned}(a \cup b)\theta(c \cup d) :\Leftrightarrow{} & (a*b)*(a*c) \equiv 1 \equiv (a*b)*(a*d) \\ & \wedge (c*d)*(c*a) \equiv 1 \equiv (c*d)*(c*b).\end{aligned}$$

This is a congruence with restriction $\equiv$ what can be shown by analogy to the proof of (3.7) and it is easily seen that there are no other congruences of $(R_1, *, :)$ with $\equiv$ as restriction. So $\equiv$ can be uniquely extended by induction, and obviously in addition its extension is a congruence relation of $(S, \cup)$. □

4. The main embedding theorem

In this section we shall show that the $\cup$-closed extension constructed above is normal, too, iff $(R, *, :)$ is normal. Thus a system of axioms is given which turns out to be necessary and sufficient for a residuation groupoid to admit an NR-lattice extension. We emphasize this result by denoting it

MAIN THEOREM. *Let $(R, *, :)$ be a normal residuation groupoid. Then the extension $(S, *, :)$ is an NR-lattice.*

PROOF. We consider the special case $[a|1]$, $[c|d]$, $[u|1]$ and shall solve the main problem by applying this special case.

(i)
$$\begin{aligned}&([a|1]*[c|d])*[u|1] \equiv [1|1]\\ &\wedge([c|d]*[a|1])*[u|1] \equiv [1|1]\\ &\Rightarrow [a*c|a*d]*[u|1] = [1|1]\\ &\wedge[(c*d)*(c*a)|1]*[u|1] = [1|1]\\ &\Rightarrow ((a*c)*(a*d))*((a*c)*u) = 1\\ &\qquad\qquad\wedge((c*d)*(c*a))*u = 1\\ &\Rightarrow ((a*c)*(a*d))*((a*c)*u) = 1\\ &\wedge((c*d)*(c*a))*((c*d)*u) = 1. \end{aligned} \tag{1.7}$$

Thus by the last line and (1.7)

$$(\alpha)\qquad (c*a)*((c*d)*u) = 1$$

and by both the last lines, (R11), and (1.7)

$$\begin{aligned}&((c*a)*(c*d))*((a*c)*((c*d)*u)) = 1\\ \wedge&((c*d)*(c*a))*((a*c)*((c*d)*u)) = 1\end{aligned}$$

from which results by (N)

$$(\beta)\qquad (a*c)*((c*d)*u) = 1.$$

Thus by (α) and (β) we obtain

$$(\gamma)\qquad (c*d)*u = 1$$

and by duality in c, d

$$(\delta)\qquad (d*c)*u = 1$$

which means $u=1$.

(ii) Let us assume now the system below to be true:

$$([a|b]*[c|d])*[u|1] \equiv [1|1]$$

and

$$([c|d]*[a|b])*[u|1] \equiv [1|1].$$

We may read $=$ instead of $\equiv$ and furthermore we have

$$([a|1]*[c|d])*[u|1] = [1|1].$$

Thus all we have to prove is the equation

$$([c|d]*[a|1])*[u|1] = [1|1].$$

We succeed as follows:

$$\begin{aligned}
&([a|b]*[c|d])*[u|1] = [1|1]\\
&\wedge([c|d]*[a|b])*[u|1] = [1|1]\\
&\Rightarrow \big(((a*b)*(a*c))*((a*b)*(a*d))\big)*\big(((a*b)*(a*c))*u\big)\\
&= \big(((a*c)*(a*b))*((a*c)*(a*d))\big)*\big(((a*b)*(a*c))*u\big)\\
&=: (P*Q)*(R*u) = 1\\
&\wedge\big(((c*d)*(c*a))*((c*d)*(c*b))\big)*\big(((c*d)*(c*a))*u\big)\\
&= \big(((c*a)*(c*d))*((c*a)*(c*b))\big)*\big(((c*d)*(c*a))*u\big)\\
&= \big(((a*c)*(a*d))*((a*c)*(a*b))\big)*\big(((c*d)*(c*a))*u\big)\\
&=: (Q*P)*(S*u) = 1\\
&\Rightarrow \qquad (P*Q)*\big(R*(S*u)\big) = 1 && (1.6,\ 1.7)\\
&\qquad \wedge(Q*P)*\big(R*(S*u)\big) = 1. && (1.6,\ 1.7)
\end{aligned}$$

Thus we obtain $$R*(S*u) = 1$$

and $$P*(S*u) = 1,$$

and since $P=(a*c)*(a*b)\wedge R=(a*b)*(a*c)$ these last equations imply

$$\begin{aligned}
([c|d]*[a|1])*[u|1] &= \big[((c*d)*(c*a))*u|1\big]\\
&= [S*u|1] = [1|1].
\end{aligned}$$

(iii) The general case with $[u|v]$ instead of $[u|1]$ is solved by (i), (ii), since $[a|b]*[u|v]=[1|1]$ implies $[a|b]*[u|1]=[1|1]\wedge[a|b]*[v|1]=[1|1]$. Hence our proof is complete. □

5. Strong residuation groupoids

In this section we assume $(R, *, :)$ to be strong, i.e., to satisfy

$$\text{(S)} \qquad a:(b*a) = (b:a)*b.$$

At once we see that in this case (R5) results from (1.4). Furthermore it is pointed out in another paper that the axioms (R31) and (R32) are consequences of (S) and the other identities. But in this paper we are only interested in the proof that (S) is carried over from $(R, *, :)$ to $(S, *, :)$. We start by

(5.1) *Let* $(R, *, :)$ *be a strong residuation groupoid. Then*

$$a:(b*a) = a\cap b.$$

PROOF.

$$\begin{aligned}
a*\big(a:(b*a)\big) &= (a*a):(b*a) = 1\\
b*\big(a:(b*a)\big) &= (b*a):(b*a) = 1
\end{aligned}$$

and
$$\begin{aligned} a*x=1=b*x \Rightarrow & \big(a:(b*a)\big)*\big((x:a)*x\big) \\ & = \big(a:(b*a)\big)*\big(a:(x*a)\big) \\ & = \big((a:(b*a))*a\big):(x*a) \\ & = (b*a):(x*a)=1. \quad \square \qquad (1.7) \end{aligned}$$

As a direct consequence of (5.1) we obtain by duality

(5.2) *Every strong residuation groupoid is normal.* □

We are now going to show that (S) is transferred from $(R, *, :)$ to $(S, *, :)$.

(5.3) *Let $(R, *, :)$ be a strong residuation groupoid. Then the extension $(S, *, :)$ is strong, too.*

PROOF. All we have to show is that (S) is transferred from $(R, *, :)$ to $(R_1, *, :)$. First we show that in $(R_1, *, :)$ we have the equation

(i) $$a\cap(c\cup d)=(c\cup d):\big(a*(c\cup d)\big)$$

for elements a, b, c of R which is equivalent to

(ii) $$\big((\cup d):(a*(c\cup d)\big)*\big(a\cap(c\cup d)\big)=1.$$

Hence (i) follows from (ii) and (ii) follows from

$$\begin{aligned} & (a*d)\cap(d*a)=1 \\ \Rightarrow\ & (a*d)\cap(d*a)\cap(d*c)=1 \\ \Rightarrow\ & \big(((c:(a*d))*c):(a*c)\big)\cap\big((d*a)\cap(d*c)\big)=1 \\ \Rightarrow\ & \big((c:(a*d))*(c:(a*c))\big)\cap\big(d*(a\cap c)\big)=1 \\ \Rightarrow\ & \big((c:(a*d))\cup d\big)*(a\cap c)=1 \\ \Rightarrow\ & \big(((c:(a*d))\cup a)\cap((c:(a*d))\cup d)\big)*(a\cap c)=1 \\ \Rightarrow\ & \big((c:(a*d))\cup(a\cap d)\big)*(a\cap c)=1 \\ \Rightarrow\ & \big((c\cup d):(a*d)\big)*(a\cap c)=1 \\ \Rightarrow\ & \big(((c\cup d):(a*c))*(c:(a*c))\big)\cup\big(((c\cup d):(a*d))*(a\cap c)\big)=1 \\ \Rightarrow\ & \big(((c\cup d):(a*c))\cap((c\cup d):(a*d))\big)*(a\cap c)=1 \\ \Rightarrow\ & \big((c\cup d):(a*(c\cup d))\big)*(a\cap c)=1 \\ \wedge\ & \big((c\cup d):(a*(c\cup d))\big)*(a\cap d)=1 \\ \Rightarrow\ & \big((c\cup d):(a*(c\cup d))\big)*\big(a\cap(c\cup d)\big)=1. \end{aligned}$$

Aplying this special formula we obtain the general formula by

$$\begin{aligned}(a\cup b):((c\cup d)*(a\cup b)) &= (a\cup b):((c*(a\cup b))\cap(d*(a\cup b)))\\ &= ((a\cup b):(c*(a\cup b)))\cup((a\cup b):(d*(a\cup b)))\\ &= ((a\cup b)\cap c)\cup((a\cup b)\cap d)\\ &= (a\cup b)\cap(c\cup d). \quad \square\end{aligned}$$

Let us assume now $(R, *, :)$ to be symmetric, i.e., to satisfy $a*b=b:a$. Let us assume moreover $(R, *)$ to satisfy $a*(b*c)=(a*b)*(a*c)$. Then $(S, *)$ satisfies this identity, too, as was shown in [3]. So if $(R, *)$ is additionally strong the extension $(S, *)$ is a generalized Boolean ring. On the other hand a generalized Boolean ring satisfies the axioms of a strong symmetric residuation groupoid and it is additionally autodistributive. This leads us to the

THEOREM. *A residuation groupoid admits a Boolean extension if and only if it is symmetric, strong, and autodistributive.* □

REFERENCES

[1] BIRKHOFF, G., *Lattice theory*, AMS Coll. Bibl. (1967).

[2] BOSBACH, B., Komplementäre Halbgruppen. Axiomatik und Arithmetik. *Fund. Math.* **64** (1969), 257—287.

[3] BOSBACH, B., Concerning representation of completely join distributive algebraic lattices, *Period. Math. Hungar.* (to appear)

[4] DILWORTH, R. P., Abstract residuation over lattices, *Bull. Amer. Math. Soc.* **44** (1938), 262—268.

[5] DILWORTH, R. P., Noncommutative residuated lattices, *Trans. Amer. Math. Soc.* **46** (1939), 426—444.

[6] WARD, M., Structure residuation, *Ann. of Math.* **39** (1938), 558—568.

[7] WARD, M., Residuated distributive lattices, *Duke Math. J.* **6** (1940), 641—651.

[8] WARD, M. and DILWORTH, R. P., Residuated lattices, *Proc. Nat. Acad. Sci. USA* **24** (1938), 162—164 and *Trans. Amer. Math. Soc.* **45** (1939), 335—354.

Fachbereich Mathematik, Gesamthochschule Kassel
Postfach 10 1380, D—3500 Kassel

(Received September 17, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 453—461.

EVADING CONVEX DISCS

by

G. FEJES TÓTH

Confirming two conjectures of L. FEJES TÓTH [4, 8] we shall prove the following theorems:

THEOREM 1. *Given a set of disjoint open squares with side-lengths not exceeding* 1, *any two points of the plane lying outside the squares at distance d from one another can be connected by a path evading the squares and having length at most* $\frac{3d+1}{2}$.

THEOREM 2. *Given a set of disjoint open unit circles, any two points of the plane lying outside the circles at distance d from one another can be connected by a path evading the circles and having length at most* $\frac{2\pi}{\sqrt{27}}(d-2)+\pi$.

Using a mean value argument, J. PACH [8] proved previously a theorem similar to Theorem 1 with the weaker bound $\frac{3}{2}d+4\sqrt{d}+1$.

The problem of finding a short path between two points of the plane evading the discs of a packing is closely related to problems concerning the permeability of a layer. We say that a set of disjoint open domains lying in a parallel strip forms a *layer*. By crossing the layer we mean drawing a path connecting the edges of the strip without entering any of the domains. Let w be the width of the strip and l the length of a path crossing the strip. The *permeability* of the layer is defined by the ratio $w/\inf l$ where the infimum extends over all paths crossing the layer [2].

It is known that the permeability of any layer of squares is greater than 2/3 [3] and the permeability of any layer of equal circles is greater than $\sqrt{27}/2\pi$ [2]. Even a little more is true: starting from any boundary-point of a parallel strip of width w containing a layer of squares with side-lengths $\leqq 1$, we can cross the layer through a path of length $\leqq 3w/2$. Analogously, starting from any boundary-point of a strip of width w containing a layer of unit circles, we can cross the layer through a path of length $\leqq\frac{2\pi}{\sqrt{27}}(w-2)+\frac{\pi}{2}+1$. (The bound $\frac{2\pi}{\sqrt{27}}(w-2)+\frac{\pi}{2}+1$ has been improved recently by A. FLORIAN [5]. His bound is sharp for all values of w.) However, the proof of these estimates does not give information about the question how far the endpoint of the path lies from the point opposite to the starting point. Theorems 1 and 2 imply that the layers can be crossed also "perpendicularly" by paths of approximately the same length as the above bounds.

Theorems 1 and 2 are sharp for infinitely many values of d. Consider a lattice-packing of open unit squares which arises from a grid of squares with horizontal rows by translating every second row of squares horizontally through 1/2. Let A be the midpoint of the upper horizontal side of a square and B a point below A at distance $2k+1$ from it $(k=0, 1, \ldots)$ (Fig. 1). It is easily seen that the length

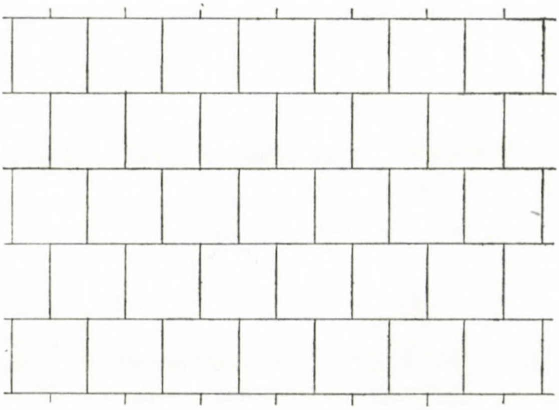


Fig. 1

of any path connecting A with B outside the squares is at least $\frac{3d+1}{2}$. Thus the bound of Theorem 1 cannot be improved if d is an odd number.

The bound of Theorem 2 is best possible if d is of the form $d=2k\sqrt{3}+2$, $k=0, 1, \ldots$. In order to see this we consider the densest lattice-packing of open unit circles with horizontal rows of circles (Fig. 2). Let A be the highest point of a circle and B the lowest point of a circle exactly below A at distance $2k\sqrt{3}+2$

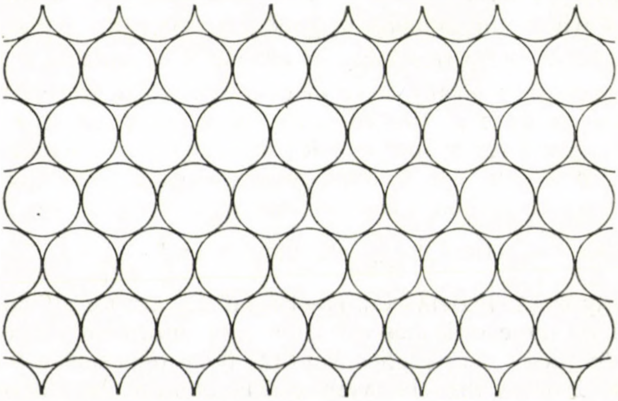


Fig. 2

from A. It is easy to check that A and B cannot be connected outside the circles by a path of length less than $\frac{2\pi}{\sqrt{27}}(d-2)+\pi$:

In contrast with Theorem 1, which is true for any packing of squares of side-lengths not exceeding 1, Theorem 2 does not hold for all packings of circles of radius $\leqq 1$. This is shown by an example of a layer of incongruent circles with permeability less than $\sqrt{27}/2\pi$ [2].

Now we turn to the proof of Theorem 1. Let $\mathcal{P}$ be a packing of open squares with side-lengths $\leqq 1$. We shall restrict ourselves to the case when the squares do not accumulate. The general case follows from this case easily by known compactness arguments. Such an argument is presented in [1].

To a given point P outside the squares and a prescribed direction $\mathbf{d}$ we construct a path described in [3]. Starting from P we go in the direction $\mathbf{d}$ until we arrive to a boundary-point X of a square of $\mathcal{P}$, say of the square $S=X_1X_2X_3X_4$, which blocks our way. Let $\lambda(XY)$ be the length of the shortest path connecting the boundary-points X and Y of S through a portion of the boundary of S. Let $\delta_{\mathbf{d}}(XY)=$ $=\delta(XY)$ be the inner product of the vector XY and the unit vector pointing in the direction $\mathbf{d}$. We evade S by going from X along the shorter arc of the boundary of S to that vertex, or to one of those vertices X_i of S for which the ratio $\delta(XX_i)/\lambda(XX_i)$ is maximal. Leaving the boundary of S at X_i we continue our way in the direction $\mathbf{d}$ until a second square blocks our way. We evade this square in a similar way as the first one, travel again in the direction $\mathbf{d}$, and so on. Since the squares of $\mathcal{P}$ do not accumulate, we can construct by this procedure an infinite path $\Pi=\Pi(P,\mathbf{d})$, which we shall call a *path emanating from P in the direction* $\mathbf{d}$. Since it may occur that $\max_i \delta(XX_i)/\lambda(XX_i)$ is attained for more than one vertex of a square, $\Pi(P,\mathbf{d})$ is generally not uniquely determined.

Let X be an arbitrary point of Π. Let $l(X)$ denote the length of the arc of Π lying between P and X. We shall show that

$$l(X) \leqq \frac{3}{2}(\delta(PX)+1). \tag{1}$$

To the proof we need the following simple

LEMMA. *Let X be a boundary-point of the square $S=X_1X_2X_3X_4$ such that the half-line emanating from X in the direction* $\mathbf{d}$ *intersects S. Then we have*

$$\max_{1\leqq i\leqq 4}\frac{\delta(XX_i)}{\lambda(XX_i)} \geqq \frac{2}{3}.$$

It will be convenient to suppose that the plane under consideration is in such a position that the direction $\mathbf{d}$ is vertical. We assume that no side of S is horizontal, because otherwise the lemma is obvious. We also suppose, without loss of generality, that S is a unit square. We choose the notations so that the highest vertex of S is X_1 and X lies on the edge X_1X_2 (Fig. 3). We introduce the notations $a=\delta(X_1X_2)$ and $b=\delta(X_1X_4)$ and observe that, by the symmetry of S and the theorem of Pythagoras, we have

$$a^2+b^2=1. \tag{2}$$

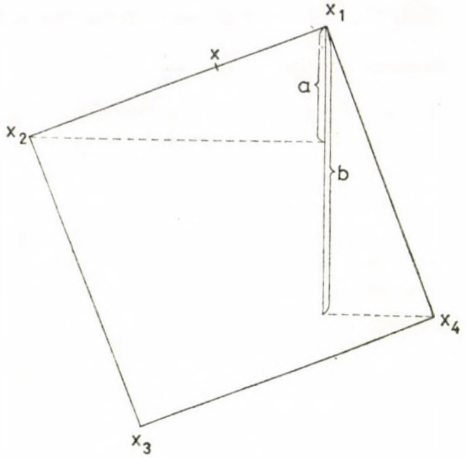


Fig. 3

If $a \geqq b$ then $a \geqq \sqrt{2}/2$, which implies that $\delta(XX_2)/\lambda(XX_2) = a \geqq \sqrt{2}/2 > 2/3$. Therefore we assume that

$$(3) \qquad 0 < a < \sqrt{2}/2 < b < 1.$$

While X travels on the segment X_1X_2 from X_1 to X_2 the ratio $\delta(XX_4)/\lambda(XX_4)$ continuously decreases. Since $\delta(X_1X_4)/\lambda(X_1X_4) = b > \sqrt{2}/2 > 2/3$ and $\delta(X_2X_4)/\lambda(X_2X_4) = (b-a)/2 < 1/2 < 2/3$, there is a point X_0 on the segment X_1X_2 for which $\delta(X_0X_4)/\lambda(X_0X_4) = 2/3$. The above considerations show that if $X \in X_1X_0$ then $\delta(XX_4)/\lambda(XX_4) > 2/3$. Thus the lemma will be proved by showing that in the case $X \in X_0X_2$ we have $\delta(XX_3)/\lambda(XX_3) > 2/3$.

We observe that $\delta(XX_3)/\lambda(XX_3)$ increases if X moves from X_0 to X_2. Therefore it is enough to show that $\delta(X_0X_3)/\lambda(X_0X_3) \geqq 2/3$. Introducing the notation $x = X_1X_0$ we have $\delta(X_0X_3) = a+b-ax$, $\delta(X_0X_4) = b-ax$, $\lambda(X_0X_3) = 2-x$ and $\lambda(X_0X_4) = 1+x$, so that the condition $\delta(X_0X_4)/\lambda(X_0X_4) = 2/3$ is equivalent with

$$(4) \qquad 2(1+x) = 3(b-ax).$$

In order to finish the proof of the lemma we have to show that under the conditions (2), (3) and (4) we have

$$2(2-x) < 3(a+b-ax).$$

But this is equivalent with the inequality $3a^2-4a+1<0$ which is, for $\sqrt{2}/2 < a < 1$, true, indeed.

We return to the proof of the inequality (1). Let $S_1, S_2, \ldots$ be the squares of $\mathscr{P}$ we succesively hit when moving from P along Π. Let V_i be the first and W_i the last point of Π common with the boundary of S_i ($i=1, 2, \ldots$). Write $W_0 = P$. Using the lemma we immediately see that if X is a point of the segment W_iV_{i+1} ($i=0, 1, \ldots$) then

$$(5) \qquad l(X) \leqq \frac{3}{2}\delta(PX).$$

Now we consider the case when X lies between V_i and W_i $(i=1, 2, \ldots)$. Since by (5) $l(V_i)\leqq\frac{3}{2}\delta(PV_i)$, it suffices to show that

$$l(X)-l(V_i) \leqq \frac{3}{2}\delta(V_iX)+\frac{1}{2}. \tag{6}$$

Let Y be the first vertex of Π we meet when travelling from W_i backwards on Π. If X lies on the closed segment YW_i then we have by the lemma and the supposition that $YW_i\leqq 1$

$$l(X)-l(V_i) = l(W_i)-l(V_i)-XW_i \leqq \frac{3}{2}\delta(V_iW_i)-\delta(XW_i) =$$

$$= \frac{3}{2}(\delta(V_iW_i)-\delta(XW_i))+\frac{1}{2}\delta(XW_i) \leqq \frac{3}{2}\delta(V_iX)+\frac{1}{2}$$

confirming the inequality (6) for the case $X\in YW_i$. Finally, suppose that $X\in V_iY$. We observe that $\frac{3}{2}\delta(V_iX)+\frac{1}{2}+l(V_i)-l(X)$ is now a linear function of the distance $x=V_iX$. We have just seen that this quantity is not negative for $X=Y$ and equal to $\frac{1}{2}$ for $X=V_i$. This completes the proof of (6) and simultaneously the proof of (1).

Let A and B be two points of the plane lying outside the squares of $\mathscr{P}$. We shall show that there are two paths, Π^+ and Π^-, emanating from A and B, respectively, in opposite directions $\mathbf{d}$ and $-\mathbf{d}$ which intersect one another. Before the proof of this proposition we shall show that, along with the inequalities (1) and (5), it implies Theorem 1.

It is easily seen that Π^+ and Π^- have a point Z in common which is the vertex of a square S of $\mathscr{P}$ touched and evaded both by Π^+ and Π^- and at least one of the paths Π^+ and Π^-, say Π^+ leaves the boundary of S at Z. By (5), the length of the arc of Π^+ between A and Z is at most $\frac{3}{2}\delta_{\mathbf{d}}(AZ)$. On the other hand, by (1) the length of the arc of Π^- between B and Z is at most $\frac{3}{2}\delta_{-\mathbf{d}}(BZ)+\frac{1}{2}=\frac{3}{2}\delta_{\mathbf{d}}(ZB)+\frac{1}{2}$. The path composed of these two arcs joins A and B and its length is at most $\frac{3}{2}AB+\frac{1}{2}$.

Thus, all we need to prove is the existence of the paths Π^+ and Π^-. As we observed above, it may occur that there are several paths emanating from a given point in a given direction. We shall consider two special paths which generally coincide. The paths $\Pi_r(P, \mathbf{d})$ and $\Pi_l(P, \mathbf{d})$ emanating from P in the direction $\mathbf{d}$ turn to the right or to the left, respectively, whenever there is a choice of equally economic possibilities to evade a square of $\mathscr{P}$. More precisely, if S is a square of $\mathscr{P}$ which is evaded by the arc VW of $\Pi_r(P, \mathbf{d})$ or $\Pi_l(P, \mathbf{d})$ and S has a vertex X other than W for which $\delta(VW)/\lambda(VW)=\delta(VX)/\lambda(VX)$ then W lies on the right-hand side or on the left-hand side of the oriented line of direction $\mathbf{d}$ through V.

Let e be the line through B perpendicular to AB. We assign in the plane a positive direction of rotation in the counter-clockwise sense and divide e into the halflines e_1 and e_2 arising from the halfline BA by a rotation through $\pi/2$ and $-\pi/2$, respectively. Let U be the set of those points X of e to which there is a direction $\mathbf{d}$ such that the path $\Pi_r(A, \mathbf{d})$ or $\Pi_l(A, \mathbf{d})$ contains X. U is not empty because, e.g., any of the paths $\Pi_r(A, \mathbf{d})$ and $\Pi_l(A, \mathbf{d})$ with $\mathbf{d}=\overrightarrow{AB}$ intersects e. Thus at least one of the halflines e_1 and e_2, say e_1, contains a point of U. We assume that $B\notin U$ since otherwise our statement is obvious. It is easily seen that U is closed, thus there is a point N of $e_1\cap U$ which lies nearest to B. Let $\mathscr{D}$ be the set of directions $\mathbf{d}$ for which one of the paths $\Pi_r(A, \mathbf{d})$ and $\Pi_l(A, \mathbf{d})$ contains N. Let $\alpha(\mathbf{d})$ be the angle of rotation carrying the direction of e_1 into $\mathbf{d}$ such that $0\leqq|\alpha(\mathbf{d})|\leqq\pi$. Let $\mathbf{d}$ be the element of $\mathscr{D}$ for which $\alpha(\mathbf{d})$ is maximal. We are going to show that, with this particular choice of the direction $\mathbf{d}$, any path emanating from B in the direction $-\mathbf{d}$ has a point in common with $\Pi_r(A, \mathbf{d})$ or $\Pi_l(A, \mathbf{d})$.

First we observe that the path $\Pi_l(A, \mathbf{d})$ does not pass through N. For, suppose that $N\in\Pi_l(A, \mathbf{d})$. Then we can choose a direction $\mathbf{d}'$ which is obtained from $\mathbf{d}$ by a rotation through a sufficiently small positive angle so that the point $N'=e\cap\Pi_l(A, \mathbf{d}')$ lies on the closed segment BN. But this is impossible because of the definition of N and $\mathbf{d}$ and the assumption that $B\notin U$.

Let f be the line through B perpendicular to the direction $\mathbf{d}$. Obviously, both $\Pi_r(A, \mathbf{d})$ and $\Pi_l(A, \mathbf{d})$ intersect f. Thus these two paths together with f enclose a bounded region R (which is generally not connected). Let M_r and M_l be the intersections of the paths $\Pi_r(A, \mathbf{d})$ and $\Pi_l(A, \mathbf{d})$ with f. According to the above considerations B lies on the open segment M_rM_l (Fig. 4).

Now we consider an arbitrary path Π emanating from B in the direction $-\mathbf{d}$. Let Q be the last point of Π common with the closed segment M_rM_l. It suffices to consider the case when Q is an interior point of the segment M_rM_l. Then travelling on Π we enter the interior of R after leaving Q and since R is bounded while Π is unbounded we shall cross the boundary of R in a point, say T. This point T is a common point of Π and one of the paths $\Pi_r(A, \mathbf{d})$ and $\Pi_l(A, \mathbf{d})$, as claimed.

This completes the proof of Theorem 1.

The proof of Theorem 2 is similar. Let now $\mathscr{P}$ be a packing of open unit circles. For a given point P outside the circles and a given direction $\mathbf{d}$ we draw a path $\hat{\Pi}$ evading the circles by a similar construction as in the case of squares: We go in the direction $\mathbf{d}$ if no circle of $\mathscr{P}$ hinders our way and evade any circle $C\in\mathscr{P}$ which blocks our way by going along the boundary of C to the nearest endpoint (or to one of the nearest endpoints) of the diameter of C perpendicular to $\mathbf{d}$. Let $C_1, C_2, \ldots$ be the circles of $\mathscr{P}$ succesively touched and evaded by $\hat{\Pi}$. Let V_i be the first and W_i the last point of $\hat{\Pi}$ common with the boundary of C_i $(i=1, 2, \ldots)$. Let C_i' be a unit circle touching the segment W_iV_{i+1} at X_i and the circle C_{i+1} at a point Y_i belonging to $\hat{\Pi}$. We note that the arc X_iY_i of C_i' does not intersect any circle of $\mathscr{P}$. We replace the arc of $\hat{\Pi}$ between the points X_i and Y_i by the circular arc X_iY_i. Rounding off all vertices of $\hat{\Pi}$ in this way we obtain a new path Π for which we again use the term *path emanating from P in the direction* $\mathbf{d}$ (Fig. 5).

Let Π be a path emanating from P in the direction $\mathbf{d}$ and X an arbitrary point of Π. Let $l(X)$ denote the length of the arc of Π lying between P and X. The arcs X_jY_j and Y_jW_{j+1} are congruent and their common length is $\arcsin\frac{1}{2}\delta(X_jW_{j+1})$.

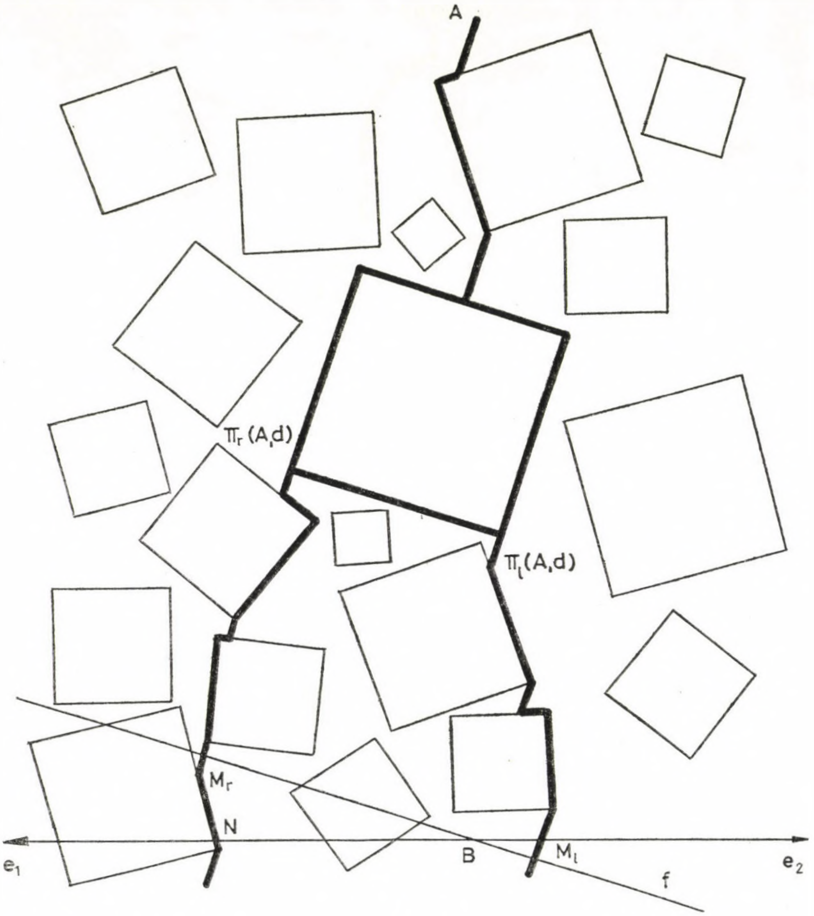


Fig. 4

The length of the arc V_1W_1 is $\arcsin\delta(V_1W_1)$. Thus we have

$$l(W_i) = PV_1+\arcsin\delta(V_1W_1)+\sum_{j=1}^{i-1} W_jX_j+2\sum_{j=1}^{i-1}\arcsin\frac{1}{2}\,\delta(X_jW_{j+1}).$$

Observe that $\delta(X_jW_{j+1})\leqq\sqrt{3}$. Since the function $\dfrac{\arcsin x}{x}$ is increasing we have

$$\frac{\arcsin\frac{1}{2}\,\delta(X_jW_{j+1})}{\delta(X_jW_{j+1})}\leqq\frac{\arcsin\frac{\sqrt{3}}{2}}{\frac{\sqrt{3}}{2}}=\frac{2\pi}{\sqrt{27}}.$$

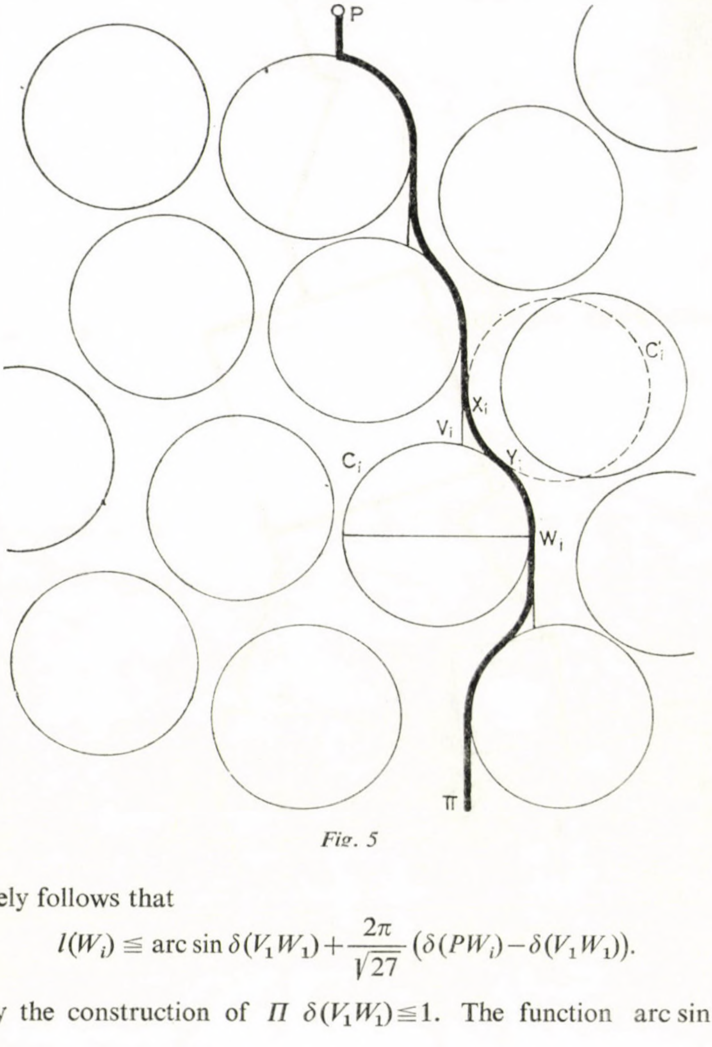


Fig. 5

It immediately follows that

$$l(W_i) \leqq \arcsin \delta(V_1 W_1) + \frac{2\pi}{\sqrt{27}}\left(\delta(PW_i) - \delta(V_1 W_1)\right).$$

We have by the construction of Π $\delta(V_1 W_1) \leqq 1$. The function $\arcsin x - \frac{2\pi}{\sqrt{27}} x$ attains its maximum $\frac{\pi}{2} - \frac{2\pi}{\sqrt{27}}$ in the interval $0 \leqq x \leqq 1$ for $x = 1$. It follows that

$$l(W_i) \leqq \frac{2\pi}{\sqrt{27}}\left(\delta(PW_i) - 1\right) + \frac{\pi}{2}. \tag{7}$$

Let now A and B be two points of the plane lying outside the circles of $\mathcal{P}$. Repeating the argument of the proof of Theorem 1 we see that there are two paths Π^+ and Π^- such that Π^+ emanates from A in a direction, say, **d** and Π^- emanates from B in the direction opposite to **d** and Π^+ and Π^- intersect one another. Moreover, there is a circle C of $\mathcal{P}$ touched and evaded both by Π^+ and Π^- such that

Π^+ and Π^- have a point W in common which is the endpoint of the diameter of C perpendicular to $\mathbf{d}$. In view of (7) the length of Π^+ between A and W is at most $\frac{2\pi}{\sqrt{27}}(\delta(AW)-1)+\frac{\pi}{2}$ and the length of the arc of Π^- between W and B is at most $\frac{2\pi}{\sqrt{27}}(\delta(WB)-1)+\frac{\pi}{2}$. The length of the path composed of these two arc is at most $\frac{2\pi}{\sqrt{27}}(AB-2)+\pi$.

Several results are known about the permeability of some special layers. Among others layers consisting of translates of a regular triangle [7], of a regular hexagon [7] and of a disc of constant width [8], as well as of similar replicas of a parallelogram [2, 4] have been studied. These results depend on the construction of suitable paths by similar methods as in the case of squares or congruent circles. These paths are "uniformly good" in the sense that they satisfy a condition analogous to (1) or (7). Also the method of intersecting paths emanating from two points in opposite directions can be applied with all these paths. Thus we can obtain good upper bounds for the length of the shortest path connecting two points also in the case of the above mentioned discs.

REFERENCES

[1] Bollobás, B., Remarks to a paper of L. Fejes Tóth, *Studia Sci. Math. Hungar.* **3** (1968), 373—379.

[2] Fejes Tóth, L., On the permeability of a circle layer, *Studia Sci. Math. Hungar.* **1** (1966), 5—10.

[3] Fejes Tóth, L., On the permeability of a layer of parallelograms, *Studia Sci. Math. Hungar.* **3** (1968), 195—200.

[4] Fejes Tóth, L., Research problem № 24, *Period. Math. Hungar.* **9** (1978), 173—174.

[5] Florian, A., On the permeability of layers of discs, *Studia Sci. Math. Hungar.* **13** (1978), 125.

[6] Florian, A., Über die Durchlässigkeit gewisser Scheibenschichten. *Sitzungsber. Österreichischen Akad. Wiss., Math.-Naturw. Klasse, Abt. II.* **188** (1979), 417—427.

[7] Hortobágyi, I., Über die Durchlässigkeit einer aus Scheiben konstanter Breite bestehenden Schicht, *Studia Sci. Math. Hungar.* **11** (1976), 383—387.

[8] Pach, J., On the permeability problem, *Studia Sci. Math. Hungar.* **12** (1977), 419—424.

Mathematical Institute of the Hungarian Academy of Sciences,
H—1053 Budapest, Reáltanoda u. 13—15

(Received November 15, 1979)

NON-DEGENERATE INNER PRODUCT SPACES SPANNED BY TWO NEUTRAL SUBSPACES

by

J. BOGNÁR

Abstract. We make some remarks in connection with the following problem. Suppose the non-degenerate inner product space E is the direct sum of two neutral subspaces L_1 and L_2. What is the necessary and sufficient condition, in terms of L_1 and L_2, that E be decomposable?

Let E be a complex vector space[1] equipped with a non-degenerate hermitian form $(\cdot, \cdot)$. The form $(\cdot, \cdot)$ will be called the *inner product*, and E a *non-degenerate inner product space.*

An element $x \in E$ is said to be *neutral* if $(x, x)=0$. A subspace $L \subset E$ is said to be *neutral* if all its elements are neutral. By the Bunjakovski—Schwarz inequality, on a neutral subspace the inner product vanishes identically.

We say the element $x \in E$ is *positive* if $(x, x)>0$. Further, we say the subspace $L \subset E$ is *positive definite* if the relations $x \in L$ and $x \neq 0$ imply $(x, x)>0$. Negativeness of elements and negative definiteness of subspaces are defined similarly.

Orthogonality of x and y means $(x, y)=0$.

If E is the orthogonal direct sum of a positive definite subspace E^+ and a negative definite subspace E^-,

$$E = E^+ \oplus E^-, \tag{1}$$

we say E is *decomposable* and (1) is a *fundamental decomposition* of E.

The first example of a non-decomposable E was given by G. W. MACKEY (see [4] or [1; Example I.11.3]). Another example (see [2; § 2.4]) has been constructed by M. L. BRODSKIĬ. A third example is due to V. I. OVČINNIKOV [3], who pointed out that the idea common to all of these examples was the following.

PROPOSITION A (Mackey, Ovčinnikov). *If the non-degenerate inner product space E is the direct sum of two non-isomorphic neutral subspaces, then E is non-decomposable.*

One may ask whether all non-decomposable spaces arise in this manner.

PROBLEM 1. Can every non-decomposable, non-degenerate inner product space be written as the direct sum of two non-isomorphic neutral subspaces?

At first glance, the answer appears to be "no": a construction of G. WITTSTOCK (see [1; Example IV.5.6]) provides non-decomposable spaces E each representable

[1] We use this opportunity to point out an error in [1; Section VIII.6]: assertion 15) is false. Namely, if the principal subspace to $\lambda=0$ of a positive operator on a Krein space is ortho-complemented, then the right-hand and left-hand limits at $\lambda=0$ of the spectral function need not exist. We thank P. JONAS for calling our attention to this fact.

as the direct sum of two *isomorphic* neutral subspaces. However, it is not at all clear that for any other direct decomposition of these spaces into the sum of two neutral subspaces the components are isomorphic. In fact, as shown by the example below, a space E may be the direct sum of two isomorphic neutral subspaces and the direct sum of two non-isomorphic neutral subspaces at the same time.

We cannot decide whether Wittstock's spaces admit direct decompositions into the sum of two non-isomorphic neutral subspaces. Thus Problem 1 remains open.

EXAMPLE 1. Consider the vector space E of complex numerical sequences $x=\{\xi_1, \xi_2, \dots\}$ such that $\xi_{2k}=0$ for $k>k_0(x)$. If $y=\{\eta_1, \eta_2, \dots\}\in E$, define the inner product of x and y by the relation

$$(x, y) = \sum_{k=1}^{\infty} (\xi_{2k-1}\bar{\eta}_{2k}+\xi_{2k}\bar{\eta}_{2k-1}).$$

Then E is Mackey's non-decomposable, non-degenerate inner product space. It is the direct sum of the neutral subspaces

$$L_1 = \{x\colon \xi_{2k-1} = 0 \text{ for } k = 1, 2, \dots\}$$

and

$$L_2 = \{x\colon \xi_{2k} = 0 \text{ for } k = 1, 2, \dots\}.$$

Here the algebraic dimension of L_1 is countable, while that of L_2 is continuum. Simultaneously, E is the direct sum of the neutral subspaces

$$N_1 = \{x\colon \xi_{4k-3} = \xi_{4k} = 0 \text{ for } k = 1, 2, \dots\}$$

and

$$N_2 = \{x\colon \xi_{4k-2} = \xi_{4k-1} = 0 \text{ for } k = 1, 2, \dots\},$$

the algebraic dimension of both N_1 and N_2 being continuum.

REMARK. The negative result just obtained can be weakened to the following: in a decomposition of a non-degenerate inner product space into the direct sum of two neutral subspaces, the algebraic dimensions of the components are not uniquely defined. On the other hand, Proposition A implies that this anomaly may occur in non-decomposable spaces only: in a decomposable E, both neutral components (provided they exist) have "half of the dimension" of E. The next theorem says something more.

THEOREM 1. *If*

$$E = E^+ \oplus E^-$$

is a fundamental decomposition of the non-degenerate inner product space E and

$$E = L_1 \dotplus L_2$$

is a decomposition of E into the direct sum of two neutral subspaces, then

$$\operatorname{Dim} L_1 = \operatorname{Dim} L_2 = \operatorname{Dim} E^+ = \operatorname{Dim} E^-,$$

where Dim *denotes algebraic dimension.*

PROOF. Consider the projections P^+ and P^- defined on E by the relations

$$x = P^+x+P^-x, \quad P^+x\in E^+, \quad P^-x\in E^-.$$

If $x\in L_1$ and $P^+x=0$, where $x\neq 0$, then $(x, x)=0$ and $(x, x)=(P^-x, P^-x)<0$ simultaneously. Thus $P^+|L_1$ is one-to-one. In particular,

$$\operatorname{Dim} L_1 \leqq \operatorname{Dim} E^+.$$

Define the projections P_1 and P_2 by

$$x = P_1x+P_2x, \quad P_1x\in L_1, \quad P_2x\in L_2.$$

If $x\in E^+$ and $P_1x=0$, where $x\neq 0$, then $(x, x)>0$ and $(x, x)=(P_2x, P_2x)=0$ at the same time. Therefore $P_1|E^+$ is one-to-one. Hence

$$\operatorname{Dim} E^+ \leqq \operatorname{Dim} L_1.$$

As a result, $\operatorname{Dim} L_1=\operatorname{Dim} E^+$. The proof of the relations $\operatorname{Dim} L_1=\operatorname{Dim} E^-$, $\operatorname{Dim} L_2=\operatorname{Dim} E^+$, $\operatorname{Dim} L_2=\operatorname{Dim} E^-$ is similar.

REMARK. Under the circumstances of Theorem 1, the cardinal numbers $\operatorname{Dim} E^+$ and $\operatorname{Dim} E^-$ do not depend on the choice of the fundamental decomposition. With the help of the four projections defined by two fundamental decompositions it can be seen that this is generally true. Theorem 1 says, among others, that a decomposable space with $\operatorname{Dim} E^+\neq\operatorname{Dim} E^-$ cannot be represented as the direct sum of two neutral subspaces at all. We do not know whether the absence of such representations is also possible for spaces with $\operatorname{Dim} E^+=\operatorname{Dim} E^-$, and for non-decomposable spaces.

PROBLEM 2. Given a fundamental decomposition $E=E^+\oplus E^-$, with $\operatorname{Dim} E^+=$ $=\operatorname{Dim} E^-$, of the non-degenerate inner product space E, how can we decide whether E admits a representation as the direct sum of two (necessarily isomorphic) neutral subspaces?

The following counterpart to this problem is related to Problem 1. Its importance turns out also from Example 1.

PROBLEM 3. Given a representation of the non-degenerate inner product space E as the direct sum of two isomorphic neutral subspaces, how can we decide whether E is decomposable?

Below, we make the first step in the study of Problems 2 and 3.

Two subspaces $M_1, M_2\subset E$ are said to be *anti-isometrically isomorphic* if there exists a one-to-one linear mapping V of M_1 onto M_2 such that $(Vx, Vy)=-(x, y)$ for all $x, y\in M_1$.

We say the neutral subspaces $L_1, L_2\subset E$ are *positively isomorphic* if there exists a one-to-one linear mapping A of L_1 onto L_2 such that $(Ax, x)>0$ for $x\in L_1$, $x\neq 0$.

THEOREM 2. *For a non-degenerate inner product space E the following two statements are equivalent:*

(i) *E is decomposable and has a fundamental decomposition into anti-isometrically isomorphic subspaces;*

(ii) *E is the direct sum of two positively isomorphic neutral subspaces.*

PROOF. Let

$$E=E^{+}\oplus E^{-}$$

be a fundamental decomposition and V a one-to-one linear mapping of E^{+} onto E^{-} such that

$$(Vx, Vy)=-(x, y) \quad \text{for all} \quad x, y \in E^{+}. \tag{2}$$

Put

$$L_1=\{x+Vx\colon\ x\in E^{+}\},$$
$$L_2=\{x-Vx\colon\ x\in E^{+}\}.$$

Then L_1 and L_2 are (linear) subspaces. They are neutral since, by (2) and the orthogonality of E^{+} and E^{-},

$$(x+Vx, x+Vx)=(x, x)+(Vx, Vx)=0,$$

$$(x-Vx, x-Vx)=(x, x)+(Vx, Vx)=0$$

for all $x\in E^{+}$.

Moreover, the relation

$$x+Vx=y-Vy$$

implies

$$x-y=-Vx-Vy$$

and, in view of $E^{+}\cap E^{-}=0$,

$$x=y, \quad Vx=-Vy=-Vx, \quad Vx=0.$$

Therefore by (2) $(x, x)=0$; using the positive definiteness of E^{+} we obtain $x=0$. Hence $x+Vx=0=y-Vy$, that is, $L_1\cap L_2=0$.

Further, for any $x\in E^{+}$

$$x=\frac{1}{2}(x+Vx)+\frac{1}{2}(x-Vx),$$

$$Vx=\frac{1}{2}(x+Vx)-\frac{1}{2}(x-Vx).$$

Therefore $E^{+}\subset L_1+L_2$ and $E^{-}\subset L_1+L_2$, which yields the relation $E=L_1+L_2$.

Now, define a mapping A in the following way:

$$A(x+Vx)=x-Vx \quad (x\in E^{+}).$$

This definition is correct, since assuming

$$x+Vx=y+Vy,$$

where $x, y\in E^{+}$, we obtain

$$x-y=-Vx+Vy$$

and, in view of the relation $E^{+}\cap E^{-}=0$,

$$x=y.$$

It is clear that A is a linear mapping of L_1 onto L_2. We have

$$(A(x+Vx), x+Vx) = (x-Vx, x+Vx) = (x, x)-(Vx, Vx) = 2(x, x) > 0$$

unless $x=0$ (or, what is the same, unless $x+Vx=0$). Thus A is one-to-one and the subspaces L_1 and L_2 are positively isomorphic.

Let, conversely, E be the direct sum of two neutral subspaces,

$$E = L_1 \dotplus L_2,$$

and A a one-to-one linear mapping of L_1 onto L_2 such that

$$(3) \qquad (Af, f) > 0 \quad \text{for} \quad f \in L_1, \quad f \neq 0.$$

Put

$$E^+ = \{f+Af\colon f \in L_1\},$$
$$E^- = \{f-Af\colon f \in L_1\}.$$

Then E^+ and E^- are subspaces. Moreover, by (3) and the neutrality of L_1 and L_2, for any non-zero elements $f, g \in L_1$ we have

$$(f+Af, f+Af) = (f, Af)+(Af, f) = 2(Af, f) > 0,$$
$$(g-Ag, g-Ag) = -(g, Ag)-(Ag, g) = -2(Ag, g) < 0$$

and

$$(f+Af, g-Ag) = -(f, Ag)+(Af, g) =$$
$$= \frac{-1}{4} \sum_{n=1}^{4} i^n(f+i^n g, A(f+i^n g))+\frac{1}{4} \sum_{n=1}^{4} i^n(A(f+i^n g), f+i^n g) = 0.$$

(At this step we used only that (f, Af) is real.) Thus the positive definite subspace E^+ and the negative definite subspace E^- are orthogonal to each other. Clearly, the direct sum $E^+ \oplus E^-$ exists.

For $f \in L_1$ we have

$$f = \frac{1}{2}(f+Af)+\frac{1}{2}(f-Af),$$

$$Af = \frac{1}{2}(f+Af)-\frac{1}{2}(f-Af).$$

Hence $L_1+L_2 \subset E^+ \oplus E^-$, that is, $E^+ \oplus E^- = E$.

Define a mapping V as follows:

$$V(f+Af) = f-Af \qquad (f \in L_1).$$

It is easy to see that the definition is correct and yields a one-to-one linear mapping of E^+ onto E^-. Finally, for any $f, g \in L_1$

$$(f+Af, g+Ag) = (f, Ag)+(Af, g),$$

$$(f-Af, g-Ag) = -(f, Ag)-(Af, g);$$

thus E^+ and E^- are anti-isometrically isomorphic.

The proof is complete.

We express our gratitude to C. BAJASGALAN, Á. BOSZNAY and D. PETZ for useful discussions.

Added in proof. 1. I. S. IOHVIDOV has called our attention to S. P. MARKIN's paper [5], where a non-decomposable space having no representation as the direct sum of two neutral subspaces was constructed. This yields a negative solution of Problem 1, and a positive answer to the second question posed in the last sentence of the Remark following Theorem 1.

2. Recently C. BAJASGALAN has constructed a decomposable space for which $\mathrm{Dim}\, E^+ = \mathrm{Dim}\, E^-$ and which cannot be written as the direct sum of two neutral subspaces; thus the answer to the first question in the sentence just mentioned is positive.

3. From an observation appearing in [5] it follows that the conclusion of Theorem 1 remains valid if E^+ and E^- are definite but not necessarily orthogonal subspaces.

REFERENCES

[1] BOGNÁR, J., *Indefinite Inner Product Spaces,* Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 78, Springer-Verlag, Berlin, 1974.

[2] GINZBURG, JU. P. and IOHVIDOV, I. S., A study of the geometry of infinite-dimensional spaces with bilinear metric, *Uspehi Mat. Nauk* **17** (1962), no. 4, 3—56 (in Russian).

[3] OVČINNIKOV, V. I., The decomposability of spaces with indefinite metric, *Mat. Issled.* **3** (1968), no. 4, 175—177 (in Russian).

[4] SAVAGE, L. J., The application of vectorial methods to metric geometry, *Duke Math. J.* **13** (1946), 521—528.

[5] MARKIN, S. P., On the decomposability of linear spaces with a sesquilinear form, *Trudy Mat. Fak. Voronež. Gos. Univ.,* no. **10** (1973), 107—112 (in Russian).

Mathematical Institute of the Hungarian Academy of Sciences
Reáltanoda u. 13—15, H—1053 Budapest,

(Received December 10, 1979)

Studia Scientiarum Mathematicarum Hungarica 13 (1978), 469—475.

ОБ ОДНОМ ЭКСТРЕМАЛЬНОМ СВОЙСТВЕ РЕПЕРОВ, ПРИВЕДЕННЫХ ПО МИНКОВСКОМУ

G. CSÓKA

Обозначим через

$$f(\bar{x}) = \sum_{i,j=1}^{n} a_{ij} x_i x_j \tag{1}$$

положительно определенную квадратичную форму (ПКФ) от n переменных. Симметричную матрицу $A=(a_{ij})$ называют матрицей формы f.

Минковский в работе [1] высказывает для $n \leqq 5$ без доказательства интересное утверждение, которое мы сформулируем следующим образом:

Утверждение M_n: *среди всех попарно эквивалентных ПКФ выражения (у которых интересное геометрическое значение тоже)*

$$s_1 = \sum_{i=1}^{n} a_{ii}; \quad s_2 = \sum_{\substack{i,k=1 \\ i \neq k}}^{n} a_{ii} a_{kk}; \ \ldots; \ s_n = \prod_{1}^{n} a_{ii} \tag{2}$$

минимальны для приведенной по Минковскому формы.

Основной результат этой статьи следующий: утверждение M_n верно для $n \leqq 6$ (§ 2) и неверно при $n \geqq 7$ (§ 3). В § 1 этой работы мы даем краткий очерк нужных понятий (см., например, [2, 3]). В § 4 даются геометрически более наглядные, чем в § 2, доказательства основной теоремы при $n \leqq 4$ и $n=5$.

§ 1. Пусть в n-мерном евклидовом пространстве E^n задан n-мерный репер $\mathscr{E}=\{\bar{\mathbf{e}}_1, \bar{\mathbf{e}}_2, \ldots, \bar{\mathbf{e}}_n\}$, то есть система n линейно независимых векторов с общим началом $\{O\}$. Реперу $\mathscr{E}$ ставится в соответствие, во-первых, его матрица Грама $A=(a_{ik})$ с элементами $a_{ik}=\bar{\mathbf{e}}_i \bar{\mathbf{e}}_k$, во-вторых, ПКФ вида (1), коэффициенты которой являются элементами a_{ik} матрицы Грама репера.

С другой стороны, каждой ПКФ можно отнести некоторый репер с точностью до его пространственного положения.

Совокупность Γ всех точек пространства, имеющих целочисленные координаты относительно репера $\mathscr{E}$, то есть множество точек с радиусами-векторами $\bar{\mathbf{x}} = \sum_{i=1}^{n} x_i \bar{\mathbf{e}}_i$, где x_i целые числа, называется n-мерной решеткой, заданной репером $\mathscr{E}$. Указанный и всякий другой репер, задающий решетку Γ, называется ее основным репером. Эти основные реперы переходят друг в друга целочисленной подстановкой переменных, матрица которой унимодулярна. Также говорят, что ПКФ f и g эквивалентны, $f \sim g$, если их матрицы A и B связаны соотношением $G^T A G = B$, где G унимодулярная целочисленная $(n \times n)$ матрица. Так устанавливается взаимная однозначная связь между решетками и классами эквивалентности ПКФ.

Говорят, что ПКФ вида (1) приведена по Минковскому, если для любого целочисленного набора $(x_1, \ldots, x_n)$, где н.о.д. $(x_k, x_{k+1}, \ldots, x_n)=1$, выполнено соотношение $f(x_1, \ldots, x_n) \geqq a_{kk}$, $k=1, 2, \ldots, n$. Выбираем форму из класса эквивалентных ей форм, то есть выбираем один или несколько из бесконечного числа основных реперов данной решетки. Геометрический смысл этого выбора состоит в следующем: берем кратчайший вектор $\bar{\mathbf{e}}_1$ решетки Γ (или один из них) и присоединяем к нему кратчайший вектор $\bar{\mathbf{e}}_2$ из Γ, образующий с $\bar{\mathbf{e}}_1$ примитивную систему, т. е. $\bar{\mathbf{e}}_1$ и $\bar{\mathbf{e}}_2$ образуют основной репер той решетки, которая получается при сечении решетки Γ плоскостью $E^2=\{\bar{\mathbf{e}}_1, \bar{\mathbf{e}}_2\}$. Таким же образом берем $\bar{\mathbf{e}}_3$ и т. д. Вообще говоря, в решетке может быть несколько различных основных реперов (M-реперов), приведенных по Минковскому [4].

Каждой форме вида (1) ставится в соответствие точка f евклидова $\mathcal{N}$-мерного, где $\mathcal{N}=\frac{1}{2}n(n+1)$, пространства коэффициентов $E^{\mathcal{N}}$ с координатами $(a_{11}, a_{22}, \ldots, a_{nn}, a_{12}, \ldots, a_{n-1,n})$. Множеству ПКФ соответствует в $E^{\mathcal{N}}$ выпуклый конус K с началом в точке $\{O\}$. ПКФ, приведенные по Минковскому, образуют в $E^{\mathcal{N}}$ конечногранное коническое тело M — область приведения по Минковскому.

§ 2. Здесь мы докажем теорему, из которой утверждение M_n при $n \leqq 6$ очевидно следует, поскольку a_{ii} суть квадраты длин векторов основного репера.

Теорема. *Пусть в решетке Γ размерности $n \leqq 6$ заданы репер Минковского $\{\bar{\mathbf{a}}_i\}$ и любой основной репер $\{\bar{\mathbf{b}}_i\}$, удовлетворяющий условию $|\bar{\mathbf{b}}_j| \leqq |\bar{\mathbf{b}}_{j+1}|$, где $j=1, \ldots, n-1$. Тогда выполняются неравенства $|\bar{\mathbf{a}}_i| \leqq |\bar{\mathbf{b}}_i|$.*

В работе [5] дается определение и доказываются свойства одной области в K, области L^*, которая не является областью приведения вообще, но для $n \leqq 6$ она совпадает с областью M (там область M обозначается через M^*).

Определение. ПКФ f принадлежит области L^*, если для любого целочисленного набора $(x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$ выполнено условие $f(x_1, \ldots, x_{k-1}, 1, x_{k+1}, \ldots, x_n) \geqq a_{kk}$ $(k=1, \ldots, n)$ и $a_{11} \leqq a_{22} \leqq \ldots \leqq a_{nn}$. Геометрический смысл этого определения: если $\mathscr{E}=\{\bar{\mathbf{e}}_1, \ldots, \bar{\mathbf{e}}_n\}$ основной репер решетки Γ_f и $f \in L^*$, тогда проекция вектора $\bar{\mathbf{e}}_k$ на плоскость решетки Γ_k, заданной репером $\mathscr{E}_k=\{\bar{\mathbf{e}}_1, \ldots, \bar{\mathbf{e}}_{k-1}, \bar{\mathbf{e}}_{k+1}, \ldots, \bar{\mathbf{e}}_n\}$ принадлежит области Дирихле начала координат решетки Γ_k. Значит, $\bar{\mathbf{e}}_k$ является кратчайшим вектором ближайшего слоя точек решетки Γ_f, параллельного с Γ_k (рис. 1). Векторы репера занумерованы по возрастанию их длин.

Доказательство теоремы. Мы приведем доказательство только для $n=6$, так как при $2 \leqq n \leqq 6$ утверждения доказываются аналогично. Для s_1 утверждение M_n доказано в [5].

Пусть $\{\bar{\mathbf{b}}_i\} \notin M$, то есть принадлежащая ему форма не входит в L^*. Но тогда существует вектор $\bar{\mathbf{b}}_k$, который не является кратчайшим в соответственной подрешетке Γ_k^1 (рис. 1). Выбирем кратчайший вектор и обозначим через $\bar{\mathbf{b}}_k \in \Gamma_k^1$. В полученном основном репере $\{\bar{\mathbf{b}}_1, \ldots, \bar{\mathbf{b}}'_k, \ldots, \bar{\mathbf{b}}_6\}$ перенумеруем векторы по возрастанию их длин, если полученный репер не приведенный по

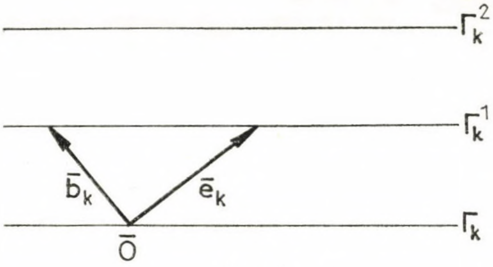


Рис. 1

Минковскому, повторяем шаг за шагом такой выбор следующего основного репера, и наконец, мы приходим к некоторому приведенному по Минковскому реперу $\{\bar{\mathbf{a}}_i'\}$ решетки Γ_f.

Действительно, рассмотрим шар радиуса $|\bar{\mathbf{b}}_6|$ в пространстве E^6 решетки Γ_f. Число векторов решетки в этом шаре конечное и только они имеются в виду при каждом шаге выбора. При каждом шаге по крайней мере у одного из векторов $\bar{\mathbf{b}}_i$ длина уменьшается. Основной репер, полученный через конечное число шагов дает уже форму, принадлежащую области $L_6^* \equiv M_6$. Если $\{\bar{\mathbf{a}}_i\} \equiv \{\bar{\mathbf{a}}_i'\}$, то теорема доказана. Пусть $\{\bar{\mathbf{a}}_i\} \not\equiv \{\bar{\mathbf{a}}_i'\}$. Докажем, что в реперах $\{\bar{\mathbf{a}}_i\}$ и $\{\bar{\mathbf{a}}_i'\}$ длины $i-$х векторов одинаковы.

ПКФ приведена по Эрмиту, если принадлежащий ей основной репер выбирается следующим образом из Γ_f: пусть первый вектор $\bar{\mathbf{e}}_1$ имеет наименьшую возможную длину (таких векторов в Γ_f может быть несколько). Каждому из выбранных векторов присоединяем кратчайший (их опять может быть несколько) вектор решетки, образующий с $\bar{\mathbf{e}}_1$ примитивную систему [6]. К дальнейшему рассмотрению мы допускаем только такие пары $\{\bar{\mathbf{e}}_1, \bar{\mathbf{e}}_2\}$, для которых вектор $\bar{\mathbf{e}}_2$ имеет из всех таких векторов наименьшую длину и т. д. Обозначим через H^* область приведения по Эрмиту. Очевидно $H^* \subset M$ (см. § 1). Доказано [5], что $L_K^* = M_K = H_K^*$ для $K \leqq 6$. Поскольку $\{\bar{\mathbf{a}}_i\} \in M_6$ и $\{\bar{\mathbf{a}}_i'\} \in M_6$, то $\{\bar{\mathbf{a}}_i\} \in H_6^*$ и $\{\bar{\mathbf{a}}_i'\} \in H_6^*$, тем самым для $i-$х векторов выполняется $|\bar{\mathbf{a}}_i| = |\bar{\mathbf{a}}_i'|$.

§ 3. С помощью контрпримера доказываем, что как наша теорема, так и утверждение M_n при $n > 6$ уже неверны.

Рассмотрим две ПКФ, где $\gamma = 10^{-10}$

$$f = \frac{7}{8}(x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2) + \frac{1}{8}(x_1 + x_2 + x_3 + x_4 + x_5)^2 + x_6^2 +$$

$$+ \left(\frac{13}{12} + 0{,}0011 - \gamma + \gamma^2\right) x_7^2 + \frac{47}{48}(x_1 + x_2 + x_3 + x_4) x_7 + \frac{5}{6} x_5 x_6 + (1 - 2\gamma) x_6 x_7 + \sum_{i=8}^{n} 7x_i^2$$

$$f' = \frac{7}{8}(x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2) + \frac{1}{8}(x_1 + x_2 + x_3 + x_4 + x_5)^2 + (1{,}0044 + 4\gamma^2) x_6^2 +$$

$$+ (1{,}0099 - 3\gamma + 9\gamma^2) x_7^2 - \frac{19}{24}(x_1 + x_2 + x_3 + x_4) x_6 + \frac{2}{3} x_5 x_6 - \frac{1}{16}(x_1 + x_2 + x_3 + x_4) x_7 -$$

$$- \frac{1}{2} x_5 x_7 + \left(\frac{3}{4} + 0{,}0132 - 4\gamma + 12\gamma^2\right) x_6 x_7 + \sum_{i=8}^{n} 7x_i^2.$$

В работе [6] доказано, что формы f и f' эквивалентны, то есть определяют реперы одной и той же решетки. Форма f приведена по Минковскому, а f' не приведена. В то же время имеем диагональные элементы матриц этих форм соответственно

$$\text{(3)}\qquad \text{для формы } f\text{: } 1;\ 1;\ 1;\ 1;\ 1;\ 1;\ \frac{13}{12}+0{,}0011-\gamma+\gamma^2;\ 7;\ 7;\ \ldots$$

$$\text{(3')}\qquad \text{для формы } f'\text{: } 1;\ 1;\ 1;\ 1;\ 1;\ 1{,}0044+4\gamma^2;\ \ 1{,}0099-3\gamma+9\gamma^2;\ 7;\ 7;\ \ldots .$$

Как видно, седьмой вектор репера у f длиннее, чем у f', это показывает, что при $n>6$ наша теорема неверна.

Пусть обозначаем число $\frac{13}{12}+0{,}0011-10^{-10}+10^{-20}$ через $1+a$, число $1{,}0044+4\cdot 10^{-20}$ через $1+b$ и $1{,}0099-3\cdot 10^{-10}+9\cdot 10^{-20}$ через $1+c$. Покажем, что утверждение M_n неверно при $n\geqq 7$.

Рассмотрим выражения s_k и s'_k — суммы k-х произведений $(1\leqq k\leqq n)$ элементов (3) и (3'). Эти элементы на $1-5$ и на $8-n$-х местах для формы f и для формы f' те же самые, поэтому при сравнении s_k и s'_k мы должны учитывать только те произведения, в которых фигурируют соответственные 6- или 7 элементы, или они оба.

Некоторыми числами A и B, которые зависят только от $1-5$ и $8-n$-х диагональных элементов, получаем не общие элементы в выражениях s_k и s'_k в следующем виде:

$$\text{в сумме } s_k\text{: } 1\cdot A+(1+a)A+1(1+a)B;$$

$$\text{в сумме } s'_k\text{: } (1+b)A+(1+c)A+(1+b)(1+c)B.$$

Отсюда получаем

$$s_k-s'_k=A(a-b-c)+B(a-b-c-bc)>0,$$

поскольку $a\approx 0{,}0844$; $b\approx 0{,}0044$ и $c\approx 0{,}0099$. Таким образом наш контрпример имеет место для любого $n\geqq 7$ и $k\leqq n$.

§ 4. В этом параграфе мы даем геометрически более наглядное доказательство теоремы для случая $n\leqq 5$. В случае $n=6$ найти такое доказательство не удалось.

Рассмотрим пример (телесно) центрированной кубической решетки, которая построена на n-мерном кубе и центр каждого куба тоже является точкой решетки. Первые n кратчайших, линейно независимых векторов решетки в случае $n\geqq 5$ будут ребрами куба, но они не образуют примитивную систему, поскольку куб, натянутый на них, содержит точку решетки, центр куба. В случае $n=4$ половина длины телесной диагонали куба равняется длине его ребер. Поэтому четыре кратчайших, линейно независимых вектора могут составлять как не примитивную, так и различные примитивные системы. Однако имеют место две нижеследующие леммы (первая из них известна [1], но, как нам кажется, мы дам её новое доказательство), в формулировках и доказательствах этих лемм мы сохраняем обозначения предыдущих параграфов.

Лемма 1. *В n-мерной, $n\leqq 4$, решетке каждый М-репер является системой n линейно независимых, последовательно кратчайших векторов. В каждой решетке, за исключением центрированной кубической, любые n таких векторов образуют М-репер.*

Следствие. В каждой решетке Γ размерности m для любых первых $n\leqq m$ $(n\leqq 4)$ линейно независимых кратчайших векторов найдется M-репер, в котором они составляют первые n элементов. При $n=4$ исключаются те решетки, которые содержают в качестве подрешетки такую 4-мерную центрированную кубическую решетку, в которой ребро куба является кратчайшим вектором решетки Γ.

Доказательство. Доказательства для $n=1, 2, 3$ аналогичны случаю $n=4$, но существенно проще, поэтому мы считаем, что Лемма 1 и её следствие уже доказаны при $n\leqq 3$ и рассмотрим только случай $n=4$. Покажем во первых, что любой вектор $\bar{\mathbf{v}}$ из $\Gamma\diagdown\Gamma_4$ не короче, чем $\bar{\mathbf{a}}_4$ (рис. 2 при $n=4$).

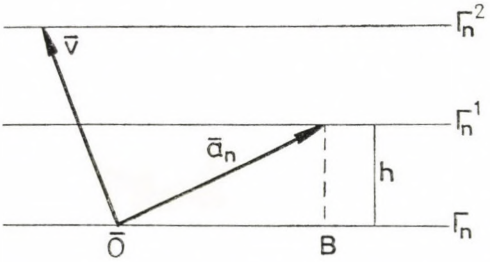


Рис. 2

Если $\bar{\mathbf{v}}\in\Gamma_4^1$, тогда это выполняется по правилам выбора $\bar{\mathbf{a}}_4$. Если $\bar{\mathbf{v}}\in\Gamma_4^i$, где $i>1$, тогда оценим снизу расстояние h между гиперплоскостями подрешеток Γ_4 и Γ_4^1, очевидно $|\bar{\mathbf{v}}|\geqq 2h$. Ортогонально проецируем решетку Γ_4^1 на подпространство E^3 подрешетки Γ_4, обозначим проекцию вектора $\bar{\mathbf{a}}_4$ через OB, длину которой мы должны оценить сверху, чтобы получить нижнюю оценку для h. Заметим, что $OB=\min\limits_{C\in\Gamma} CB=\varrho(B, \Gamma_4)$. Возьмём то параллелепипед с вершиной в точке решетки и ребрами $\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2, \bar{\mathbf{a}}_3$, в которой попала точка B. Поскольку нам нужна только подходящая верхняя оценка $\varrho(B, \Gamma_4)$, то мы берём только те точки из Γ_4, которые принадлежат подрешетке векторов $\bar{\mathbf{a}}_1$ и $\bar{\mathbf{a}}_2$, и точки подрешетки, полученной из этой параллельным перенесом на $\bar{\mathbf{a}}_3$ (рис. 3).

Проецируем точку B на плоскость векторов $\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2$ или на верхнюю плоскость, если последняя ближе. Полученная проекция B' лежит в некотором параллелограмме со сторонами $\bar{\mathbf{a}}_1$ и $\bar{\mathbf{a}}_2$ имеющем вершину в точке решетки (два варианта см. на рис. 3). Точку B' проецируем на ту прямую, несущую одну из параллельных векторов $\bar{\mathbf{a}}_1$ сторон указанного параллелограмма, к которой точка B' ближе. Расстояние от полученной проекции B'' до ближайшей

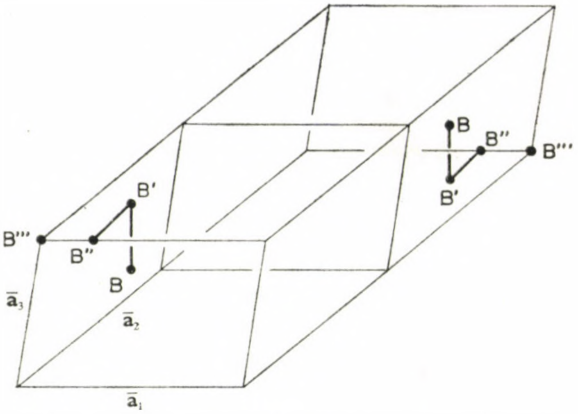


Рис. 3

точки $B'''\in\Gamma$ не больше, чем $\frac{1}{2}|\bar{\mathbf{a}}_1|$. Имеем $OB\leqq BB'''$, далее

$$B'''B^2 = BB'^2+B'B''^2+B''B'''^2 \leqq \frac{1}{4}\bar{a}_3^2+\frac{1}{4}\bar{a}_2^2+\frac{1}{4}\bar{a}_1^2 \leqq \frac{3}{4}\bar{a}_3^2 \tag{4}$$

отсюда

$$h^2 = \bar{a}_4^2-OB^2 \Rightarrow h^2 \geqq a_4^2-\frac{3}{4}a_3^2 \Rightarrow |\bar{\mathbf{v}}| \geqq 2h \geqq |\bar{\mathbf{a}}_4| \tag{5}$$

аналогичные утверждения относительно $\bar{\mathbf{a}}_3, \bar{\mathbf{a}}_2$ и $\bar{\mathbf{a}}_1$ являются фактически первым утверждением нашей леммы для $n<4$.

Пусть теперь даны некоторые три последовательно кратчайших вектора четырёхмерной решетки Γ, используя Лемму 1 и её Следствие мы считаем их первыми тремя векторами некоторого M-репера решетки Γ и обозначим через $\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2$ и $\bar{\mathbf{a}}_3$. Если последующий линейно независимый кратчайший вектор $\bar{\mathbf{v}}\in\Gamma_4^1$, то его можно принять за $\bar{\mathbf{a}}_4$ и лемма доказана. Пусть $\bar{\mathbf{v}}\notin\Gamma_4^1$ и $|\bar{\mathbf{v}}|\leqq|\bar{\mathbf{a}}_4|$, как видно из (4) и (5), это может быть только при следующих условиях: $OB=BB'''$; $|\bar{\mathbf{a}}_i|=|\bar{\mathbf{v}}|$ и векторы $\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2, \bar{\mathbf{a}}_3$ и $\bar{\mathbf{v}}$ попарно перпендикулярны друг другу; точка A_4 в результате перенёса её вектором $-\frac{1}{2}\bar{\mathbf{v}}$ даёт центр 3-мерного куба с ребрами $\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2, \bar{\mathbf{a}}_3$. Как нетрудно видеть, этим условиям удовлетворяет только случай, когда $\{\bar{\mathbf{a}}_i\}$ репер, определяющий центрированную кубическую решетку, в которой ребро куба является кратчайшим вектором Γ.

Этим полностью доказаны Лемма 1 и её Следствие.

Докажем лемму, которая вместе со Следствием леммы 1 даст доказательство нашей теоремы при $n=5$.

Лемма 2. *В 5-мерной решетке* $|\bar{\mathbf{a}}_5|\leqq|\bar{\mathbf{b}}_5|$.

Доказательство. Пусть вектор $\bar{\mathbf{v}}\notin\Gamma_5$ и $|\bar{\mathbf{v}}|<|\bar{\mathbf{a}}_5|$ (рис. 2 при $n=5$). Если записать вектор $\bar{\mathbf{v}}$ в репере $\{\bar{\mathbf{a}}_i\}$, то последняя его координата v_5 не равняется 0 и 1, поскольку $\bar{\mathbf{v}}\notin\Gamma_5^1$, если $|\bar{\mathbf{v}}|<|\bar{\mathbf{a}}_5|$. В то же время $v_5<3$, как следует из работы [4] то есть $v_5=2$. Но $\{\bar{\mathbf{b}}_i\}$ основной репер, поэтому существует по крайней мере один вектор $\bar{\mathbf{b}}_k \notin \Gamma_5$

— если $|\bar{\mathbf{b}}_k| \geqq |\bar{\mathbf{b}}_5|$, тогда $|\bar{\mathbf{b}}_5| \geqq |\bar{\mathbf{a}}_5|$,

— если $|\bar{\mathbf{b}}_k| < |\bar{\mathbf{a}}_5|$, тогда $\bar{\mathbf{b}}_k$ вектор типа $\bar{\mathbf{v}}$ и $b_{k5}=2$.

Вектор $\bar{\mathbf{a}}_5=(0, 0, 0, 0, 1)$ целочисленно выражается через векторы $\bar{\mathbf{b}}_i$. Отсюда следует, что пятые координаты у всех векторов $\bar{\mathbf{b}}_i$ не все могут быть четными числами, найдется хоть один $\bar{\mathbf{b}}_j$, у которого пятая координата — нечетное число. Это значит, что $\bar{\mathbf{b}}_j \notin \Gamma_5$ и $\bar{\mathbf{b}}_j \notin \Gamma_5^2$; откуда следует $|\bar{\mathbf{b}}_j| \geqq |\bar{\mathbf{a}}_5|$, т. е. $|\bar{\mathbf{b}}_5| \geqq |\bar{\mathbf{b}}_j| \geqq |\bar{\mathbf{a}}_5|$, что и требовалось доказать.

ЛИТЕРАТУРА

[1] Minkowski, H., Diskontinuitätsbereich für arithmetische Aquivalenz, *J. reine und angew. Math.* **129** (1905), 220—294.

[2] Делоне, Б. Н., Геометрия положительных квадратичных форм, *Успехи Мат. Наук* **3** (1937), 16—62; **4** (1938), 102—164.

[3] Рышков, С. С.—Барановский, Е. П., Классические методы теории решетчатых упаковок, *Успехи Мат. Наук* **34** 4 (1979), 3—63.

[4] Таммела, П. П., К теории приведения положительных квадратичных форм, *Исследования по теории чисел*, 3, *Зап. научных семинаров ЛОМО* **50** (1975), 6—97.

[5] Рышков, С. С., К теории приведения положительных квадратичных форм по Эрмиту—Минковскому, *Исследования по теории чисел*, 2, *Зап. научн. семинаров ЛОМИ* **33** (1973), 37—64.

[6] Рышков, С. С., О приведении положительных квадратичных форм от переменных по Эрмиту, по Минковскому и по Венкову, *Доклады АН СССР*, **207** (1972), 1054—56.

[7] Waerden, B. L. van der, Die Reduktionstheorie der quadratischen Formen, *Acta Mathem.* **96** (1956), 265—309.; **98** (1957), 3—4.

Кафедра Геометрии Университета им. Л. Этвеша, Будапешт

(Поступила 10-ого декабря 1979 г.)