# Contents

The aim of the journal **"Algebra and Discrete Mathematics"** is to present timely the state-of-the-art accounts on modern research in all areas of algebra (general algebra, semigroups, groups, rings and modules, linear algebra, algebraic geometry, universal algebras, homological algebra etc.) and discrete mathematics (combinatorial analysis, graphs theory, mathematical logic, theory of automata, coding theory, cryptography etc.)

***Languages:*** English.

***Preparing and submitting papers:***

Papers submitted for publication and editorial correspondence should be addressed to:

*Editorial office "Algebra and Discrete Mathematics",*
*Department of Algebra and Analysis,*
*Lugansk State Pedagogical Taras Shevchenko University,*
*vul. Oboronna 2, 91011, Lugansk, UKRAINE.*
The e-mail address is: `algebra@lgpu.lg.ua`
Web-site: `http://adm.lgpu.lg.ua`

Papers submitted to the journal **"Algebra and Discrete Mathematics"** (as ADM below) can be either compuscripts (done on a computer) or manuscripts (done on a typewriter). Handwritten papers will not be considered.

Compuscripts should be prepared in TeX, LaTeX or AMS–LaTeX; these will be used for production if accepted for publication. A macro package is available from our web-site for authors to prepare their papers in the ADM style.

The following information is required:

1. full postal addresses of all authors;
2. e-mails of all authors;
3. abstract;
4. 2001 AMS subject classification;
5. keywords and phrases.

Authors who wish to submit their papers electronically should send the TeX-file either to the ADM e-mail address or to anyone of the Editors as separately attached file (not in the letter body). The electronic addresses of the Editors are listed on the Editorial Board page.

Manuscripts prepared by other means must be typed on one side of the page, with double-spaced lines, long formulas displayed. Greek, German and script letters must be clearly differentiated from the text, and special signs must be explicitly explained.

***Proof-sheets:***

Authors receive only one set of proof-sheets (in PDF or PS format) via e-mail for corrections. Only correction of misprints and minor changes can be made during proofreading.

Algebra and Discrete Mathematics RESEARCH ARTICLE
Number 2. **(2003).** pp. 1–5


# Almost all derivative quivers of artinian biserial rings contain chains

**Tetjana Avdeeva and Olexandr Ganyushkin**

Communicated by V. V. Kirichenko

Abstract. A lower estimate for the number $M_n$ of all labelled quivers with $n$–vertex parts of Artinian biserial rings is given and the asymptotic of the relation $M_n/B_n$, where $B_n$ denotes the number of those quivers all connected components of which are cycles, is studied.

In the beginning of 70-s P.Gabriel [1] introduced a notion of a quiver of a finite dimensional algebra over an algebraically closed field — an directed graph of special type which in concise form preserves some very important information about the algebra. Using these graphs in [1] (see also Krugliak [2]) all finite dimensional algebras of finite type over an algebraically closed field with square zero radical are described. Later V.Kirichenko has expanded the construction of such an directed graph to right Noetherian semiperfect rings [3], and then to several other classes of rings (see, for example, [4], [5] and bibliography there). For some classes of rings it is convenient to consider a so called derivative quiver $RQ(A)$ (see [6]), which for the rings under consideration always turns out to be a simple bipartite graph with equicardinal part, instead of a quiver $Q(A)$ of a ring $A$.

In this connection there arises a natural problem of investigation of graphs which can be quivers of rings of some class. We will deal with Artinian biserial rings, first introduced by .Fuller [7]. A starting point for this paper is the following statement( [4], Corollary 5.15): *An Artinian ring A, with square zero Jacobson radical is biserial if and only if its derivative quiver $RQ(A)$ is a disconnected union of chains and cycles.*

**2000 Mathematics Subject Classification:** *05C30, 05C38, 16P20.*
**Key words and phrases:** *artinian rings, quivers, bipartite graphs.*

Therefore, a derivative quiver of an Artinian biserial ring is a simple bipartite graph with parts of the same cardinality, in which the degree of each vertex does not exceed 2. In [8] such graphs have been called *Artinian–biserial*, or just *AB–graphs*. An $AB$–graph with $n$–vertex parts is called *labelled*, if the vertices of each part are numbered from 1 to $n$ and it is indicated, which of the parts is *lower*, and which is *upper*. In what follows we consider only labelled $AB$–graphs.

In [8] the number $B_n$ of those $AB$–graphs with $n$–vertex parts, all connected components of which are cycles, is counted:

$$B_n = \sum_{\substack{(l_1,l_2,\dots,l_n)\\ 1l_1+2l_2+\dots+nl_n=n}} \frac{(n!)^2}{(l_1!)^2 \prod\limits_{k=2}^{n} (2^{l_k}\cdot k^{l_k}\cdot l_k!)}, \qquad (1)$$

and an upper bound for the number $M_n$ of all labelled $AB$–graphs with $n$–vertex parts is obtained:

$$M_n < \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{n!}{(n-k)!} \left|IS_{n-k}\right|^2,$$

where $\left|IS_n\right|$ — is the order of the inverse symmetric semigroup $IS_n$ of degree $n$. This estimate, however, is rather rough. Beside this, to give an estimate for the order in the right-hand side of the latter inequality for large values of $n$ is itself a difficult problem.

A more effective lower estimate for $M_n$ is given in the following

**Lemma.** *The number $M_n$ of all labelled $AB$–graphs with $n$–vertex parts satisfies the inequality $M_n > (n!)^2 \cdot \frac{n}{2}$.*

*Proof.* Since $M_1 = 2$ and $M_2 = 16$ then the statement is obvious for $n = 1$ and for $n = 2$. Let now $n \geq 3$ and suppose that for all $k < n$ the statement of Lemma is true. Consider those $AB$–graphs, which contain a sufficiently long chain of an odd length. Then exactly one of the endpoints of such a chain will belong to the lower part. To determine a chain of length $2n-2k-1$, one has to choose its endpoint in the lower part, then a vertex in the upper part incident to this endpoint, then the next vertex of a chain in the lower part, and so on, each time switching the part of the next vertex choice, till one reaches the $2n-2k$–s vertex of the chain which is its endpoint from the upper part. Since in this way one will get every chain of length $2n-2k-1$ exactly one time then the number of different chains of length $2n-2k-1$ equals

$$n\cdot n\cdot(n-1)\cdot(n-1)\cdot(n-2)\cdot(n-2)\cdots(k+1)\cdot(k+1) = \frac{(n!)^2}{(k!)^2}.$$

Since for $k \le \dfrac{n-1}{2}$ an $AB$–graph with $n$–vertex parts can not contain more than one chain of length $2n-2k-1$ then for such values of $k$ the number of $AB$–graphs, containing a chain of length $2n-2k-1$, equals $\dfrac{(n!)^2}{(k!)^2}\cdot M_k$. By the inductive assumption $M_k > (k!)^2 \cdot \dfrac{k}{2}$. Using the equality $M_1 = 2$, we can assume $M_k > (k!)^2$. It is easily seen, that an $AB$–graph with $n$–vertex parts can contain only one chain of length $\ge n$. Therefore,

$$M_n > \sum_{k=0}^{[(n-1)/2]} \left(\frac{n!}{k!}\right)^2 (k!)^2 = (n!)^2 \cdot \sum_{k=0}^{[(n-1)/2]} 1 = (n!)^2 \cdot \left[\frac{n-1}{2}\right] > (n!)^2 \cdot \frac{n}{2}.$$

□

**Theorem.** *Let $M_n$ be the number of all labelled $AB$–graphs with $n$–vertex parts, and $B_n$ — the number of those of such graphs, all connected components of which are cycles. Then* $\lim\limits_{n\to\infty} \dfrac{B_n}{M_n} = 0.$

*Proof.* Let us calculate an upper bound for $B_n$. It is known, that the number of permutations of a cycle type $(l_1, l_2, \dots, l_n)$ equals $n! \cdot \left(\prod_{k=1}^{n} (k^{l_k} \cdot l_k!)\right)^{-1}$. Since the number of all permutations is $n!$, then

$$\sum_{\substack{(l_1,l_2,\dots,l_n)\\ 1l_1+2l_2+\dots+nl_n=n}} \frac{n!}{\prod\limits_{k=1}^{n} (k^{l_k} \cdot l_k!)} = n!.$$

After cancellation of both sides by $n!$ we obtain:

$$\sum_{\substack{(l_1,l_2,\dots,l_n)\\ 1l_1+2l_2+\dots+nl_n=n}} \frac{1}{\prod\limits_{k=1}^{n} (k^{l_k} \cdot l_k!)} = 1.$$

This equality and an obvious inequality $l_1! \cdot \prod_{k=2}^{n} 2^{l_k} \ge 1$, imply:

$$\sum_{\substack{(l_1,l_2,\dots,l_n)\\ 1l_1+2l_2+\dots+nl_n=n}} \frac{1}{(l_1!)^2 \prod\limits_{k=2}^{n} (2^{l_k} \cdot k^{l_k} \cdot l_k!)} =$$

$$= \sum_{\substack{(l_1,l_2,\dots,l_n)\\ 1l_1+2l_2+\dots+nl_n=n}} \frac{1}{\prod\limits_{k=1}^{n} (k^{l_k} \cdot l_k!)} \cdot \frac{1}{l_1! \prod\limits_{k=2}^{n} 2^{l_k}} <$$

$$< \sum_{\substack{(l_1,l_2,\dots,l_n)\\ 1l_1+2l_2+\dots+nl_n=n}} \frac{1}{\prod\limits_{k=1}^{n} (k^{l_k} \cdot l_k!)} = 1.$$

This inequality and inequality 1 now imply, that $B_n < (n!)^2$. Therefore, using Lemma, we obtain:

$$0 \le \frac{B_n}{M_n} \le \frac{(n!)^2}{(n!)^2 \cdot \frac{n}{2}} = \frac{2}{n}.$$

Thus, $\lim\limits_{n\to\infty} \frac{B_n}{M_n} = 0.$ □

Following the tradition for usage of the expression 'almost all' (see, for example, [9]), we obtain the following

**Corollary.** *Almost all AB–graphs with n–vertex parts contain chains.*

We conclude by stating the values of $B_n$ and $M_n$ and of the relation $B_n/M_n$ for small values of $n$:

| $n$ | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| $B_n$ | 2 | 16 | 151 | 4991 |
| $M_n$ | 16 | 265 | 7343 | 304186 |
| $B_n/M_n$ | 0.125 | 0.0603773 | 0.0205638 | 0.0164077 |

## References


[1] *Gabriel P.* Unzerlegbare Darstellungen 1. Manuscript Math. **6** (1972), 71–103. [5]

[2] *Krugliak.S.A.* Representations of algebras with zero sguare radical. Zapiski nauchn. seminarov LOMI Ac.Sci. USSR **28** (1972), 80–89. (in Russian)

[3] *Kirichenko, V.V.* Generalized uniserial rings, Math. Sb. (N.S.), 99 (141), N 4, (1976), pp. 559-581. English translation, Math USSR Sb., v. 28, N 4, 1976, pp. 501-520.

[4] *Danlyev Kh.M., Kirichenko V.V., Haletskaja Z.P., Jaremenko Ju.V.* Weakly prime semiperfect 2–rings and modules over them. Algebraicheskie issledovania (collection of papers), Institut mathematiki NAN Ukrainy, Kiev, 1995, 5–32.

[5] *Gubareni N.M., Kirichenko V.V.* Rings and Modules. - Czestochowa, 2001.

[6] *Kirichenko V.V., Bernik O.Ja.* Semiperfect rings of distributive module type. Dopovidi Ac.Sci. URSR. 1988, no. 3, 15–17.

[7] *Fuller K.R.* Weakly symmetric rings of distributive module type. Comm. in Algebra. **5** (1977), 997–1008.

[8] *Avdeeva T.V., Ganyushkin O.G.* The number of quivers of Artinian biserial rings. Visnyk Kyivs'kogo Univ., ser. phiz.–math. sci., 1999, no. 3, 18–27. (in Ukrainian)

[9] *Korshunov A.D.* Main properties of random graphs with large number of vertices and edges. Uspehi mathem.nauk **40** (1985), 1(241), 107–173.(in Russian)

Contact information

**T. Avdeeva** Department of Physics and Mathematics, National of Technical University of Ukraine "Kyiv of Polytechnical Institute",37, Peremohy pr., Kyiv,UKRAINE

**O. Ganyushkin** Department of Mechanics and Mathematics, Kyiv Taras Shevchenko University, 64, Volodymyrska st., 01033, Kyiv, UKRAINE
*E-Mail:* ganiyshk@mechmat.univ.kiev.ua

Received by the editors: 06.12.2002.

Algebra and Discrete Mathematics
Number 2. **(2003).** pp. 6–19 RESEARCH ARTICLE


# Green’s relations on the seminearring of full hypersubstitutions of type $(n)$

**Th. Changphas, K. Denecke**

Communicated by V. M. Usenko

Abstract. Hypersubstitutions are mappings which are used to define hyperidentities and solid varieties. In this paper we will show that the set of all hypersubstitutions of a given type forms a seminearring. We will give a full characterization of Green’s relation $\mathcal{R}$ on a sub-seminearring of the seminearring $Hyp(n)$ of all hypersubstitutions of type $(n)$.

## 1. Introduction

Hypersubstitutions were introduced to make precise the concept of a hyperidentity and generalizations to $M$-hyperidentities. Let $\tau = (n_i)_{i \in I}$ be a type indexed by a set $I$, with operation symbols $f_i$ of arity $n_i \in \mathbb{N}$. Let $X = \{x_1, x_2, \ldots\}$ be a countably infinite set of variables and let $X_n = \{x_1, \ldots, x_n\}$ be a finite set. We denote by $W_\tau(X_n)$ the set of all $n$-ary terms of type $\tau$ over the alphabet $X_n$ and by $W_\tau(X)$ the set of all terms of type $\tau$.
An identity $s \approx t, s, t \in W_\tau(X)$, of type $\tau$ is called a *hyperidentity* of a variety $V$ of type $\tau$ if for every substitution of $n$-ary terms for the $n$-ary operation symbols in $s \approx t$, the resulting identity holds in $V$. This shows that we are interested in mappings which associate to every operation symbol $f_i$ of a given type $\tau$ a term $\sigma(f_i)$ of type $\tau$ of the same arity as $f_i$. Any such map is called a *hypersubstitution of type* $\tau$.

**2001 Mathematics Subject Classification:** *08B15.*
**Key words and phrases:** *full hypersubstitutions, seminearring, Green’s relations.*

Hypersubstitutions can be uniquely extended to mappings $\hat{\sigma} : W_\tau(X) \to W_\tau(X)$ which are inductively defined by the following steps:

(i) $\hat{\sigma}[x] := x$ if $x \in X$,

(ii) $\hat{\sigma}[t] := \sigma(f_i)(\hat{\sigma}[t_1], \ldots, \hat{\sigma}[t_{n_i}])$ if $t = f_i(t_1, \ldots, t_{n_i})$.

Using this extension we can define a binary operation $\circ_h$ on the set $Hyp(\tau)$ of all hypersubstitutions of type $\tau$ by $\sigma_1 \circ_h \sigma_2 := \hat{\sigma}_1 \circ \sigma_2$, where $\circ$ is the usual composition of operations. By $\sigma_{id}$ we denote the identity hypersubstitution, mapping each operation symbol $f_i$ to the term $f_i(x_1, \ldots x_{n_i})$. This gives the monoid $(Hyp(\tau); \circ_h, \sigma_{id})$.

If $M$ is any submonoid of the monoid $Hyp(\tau)$, then an identity $u \approx v$ of a variety $V$ is called an *$M$-hyperidentity* of $V$ if for every $\sigma \in M$ the equation $\hat{\sigma}[u] \approx \hat{\sigma}[v]$ holds in $V$. A variety $V$ is called *$M$-solid* if every identity of $V$ is an $M$-hyperidentity in $V$. The collection of all $M$-solid varieties of type $\tau$ forms a complete sublattice of the lattice of all varieties of type $\tau$. Actually, there is a Galois connection between submonoids of $Hyp(\tau)$ and complete sublattices of the lattice of all varieties of type $\tau$. For more background see [3]. This shows the importance of studying the properties of the monoid $Hyp(\tau)$ and its submonoids. In [2] the authors started to investigate the semigroup properties of the monoid $Hyp(2)$ of all hypersubstitutions of type $\tau = (2)$, especially Green's relations on $Hyp(2)$. We want to continue these investigations with $Hyp(n)$, the monoid of all hypersubstitutions of type $\tau = (n)$, and a particular submonoid of $Hyp(n)$, the monoid of all so-called *full hypersubstitutions*.

It turns out that one can define a second binary operation on $Hyp(\tau)$ such that $Hyp(\tau)$ forms a seminearring. By

$$(\sigma_1 + \sigma_2)(f_i) := \sigma_2(f_i)(\sigma_1(f_i), \ldots, \sigma_1(f_i))$$

we define a hypersubstitution which maps the $n_i$-ary operation symbol $f_i$ to the $n_i$-ary term $\sigma_2(f_i)(\sigma_1(f_i), \ldots, \sigma_1(f_i))$ for every $i \in I$. The term $\sigma_2(f_i)(\sigma_1(f_i), \ldots, \sigma_1(f_i))$ is $n_i$-ary. Therefore $\sigma_1 + \sigma_2$ is a hypersubstitution of $Hyp(\tau)$. We show that the operation $+$ is associative. Indeed, we have

$$\begin{gathered}((\sigma_1 + \sigma_2) + \sigma_3)(f_i) = \sigma_3(f_i)((\sigma_1 + \sigma_2)(f_i), \ldots, (\sigma_1 + \sigma_2)(f_i)) = \\ = \sigma_3(f_i)(\sigma_2(f_i)(\sigma_1(f_i), \ldots, \sigma_1(f_i)), \ldots, \sigma_2(f_i)(\sigma_1(f_i), \ldots, \sigma_1(f_i))) = \\ = \sigma_3(f_i)(\sigma_2(f_i), \ldots, \sigma_2(f_i))(\sigma_1(f_i), \ldots, \sigma_1(f_i)) = \\ = (\sigma_2 + \sigma_3)(f_i)(\sigma_1(f_i), \ldots, \sigma_1(f_i)) = (\sigma_1 + (\sigma_2 + \sigma_3))(f_i).\end{gathered}$$

Here we used properties of the superposition of terms. In a similar way we prove that $\sigma \circ_h (\sigma_1 + \sigma_2) = (\sigma \circ_h \sigma_1) + (\sigma \circ_h \sigma_2)$.

Indeed, we have

$$\begin{gathered}(\sigma \circ_h (\sigma_1 + \sigma_2))(f_i) = \hat{\sigma}[(\sigma_1 + \sigma_2)(f_i)] = \\ = \hat{\sigma}[\sigma_2(f_i)(\sigma_1(f_i), \ldots, \sigma_1(f_i))] = \\ = \hat{\sigma}[\sigma_2(f_i)](\hat{\sigma}[\sigma_1(f_i)], \ldots, \hat{\sigma}[\sigma_1(f_i)]) = \\ = (\sigma \circ_h \sigma_2)(f_i)((\sigma \circ_h \sigma_1)(f_i), \ldots, (\sigma \circ_h \sigma_1)(f_i)) \\ ((\sigma \circ_h \sigma_1) + (\sigma \circ_h \sigma_2))(f_i)\end{gathered}$$

if we use that hypersubstitution and superposition are permutable. This shows the left distributivity.
The following counterexample shows that the right distributive identity is not satisfied.
Assume that $\tau = (2)$, with a binary operation symbol $f$, and that $\sigma_1, \sigma_2, \sigma_3$ are defined by $\sigma_1(f) = f(x, y)$, $\sigma_2(f) = f(y, x)$, $\sigma_3(f) = f(x, f(y, y))$. Then

$$\begin{gathered}(\sigma_1 + \sigma_2)(f) = f(f(x, y), f(x, y)), \\ ((\sigma_1 + \sigma_2) \circ_h \sigma_3)(f) = (\sigma_1 + \sigma_2)\hat{}\,[f(x, f(y, y))], \\ = f(f(x, f(f(y, y), f(y, y))), f(x, f(f(y, y), f(y, y)))), \\ (\sigma_1 \circ_h \sigma_3)(f) = f(x, f(y, y)), \\ (\sigma_2 \circ_h \sigma_3)(f) = f(f(y, y), x), \\ ((\sigma_1 \circ_h \sigma_3) + (\sigma_2 \circ_h \sigma_3))(f) = \\ (\sigma_2 \circ_h \sigma_3)(f)((\sigma_1 \circ_h \sigma_3)(f), (\sigma_1 \circ_h \sigma_3)(f)) \\ = f(f(y, y), x)(f(x, f(y, y)), f(x, f(y, y))) \\ = f(f(f(x, f(y, y)), f(x, f(y, y))), f(x, f(y, y))).\end{gathered}$$

On the set $Hyp(\tau)$ not only operations, but also relations can be defined. Let $\sigma_1, \sigma_2 \in Hyp(\tau)$. Then we define $\sigma_1 \preceq_{\mathcal{R}} \sigma_2$ if and only if there is a hypersubstitution $\sigma$ such that $\sigma_1 = \sigma_2 \circ_h \sigma$. Since $Hyp(\tau)$ is a monoid, $\preceq_{\mathcal{R}}$ is reflexive and transitive, i.e. a quasiorder.
Similarly, we define $\sigma_1 \preceq_{\mathcal{L}} \sigma_2$ if and only if there is a hypersubstitution $\sigma$ such that $\sigma_1 = \sigma \circ_h \sigma_2$. The relation $\preceq_{\mathcal{L}}$ is also a quasiorder. Then it is easy to see (and well-known) that $\mathcal{R} = \preceq_{\mathcal{R}} \cap \preceq_{\mathcal{R}}^{-1}$ and $\mathcal{L} = \preceq_{\mathcal{L}} \cap \preceq_{\mathcal{L}}^{-1}$ are equivalence relations and are called *Green's relations* $\mathcal{R}$ and $\mathcal{L}$. The relations $\preceq_{\mathcal{L}}$ and $\preceq_{\mathcal{R}}$ induce partial order relations on the quotient sets $Hyp(\tau)/\mathcal{R}$ and $Hyp(\tau)/\mathcal{L}$, respectively.
Green's relations $\mathcal{H}$ and $\mathcal{D}$ are defined by $\mathcal{H} := \mathcal{R} \cap \mathcal{L}$ and $\mathcal{D} := \mathcal{R} \circ \mathcal{L}$ and $\mathcal{J}$ is defined by

$$\sigma_1 \mathcal{J} \sigma_2 :\Leftrightarrow \exists \sigma, \sigma', \gamma, \gamma' \in Hyp(\tau)\ (\sigma_1 = \sigma \circ_h \sigma_2 \circ_h \sigma', \sigma_2 = \gamma \circ_h \sigma_1 \circ_h \gamma').$$

We recall the following properties of Green's relations $\mathcal{R}$ and $\mathcal{L}$:

**Proposition 1.1.** *([7])*

*(i)* $\mathcal{R} \circ \mathcal{L} = \mathcal{L} \circ \mathcal{R}$,

*(ii)* $\mathcal{R}$ *is a left congruence on* $Hyp(\tau)$,

*(iii)* $\mathcal{L}$ *is a right congruence on* $Hyp(\tau)$.

In [5] we considered special submonoids of $Hyp(\tau)$. The images of the hypersubstitutions from these submonoids are called *full terms* and *strongly full terms* and are inductively defined as follows:

**Definition 1.2.** *Let* $f_i, i \in I$, *be an* $n_i$*-ary operation symbol and let* $s : \{1, \ldots, n_i\} \to \{1, \ldots, n_i\}$ *be a permutation, then*

*(i)* $f_i(x_{s(1)}, \ldots, x_{s(n_i)})$ *is a full term and*

*(ii) if* $f_j, j \in J$, *is an* $n_j$*-ary operation symbol and if* $t_1, \ldots, t_{n_j}$ *are full terms, then* $f_j(t_1, \ldots, t_{n_j})$ *is a full term.*

*Let* $W_\tau^f(X)$ *be the set of all full terms of type* $\tau$.

*Strongly full* terms are defined as follows:

**Definition 1.3.** *(i) For every* $n_i$*-ary operation symbol* $f_i, i \in I$ *the term* $f_i(x_1, \ldots, x_{n_i})$ *is strongly full,*

*(ii) if* $t_1, \ldots, t_{n_i}$ *are strongly full and if* $f_i, i \in I$, *is an* $n_i$*-ary operation symbol, then* $f_i(t_1, \ldots, t_{n_i})$ *is strongly full.*

*We denote by* $W_\tau^{sf}(X)$ *the set of all strongly full terms of type* $\tau$.

**Definition 1.4.** *A hypersubstitution* $\sigma$ *is called full if* $\sigma(f_i) \in W_\tau^f(X_{n_i})$ *for all* $i \in I$ *and strongly full if for all* $i \in I$ *the images of the operation symbols* $f_i$ *belong to* $W_\tau^{sf}(X_{n_i})$. *By* $Hyp^f(\tau)$ *and* $Hyp^{sf}(\tau)$, *respectively, we denote the set of all full and the set of all strongly full hypersubstitutions of type* $\tau$.

Then in [5] it was proved:

**Lemma 1.5.** *The sets* $Hyp^f(\tau)$ *and* $Hyp^{sf}(\tau)$ *form submonoids of the monoid* $Hyp(\tau)$.

## 2. Green's Relations and the Complexity of a Hypersubstitutions

There are different ways to measure the complexity of a term. If $vb(t)$ denotes the number of variables occurring in the term $t$, then $vb$ is an example for a complexity measure of terms. The most commonly used measurement of the complexity of terms of type $\tau$ is the usual depth of a term which is defined inductively by the following steps:

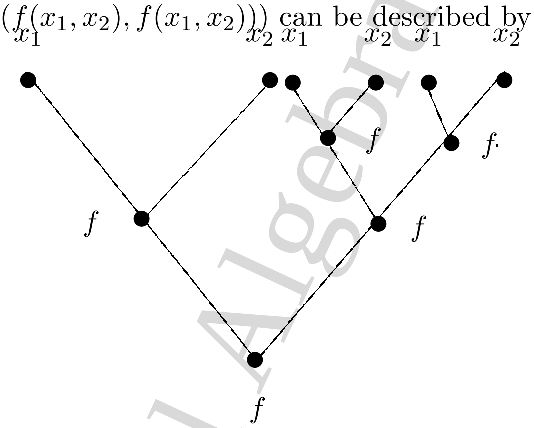
(i) If $t = x \in X$, then

$$depth(t) = 0.$$

(ii) If $t = f_i(t_1, \ldots, t_{n_i})$, then

$$depth(t) = max\{depth(t_1), \ldots, depth(t_{n_i})\} + 1.$$

To decribe some more complexity functions we often identify a term with the tree diagram used to represent it. For instance, the term $f(f(x_1, x_2), f(f(x_1, x_2), f(x_1, x_2)))$ can be described by the diagram:




Then the depth of a term $t$ is the length of the longest path from the root to a vertex labelled by a variable (a leaf) in the tree. Instead of $depth(t)$ one can also speak of the $maxdepth(t)$. The minimum depth of a term $t$, denoted by $mindepth(t)$, is the length of the shortest path from the root to a leaf in the tree and is defined inductively by

(i) If $t$ is a variable,then

$$mindepth(t) = 0.$$

(ii) If $t = f_i(t_1, \ldots, t_{n_i})$,then

$$mindepth(t) = min\{mindepth(t_1), \ldots, mindepth(t_{n_i})\} + 1.$$

Using the depth and the mindepth we can define another kind of terms.

**Definition 2.1.** *Let $t \in W_\tau(X_n)$ be an n-ary term of type $\tau$. Then t is called path-regular (for short a pr-term) if $mindepth(t) = depth(t)$. Let $W_\tau^{pr}(X_n)$ be the set of all n-ary pr-terms of type $\tau$. A hypersubstitution is called path-regular, if the terms $\sigma(f_i)$ are path-regular for every $i \in I$. Let $Hyp^{pr}(\tau)$ be the set of all path-regular hypersubstitutions.*

In [5] it was proved:

**Proposition 2.2.** *$Hyp^{pr}(\tau)$ forms a submonoid of $Hyp(\tau)$.*

All these complexity measures are particular cases of the following valuation of terms:

**Definition 2.3.** *Let $\mathcal{F}_\tau(X) = (W_\tau(X); (\bar{f}_i)_{i\in I})$ with $\bar{f}_i : \{t_1, \ldots, t_{n_i}\} \mapsto f_i(t_1, \ldots, t_{n_i})$ be the absolutely free term algebra of type $\tau$ on a countable set $X$, and let $\mathbb{N}_\tau = (\mathbb{N}; (f_i^{\mathbb{N}})_{i\in I})$ be an algebra of type $\tau$ defined on the set of all natural numbers. Then a mapping $v : X \to \mathbb{N}_\tau$ is called a valuation of terms of type $\tau$ into $\mathbb{N}_\tau$ if the following conditions are satisfied:*

*(i) $v(x) = a$ if $x \in X$ and $a \in \mathbb{N}$.*

*(ii) $v(t) \geq v(x)$ for every variable $x$ and every term $t$ (see [6]).*

From the freeness of $\mathcal{F}_\tau(X)$ we obtain a uniquely determined homomorphism $\hat{v} : \mathcal{F}_\tau(X) \to \mathbb{N}_\tau$ which extends $v$. For short, we denote this homomorphism also by $v$ and will call it valuation of terms. In the case of $depth(t)$ the operations $f_i^{\mathbb{N}}$ are defined by $f_i^{\mathbb{N}}(a_1, \ldots, a_{n_i}) = max\{a_1, \ldots, a_{n_i}\} + 1$ and for $mindepth(t)$ we have $f_i^{\mathbb{N}}(a_1, \ldots, a_{n_i}) = min\{a_1, \ldots, a_{n_i}\} + 1$. Both kinds of operations are monotone with respect to the usual order $\leq$ on $\mathbb{N}$.

So, in many case the mapping $v$ satisfies the following condition

(OC) If $a_j \leq b_j$ for $1 \leq j \leq n_i$ and $f_i$ is an $n_i$-ary operation symbol of type $\tau$ then for the corresponding operations $f_i^{\mathbb{N}}$ we have $f_i^{\mathbb{N}}(a_1, \ldots, a_{n_i}) \leq f_i^{\mathbb{N}}(b_1, \ldots, b_{n_i})$ Here we denote by $\leq$ the usual order on the set of natural numbers.

For more background on valuation of terms see [6].

This can be applied to hypersubstitutions as follows:

**Definition 2.4.** *Let $\sigma$ be a hypersubstitution of type $\tau$. Then*

$$depth(\sigma) := min\{depth(\sigma(f_i)) \mid i \in I\}$$
$$v(\sigma) := min\{v(\sigma(f_i)) \mid i \in I\}.$$

For the type $\tau = (n)$ and for full terms in [4] it was proved:

**Proposition 2.5.** *Let $\tau = (n), n \geq 1$ and let $t \in W_\tau^f(X)$ be a full term. Then*

$$depth(\hat{\sigma}[t]) = depth(\sigma(f))depth(t).$$

As a consequence we obtain:

**Proposition 2.6.** *Assume that $\tau = (n)$. The mapping $depth : Hyp^f(n) \to \mathbb{N}^+$ defined by $\sigma \mapsto depth(\sigma)$ is a homomorphism of $(Hyp^f(n); \circ_h, \sigma_{id})$ onto the monoid $(\mathbb{N}^+; \cdot, 1)$ of all positive integers.*

*Proof.* The mapping *depth* is well-defined and for every natural number $n \geq 1$ there is a full hypersubstitution $\sigma$ with $depth(\sigma) = n$. Therefore *depth* is surjective. Further we have

$$depth(\sigma_1 \circ_h \sigma_2) = depth((\sigma_1 \circ_h \sigma_2)(f)) = depth(\hat{\sigma_1}[\sigma_2(f)])$$
$$= depth(\sigma_1(f))depth(\sigma_2(f)) = depth(\sigma_1)depth(\sigma_2)$$

by Proposition 2.5 and

$$depth(\sigma_{id}) = depth(\sigma_{id}(f)) = depth(f(x_1, \ldots, x_n)) = 1.$$

□

By the homomorphism theorem $(\mathbb{N}^+; \cdot, 1)$ is isomorphic to the quotient monoid $Hyp^f(n)/kerdepth$ with $kerdepth = \{(\sigma_1, \sigma_2) \mid \sigma_1, \sigma_2 \in Hyp^f(n), depth(\sigma_1) = depth(\sigma_2)\}$.

Proposition 2.5 has some consequences for Green's relations. First of all, if $\sigma_1 = \sigma_2 \circ_h \sigma_3$, then by Proposition 2.5 $depth(\sigma_2)$ divides $depth(\sigma_1)$ and $depth(\sigma_3)$ divides $depth(\sigma_1)$. One more consequence of Proposition 2.5 is that the monoid $(Hyp^f(n); \circ_h, \sigma_{id})$ is not finitely generated. The homomorphism *depth* maps any generating system of $Hyp^f(n)$ onto a generating system of $(\mathbb{N}^+; \cdot, 1)$. Every generating system of $(\mathbb{N}; \cdot, 1)$ contains the infinite set of all prime numbers. This shows that there is no finite generating system of $(Hyp^f(n); \circ_h, \sigma_{id})$.
The homomorphism *depth* preserves also the quasiorders $\preceq_\mathcal{R}$ and $\preceq_\mathcal{L}$ on $Hyp(\tau)$ since

$$\sigma_1 \preceq_\mathcal{R} \sigma_2 \Rightarrow \exists \rho \in Hyp(\tau)(\sigma_1 = \sigma_2 \circ_h \rho) \Rightarrow$$
$$\Rightarrow depth(\sigma_2) | depth(\sigma_1) \Rightarrow depth(\sigma_2) \leq depth(\sigma_1)$$

(Here $\leq$ denotes the usual order on $\mathbb{N}$).

**Corollary 2.7.** *Let $\sigma_1, \sigma_2 \in Hyp^f(n)$ and let $\mathcal{K} \in \{\mathcal{R}, \mathcal{L}, \mathcal{H}, \mathcal{D}, \mathcal{J}\}$, then $\sigma_1 \mathcal{K} \sigma_2$ implies $depth(\sigma_1) = depth(\sigma_2)$.*

*Proof.* For $\mathcal{K} = \mathcal{R}$ we have

$$\begin{aligned} \sigma_1 \mathcal{R} \sigma_2 &\Longrightarrow \exists \rho, \rho' \in Hyp(\tau)\ (\sigma_1 = \sigma_2 \circ_h \rho \wedge \sigma_2 = \sigma_1 \circ_h \rho') \\ &\Longrightarrow depth(\sigma_2) \leq depth(\sigma_1) \wedge depth(\sigma_1) \leq depth(\sigma_2) \\ &\Longrightarrow depth(\sigma_1) = depth(\sigma_2). \end{aligned}$$

For the relation $\mathcal{L}$ we conclude in a similar way.
For $\mathcal{H}$ we have:

$$\sigma_1 \mathcal{H} \sigma_2 \Longrightarrow \sigma_1 (\mathcal{R} \cap \mathcal{L}) \sigma_2 \Longrightarrow \sigma_1 \mathcal{R} \sigma_2 \wedge \sigma_1 \mathcal{L} \sigma_2 \Longrightarrow depth(\sigma_1) = depth(\sigma_2).$$

Considering $\mathcal{D}$ we obtain:

$$\begin{aligned} \sigma_1 \mathcal{D} \sigma_2 \Longrightarrow \sigma_1 (\mathcal{R} \circ \mathcal{L}) \sigma_2 &\Longrightarrow \exists \sigma \in Hyp(\tau)(\sigma_1 \mathcal{L} \sigma \wedge \sigma \mathcal{R} \sigma_2) \\ &\Longrightarrow depth(\sigma_1) = depth(\sigma) \wedge depth(\sigma) = depth(\sigma_2). \end{aligned}$$

Finally, for $\mathcal{J}$ we get

$$\begin{aligned} \sigma_1 \mathcal{J} \sigma_2 &\Rightarrow \exists \sigma, \sigma', \gamma, \gamma' \in Hyp(\tau)\ (\sigma_1 = \sigma \circ_h \sigma_2 \circ_h \sigma' \wedge \sigma_2 = \gamma \circ_h \sigma_1 \circ_h \gamma') \\ &\Rightarrow depth(\sigma_1) | depth(\sigma_2) \wedge depth(\sigma_2) | depth(\sigma_1) \\ &\Rightarrow depth(\sigma_1) = depth(\sigma_2). \end{aligned}$$

□

For the depth of the hypersubstitutions $\rho, \rho', \gamma, \gamma'$ which are needed in the definitions of $\mathcal{R}$, $\mathcal{L}$ and $\mathcal{J}$ we have $depth(\rho) = depth(\rho') = depth(\gamma) = depth(\gamma') = 1$.

Further we have

**Corollary 2.8.** *$\sigma \in Hyp^f(n)$ is invertible if and only if $depth(\sigma) = 1$ and idempotent if and only if $\sigma = \sigma_{id}$.*

*Proof.* If $\sigma$ is invertible, then there exists a hypersubstitution $\sigma'$ such that $\sigma \circ \sigma' = \sigma' \circ \sigma = \sigma_{id}$. Now from Proposition 2.5 we obtain $depth(\sigma) \cdot depth(\sigma') = 1$ and then $depth(\sigma) = 1$. If conversely $depth(\sigma) = 1$, then there is a permutation $s$ on the set $\{1, \ldots, n\}$ such that $\sigma(f) = f(x_{s(1)}, \ldots, x_{s(n)})$, but then the hypersubstitution $\sigma'$ with $\sigma'(f) = = f(x_{s^{-1}(1)}, \ldots, x_{s^{-1}(n)})$ satisfies

$$\begin{aligned} (\sigma \circ_h \sigma')(f) &= f(x_{(s \circ s^{-1})(1)}, \ldots, x_{(s \circ s^{-1})(n)}) = f(x_1, \ldots, x_n) = \\ &= \sigma_{id}(f) = f(x_{(s^{-1} \circ s)(1)}, \ldots, x_{(s^{-1} \circ s)(n)}) = (\sigma' \circ \sigma)(f). \end{aligned}$$

Therefore $\sigma$ is invertible. The second proposition follows from

$$\sigma^2 = \sigma \Rightarrow depth(\sigma^2) = depth(\sigma) \Rightarrow depth(\sigma) = 1 \Rightarrow$$
$$\Rightarrow \exists s \in S_n(\sigma(f) = f(x_{s(1)}, \ldots, x_{s(n)})) \Rightarrow s = \sigma_{id}.$$

□

Clearly, the set of all invertible elements of $(Hyp^f(n); \circ_h, \sigma_{id})$ is the maximal subgroup of $(Hyp^f(n); \circ_h, \sigma_{id})$ and is isomorphic to the full symmetric group $S_n$ of all permutations on $\{1, \ldots, n\}$.

Assume that $\sigma_1, \sigma_2$ are full hypersubstitutions. In [4] it was proved that the superposition of full terms is full. Therefore $\sigma_1 + \sigma_2$ is a full hypersubstitution and we have:

**Theorem 2.9.** $(Hyp^f(n); \circ_h, +, \sigma_{id})$ *is a left-seminearring with identity and the function depth is a homomorphism onto the semiring* $(\mathbb{N}^+; \cdot, +, 1)$ *of natural numbers with identity.*

*Proof.* We proved already that all defining identities of a left-seminearring are satisfied. By Definition 2.4 we have

$$\begin{aligned} depth(\sigma_1 + \sigma_2) &= depth(\sigma_1 + \sigma_2)(f) \\ &= depth(\sigma_2(f)(\sigma_1(f), \ldots, \sigma_1(f))) \\ &= depth(\sigma_1) + depth(\sigma_2). \end{aligned}$$

The rest follows from Proposition 2.5. □

Seminearrings were considered in [8] and [9].

Further we have

**Proposition 2.10.** *The structures* $(Hyp^{sf}(n); \circ_h, +, \sigma_{id})$ *and* $(Hyp^{pr}(n); \circ_h, +, \sigma_{id})$ *are left-seminearrings and* $(Hyp^{sf}(n); \circ_h, +, \sigma_{id})$ *is a sub-left-seminearring of the left-seminearring* $(Hyp^f(n); \circ_h, +, \sigma_{id})$.

*Proof.* $(Hyp^{sf}(n); \circ_h, \sigma_{id})$ is a submonoid of $(Hyp^f(n); \circ_h, +, \sigma_{id})$. Assume that $\sigma_1, \sigma_2 \in Hyp^f(n)$. Then $\sigma_2(f)(\sigma_1(f), \ldots, \sigma_1(f))$. If we substitute for $x_1$ the strongly full term $\sigma_1(f)$, etc., and finally for $x_n$ the strongly full term $\sigma_1(f)$, then by the inductive definition of strongly full terms the resulting term $\sigma_2(f)(\sigma_1(f)), \ldots, \sigma_1(f))$ is strongly full and $\sigma_1 + \sigma_2 \in Hyp^{sf}(n)$.
Assume now that $\sigma_1, \sigma_2 \in Hyp^{pr}(n)$, i.e.

$$mindepth(\sigma_2(f)) = depth(\sigma_j(f)), j = 1, 2.$$

Then we have also

$$\begin{aligned} mindepth((\sigma_1 + \sigma_2)(f)) &= mindepth(\sigma_2(f)(\sigma_1(f), \ldots, \sigma_1(f))) \\ &= mindepth(\sigma_1(f)) + mindepth(\sigma_2(f)) \end{aligned}$$

and thus

$$mindepth(\sigma_1+\sigma_2) = mindepth(\sigma_1) + mindepth(\sigma_2) = \\ = depth(\sigma_1) + depth(\sigma_2) = depth(\sigma_1+\sigma_2)$$

and then

$$\sigma_1+\sigma_2 \in Hyp^{pr}(n).$$

□

Further we have

**Proposition 2.11.** *$(Hyp^{sf}(n); \circ_h, +) \cap (Hyp^{pr}(n); \circ_h, +)$ is the left-seminearring generated by $\sigma_{id}$.*

*Proof.* The one-element set $\sigma_{id}$ is closed under the multiplication. Therefore, since the addition of hypersubstitutions is associative, every element of the left-seminearring generated by $\sigma_{id}$ can be written as $n\sigma_{id}$ for some natural number $n$. We show by induction on $n$ that every hypersubstitution of the form $n\sigma_{id}$ belongs to $Hyp^{sf}(n)$ and to $Hyp^{pr}(n)$ and therefore to the intersection. This is clear for $\sigma_{id}$. Assume that $n\sigma_{id} \in Hyp^{sf}(n) \cap Hyp^{pr}(n)$, then $(n+1)\sigma_{id}(f) = n\sigma_{id} + \sigma_{id}(f) = \sigma_{id}(f)(n\sigma_{id}(f), \ldots, n\sigma_{id}(f)) = f(n\sigma_{id}(f), \ldots, n\sigma_{id}(f))$ and by the definition of strongly full and of path-regular hypersubstitutions we have $(n+1)\sigma_{id} \in Hyp^{sf}(n) \cap Hyp^{pr}(n)$.
Conversely, assume that $\sigma \in Hyp^{sf}(n) \cap Hyp^{pr}(n)$. Since $\sigma$ is full, there are terms $t_1, \ldots, t_n$ such that $\sigma(f) = f(t_1, \ldots, t_n)$. We give a proof by $depth(\sigma)$. If $depth(\sigma) = 1$, i.e. $depth(\sigma(f)) = 1$, then $\sigma(f) = f(x_1, \ldots, x_n) = \sigma_{id}(f)$ and thus $\sigma \in \langle\sigma_{id}\rangle$. Assume that every $\sigma$ with $depth(\sigma) = n$ belongs to $\langle\sigma_{id}\rangle$ and let $\sigma'$ be a hypersubstitution with $depth(\sigma') = n+1$. Thus $\sigma'(f) = f(t_1, \ldots, t_n)$ where $t_1, \ldots, t_n$ are full and path-regular terms. Consider the hypersubstitutions $\sigma_1, \ldots, \sigma_n$ with $\sigma_i(f) = t_i$. Since $depth(\sigma_i) = n$, we have $\sigma_i \in \langle\sigma_{id}\rangle$. If there numbers $n_i$ such that $\sigma_j = n_i\sigma_{id}, n_i \neq n_j$ for $i \neq j$, then

$$min\{mindepth(n_1\sigma_{id}), \ldots, mindepth(n_n\sigma_{id})\} = \\ = max\{depth(n_1\sigma_{id}), \ldots, depth(n_n\sigma_{id})\}.$$

Therefore for every $i = 1, \ldots, n$ we have $\sigma_i = n\sigma_{id}$ and therefore $\sigma' = (n+1)\sigma_{id} \in \langle\sigma_{id}\rangle$. This shows that $\langle\sigma_{id}\rangle = Hyp^{sf}(n) \cap Hyp^{pr}(n)$. □

We remark that the set of all hypersubstitutions of arbitrary type $\tau$ is also closed under our addition and is called a left-seminearring since the proof of the associativity and left distributivity did not use the type

$(n)$. Then a consequence of Proposition 2.10 is that the left- seminearring $(Hyp^f(n); \circ_h, +, \sigma_{id})$ has no finite sub-left-seminearring and that every left-seminearring of hypersubstitutions contains the infinite left-seminearring $\langle \sigma_{id} \rangle = Hyp^{sf}(n)$.

Remark further that the mapping $n\sigma_{id} \mapsto n$ defines an isomorphism between $\langle \sigma_{id} \rangle$ and $(\mathbb{N}; +, \cdot, 1)$. This show that $\langle \sigma_{id} \rangle$ is a semiring with cancellation rules for both operations, and with commutative addition. Furthermore $\langle \sigma_{id} \rangle$ has no zero-divisors. Now we want to generalize Corollary 2.7 to the valuation of terms of type $\tau$ into $\mathcal{N}_\tau$ introduced in Definition 2.3.

We mention the following Fact which was proved in [6]:

**Fact**: Let $v$ be a valuation of terms into $\mathcal{N}_\tau$ which satisfies the condition (OC). Then for any $n$-ary terms $t_1, \ldots, t_m$ and for an arbitrary $m$-ary term $t$ we have

$$v(t(t_1, \ldots, t_m)) \geq v(t).$$

Now we prove:

**Proposition 2.12.** *Let $\sigma_1, \sigma_2 \in Hyp^f(n)$. If $\sigma_1 \mathcal{R} \sigma_2$, then for every valuation which satisfies (OC) we have $v(\sigma_1) = v(\sigma_2)$.*

*Proof.* If $\sigma_1 \mathcal{R} \sigma_2$, then there exist hypersubstitutions $\sigma, \sigma' \in Hyp^f(n)$ such that $\sigma_1 = \sigma_2 \circ_h \sigma$ and $\sigma_2 = \sigma_1 \circ_h \sigma'$ and therefore from $\sigma_1(f) = (\sigma_2 \circ_h \sigma)(f)$ follows $\sigma_1(f) = \hat{\sigma}_2[\sigma(f)]$ and $\sigma_2(f) = \hat{\sigma}_1[\sigma'(f)]$. Since $\tau = n$ and since $\sigma \in Hyp^f(n)$, the term $\sigma(f)$ has the form $f(t_1, \ldots, t_n)$ and then $\hat{\sigma}_2[\sigma(f)]$ has the form $\sigma_2(f)(\hat{\sigma}_2[t_1], \ldots, \hat{\sigma}_2[t_n])$. Applying the Fact we see that $v((\sigma_2 \circ \sigma)(f)) \geq v(\sigma_2(f))$ and then $v(\sigma_1) \geq v(\sigma_2)$. Using $\sigma_2 = \sigma_1 \circ_h \sigma'$ we get $v(\sigma_2) \geq v(\sigma_1)$. Altogether, we have $v(\sigma_1) = v(\sigma_2)$. □

Clearly, if $\sigma_1 \mathcal{H} \sigma_2$ and if $v$ satisfies (OC) we have also $v(\sigma_1) = v(\sigma_2)$. Because of $(\sigma_1 + \sigma_2)(f) = \sigma_2(f)(\sigma_1(f), \ldots, \sigma_1(f))$ from the fact follows that $v(\sigma_1 + \sigma_2) \leq v(\sigma_2)$, while $v(\sigma_1 \circ_h \sigma_2) \leq v(\sigma_1)$.

## 3. A Characterization of Green's relation $\mathcal{R}$

The condition $depth(\sigma) = depth(\sigma')$ turns out to be necessary, but not sufficient for $\sigma_1 \mathcal{R} \sigma_2$. Indeed, if $\sigma_1, \sigma_2 \in Hyp^f(2)$ and $\sigma_1 \neq \sigma_2$ then $\sigma_1 \mathcal{R} \sigma_2$ implies $\sigma_1 = \sigma_2 \circ_h \sigma_{x_1 x_2}$ or $\sigma_2 = \sigma_1 \circ_h \sigma_{x_2 x_1}$. For instance, for type $\tau = (2)$ the hypersubstitutions $\sigma_1$ with

$$\sigma_1(f) = f(f(x_1, x_2), f(f(x_1, x_2), f(x_2, x_1)))$$

and $\sigma_2$ with

$$\sigma_2(f) = f(f(f(x_1, x_2), f(x_2, x_1)), f(x_1, x_2))$$

satisfy $depth(\sigma_1) = depth(\sigma_2)$, but $\sigma_1$ and $\sigma_2$ are not $\mathcal{R}$-related. Therefore we need some more conditions.

For any $n$-ary term $t$ we denote by $S_n^{n,x}(t)$ the term arising from $t$ if we substitute for each variable the variable $x$.

Then we get:

**Proposition 3.1.** *Assume that $\sigma_1, \sigma_2 \in Hyp^f(n)$. If $\sigma_1 \mathcal{R} \sigma_2$, then*

$$S_n^{n,x}(\sigma_1(f)) = S_n^{n,x}(\sigma_2(f)).$$

*Proof.* If $\sigma_1 \mathcal{R} \sigma_2$ then there are hypersubstitutions $\sigma$ and $\sigma'$ such that $\sigma_1 = \sigma_2 \circ_h \sigma$ and $\sigma_2 = \sigma_1 \circ_h \sigma'$ and then $\sigma_1 = \sigma_1 \circ_h (\sigma \circ_h \sigma')$. By Proposition 2.5 we get $depth(\sigma_1) = depth(\sigma_1) depth(\sigma) depth(\sigma')$. From this we obtain $depth(\sigma) = depth(\sigma') = 1$. Since $\sigma$ and $\sigma'$ are full hypersubstitutions, we have $\sigma(f) = f(x_{s(1)}, \ldots, x_{s(n)})$, $\sigma'(f) = f(x_{s'(1)}, \ldots, x_{s'(n)})$ for permutations $s$,$s'$ on $\{1, \ldots, n\}$. From this there follows

$$\begin{aligned} S_n^{n,x}(\sigma_1(f)) &= S_n^{n,x}((\sigma_2 \circ_h \sigma)(f)) \\ &= S_n^{n,x}(\hat{\sigma_2}[\sigma(f)]) \\ &= S_n^{n,x}(\hat{\sigma_2}[f(x_{s(1)}, \ldots, x_{s(n)})]) \\ &= S_n^{n,x}(\hat{\sigma_2}[f(x_{s'(1)}, \ldots, x_{s'(n)})]) \\ &= S_n^{n,x}(\sigma_2(f)). \end{aligned}$$

□

Every full term $t \in W_n^f(X_n)$ contains subterms of the form

$$f(x_{s_j(1)}, \ldots, x_{s_j(n)})$$

for permutations $s_j$ on $\{1, \ldots, n\}$. Considering the tree of the term $t$ by $P(t)$ we denote the sequence of all permutations on $\{1, \ldots, n\}$ which are needed in $t$, written from the left to the right, i.e.,

$$P(t) := \{(s_1, \ldots, s_m) \mid f(x_{s_i(1)}, \ldots, x_{s_i(n)}) \text{ is a subterm of } t \text{ for } 1 \leq i \leq m \text{ and where these subterms are ordered in the tree of } t \text{ from the left to the right}\}.$$

**Example**: Consider for $\tau = (3)$ the term

$$t = f(f(x_2, x_1, x_3), f(f(x_1, x_2, x_3), f(x_2, x_1, x_3), f(x_3, x_2, x_1)), f(x_3, x_1, x_2)).$$

Then $P(t) = ((12), (1), (12), (13), (132))$ if we write the permutations which are needed as cycles. Clearly, two terms $t_1, t_2 \in W_{(n)}^f(X_n)$ are equal if and only if $S_n^{n,x}(t_1) = S_n^{n,x}(t_2)$ and $P(t_1) = P(t_2)$.

Now we have

**Proposition 3.2.** *Let $\sigma_1, \sigma_2 \in Hyp^f(n)$ and assume that $P(\sigma_1(f)) = (u_1, \ldots, u_m)$ and $P(\sigma_2(f)) = (v_1, \ldots, v_l)$. If $\sigma_1 \mathcal{R} \sigma_2$ then $m = l$ and $u_1 v_1^{-1} = \cdots = u_m v_m^{-1}$.*

*Proof.* Assume that $\sigma_1 \mathcal{R} \sigma_2$. There are hypersubstitutions $\sigma, \sigma'$ such that $\sigma_1 = \sigma_2 \circ_h \sigma$ and $\sigma_2 = \sigma_1 \circ_h \sigma'$ and therefore $depth(\sigma) = depth(\sigma') = 1$. It follows that there are permutations $s, s' : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that $\sigma_1(f) = \hat{\sigma_2}[f(x_{s(1)}, \ldots, x_{s(n)})]$ and $\sigma_2(f) = \hat{\sigma_1}[f(x_{s'(1)}, \ldots, x_{s'(n)})]$. Then

$$P(\sigma_1(f)) = P(\hat{\sigma_2}[f(x_{s(1)}, \ldots, x_{s(n)}]) = \\ = P(\sigma_2(f)(x_{s(1)}, \ldots, x_{s(n)})) = (s \circ v_1, \ldots, s \circ v_l).$$

Similarly, we get

$$P(\sigma_2(f)) = P(\hat{\sigma_1}[f(x_{s'(1)}, \ldots, x_{s'(n)}]) = \\ = P(\sigma_1(f)(x_{s'(1)}, \ldots, x_{s'(n)})) = (s' \circ u_1, \ldots, s' \circ u_m).$$

By Proposition 3.1 from $\sigma_1 \mathcal{R} \sigma_2$ there follows $S_n^{n,x}(\sigma_1(f)) = S_n^{n,x}(\sigma_2(f))$. Since the trees of $S_n^{n,x}(\sigma_1(f))$ and $\sigma_1(f)$ differ only in the labeling of the leaves, the structure of the tree of $\sigma_1(f)$ and of $\sigma_2(f)$ is equal and therefore the number of permutations $s$ occurring in $\sigma_1(f)$ and $\sigma_2(f)$ is equal, i.e. we have $m = l$ and then $(s \circ v_1, \ldots, s \circ v_m) = (s' \circ u_1, \ldots, s' \circ u_m)$ implies $s \circ v_j = s' \circ u_j$ for every $1 \leq j \leq m$. From this equation we obtain $u_j \circ v_j^{-1} = s \circ s'^{-1}$ for every $1 \leq j \leq m$ and this means $u_1 \circ v_1^{-1} = \cdots = u_m \circ v_m^{-1}$. □

It turns out that both conditions, Proposition 3.1 and Proposition 3.2 together, characterize Green's relation $\mathcal{R}$.

**Theorem 3.3.** *Let $\sigma_1, \sigma_2 \in Hyp^f(n)$. Then the following conditions are equivalent:*

*(i)* $\sigma_1 \mathcal{R} \sigma_2$,

*(ii)* $S_n^{n,x}(\sigma_1(f)) = S_n^{n,x}(\sigma_2(f))$ *and* $u_1 \circ v_1^{-1} = \cdots = u_m \circ v_m^{-1}$ *where* $P(\sigma_1(f)) = (u_1, \ldots, u_m)$ *and* $P(\sigma_2(f)) = (v_1, \ldots, v_m)$.

*Proof.* $(i) \Rightarrow (ii)$ was already proved.

$(ii) \Rightarrow (i)$: We form $s = u_1 \circ v_1^{-1}, s^{-1} = v_1 \circ u_1^{-1}$ and consider $\sigma, \sigma' \in Hyp^f(n)$ defined by $\sigma(f) = f(x_{s(1)}, \ldots, x_{s(n)})$ and $\sigma'(f) = f(x_{s^{-1}(1)}, \ldots, x_{s^{-1}(n)})$. Clearly, $\sigma' = \sigma^{-1}$. Now we prove that $\sigma_1 = \sigma_2 \circ_h \sigma$ using the remark before Proposition 3.2 and showing that $S_n^{n,x}(\sigma_1(f)) = S_n^{n,x}((\sigma_2 \circ \sigma)(f))$ and $P(\sigma_1(f)) = P((\sigma_2 \circ_h \sigma)(f))$. Indeed, we have

$$S_n^{n,x}((\sigma_2 \circ \sigma)(f)) = S_n^{n,x}(\hat{\sigma_2}[\sigma(f)]) = S_n^{n,x}(\hat{\sigma_2}[f(x_{s(1)}, \ldots, x_{s(n)})]) = \\ = S_n^{n,x}(\sigma_2(f)(x_{s(1)}, \ldots, x_{s(n)})) = S_n^{n,x}(\sigma_2(f)).$$

Further $P((\sigma_2 \circ_h \sigma)(f)) = P(\sigma_2(f)(x_{s(1)}, \ldots, x_{s(n)})) = (s \circ v_1, \ldots, s \circ v_m) = (u_1, \ldots, u_m) = P(\sigma_1(f))$. Since the inverse of $\sigma$ exists, we get $\sigma_2 = \sigma_1 \circ \sigma^{-1}$ and altogether we have $\sigma_1 R \sigma_2$. □

## References


[1] K. Denecke, J. Koppitz, R. Marszałek, Derived varieties and derived equational theories, International Journal of Algebra and Computation, Vol. **8**, No. 2 (1998), 153 -169.

[2] K. Denecke, S.L. Wismath, The monoid of hypersubstitutions of type $\tau = (2)$, Contributions to General Algebra, 10, Klagenfurth, (1998), 109-127.

[3] K. Denecke, S. L. Wismath, Hyperidentities and Clones, Gordon and Breach Science Publishers, 2000.

[4] K. Denecke, J. Koppitz, Sl. Shtrakov, The Depth of a hypersubstitution, Journal of Automata, Languages and Combinatorics, 6 (2001)3, 253-262.

[5] K. Denecke, Th. Changphas, Full hypersubstitutions and Fully Solid Varieties of Semigroups, preprint 2002.

[6] K. Denecke, S.L. Wismath, Valuations of Terms, preprint, 2002.

[7] J. M. Howie, Fundamentals of Semigroups, Academic Press,1995.

[8] G. Pilz, Nearrings, North-Holland Publ. Comp., 1977.

[9] H. J. Weinert, Semirings, Seminearfields and their Semigroup-theoretical Background, Semigroup Forum,Vol. 24 (1982),231-254.



Contact information

**Th. Changphas** Universität Potsdam, Institut für Mathematik, D-14415 Potsdam (Germany), PF 601553
*E-Mail:* thacha@kku.ac.th

**K. Denecke** Universität Potsdam, Institut für Mathematik, D-14415 Potsdam (Germany), PF 601553
*E-Mail:* kdenecke@rz.uni-potsdam.de


Received by the editors: 23.09.2002.

Algebra and Discrete Mathematics
Number 2. **(2003).** pp. 20–31 RESEARCH ARTICLE


# Multi-algebras from the viewpoint of algebraic logic

**Jānis Cīrulis**

Communicated by L. Márki

Abstract. Where $\boldsymbol{U}$ is a structure for a first-order language $\mathcal{L}^{\approx}$ with equality $\approx$, a standard construction associates with every formula $f$ of $\mathcal{L}^{\approx}$ the set $\|f\|$ of those assignments which fulfill $f$ in $\boldsymbol{U}$. These sets make up a (cylindric like) set algebra $Cs(\boldsymbol{U})$ that is a homomorphic image of the algebra of formulas. If $\mathcal{L}^{\approx}$ does not have predicate symbols distinct from $\approx$, i.e. $\boldsymbol{U}$ is an ordinary algebra, then $Cs(\boldsymbol{U})$ is generated by its elements $\|s \approx t\|$; thus, the function $(s,t) \mapsto \|s \approx t\|$ comprises all information on $Cs(\boldsymbol{U})$.

In the paper, we consider the analogues of such functions for multi-algebras. Instead of $\approx$, the relation $\varepsilon$ of singular inclusion is accepted as the basic one ($s\varepsilon t$ is read as '$s$ has a single value, which is also a value of $t$'). Then every multi-algebra $\boldsymbol{U}$ can be completely restored from the function $(s,t) \mapsto \|s \ \varepsilon \ t\|$. The class of such functions is given an axiomatic description.



## 1. Introduction

We begin, in the first subsection, with reviewing a few standard constructions used in algebraic logic. Then we outline the problem which we deal with in the paper.

**1.1** Let $\mathcal{L}^{\approx}$ be a first-order language with equality over the set of variables $X$. For the sake of definiteness, we assume that the logical primitives of $\mathcal{L}^{\approx}$ are $\neg, \wedge, \vee, \exists$. Let, furthermore, $\boldsymbol{U} := (U, \ldots)$ be a structure

*This research was supported by Latvian Science Council Grant No. 01.0254*
**2001 Mathematics Subject Classification:** *08A99; 03G15, 08A62.*
**Key words and phrases:** *cylindric algebra, linear term, multi-algebra, resolvent, singular inclusion.*

for $\mathcal{L}^{\approx}$. For every formula $f$ of $\mathcal{L}^{\approx}$, we denote by $\|f\|$ the set of those assignments from $U^X$ which satisfy $f$ in $\boldsymbol{U}$. Then

$$\|\neg f\| = -\|f\|,\ \|f \wedge g\| = \|f\| \cap \|g\|,\ \|f \vee g\| = \|f\| \cup \|g\|,$$

$$\|\exists x\, f\| = C_x\|f\|,\ \|x \approx y\| = D_{xy}.$$



Here $-$ is the set complementation, $C_x$ is the *cylindrification* along $x$-axis in the "space" $U^X$ and is defined by

$$C_x(A) := \{\varphi \in U^X\colon\ \varphi^x_u \in A \text{ for some } u \in U\} = \{\psi^x_u\colon\ \psi \in A, u \in U\}, \tag{1}$$

where $\varphi^x_u$ is the assignment that assigns $u$ to $x$ and $\varphi(y)$ to every other variable $y$, and the sets

$$D_{xy} := \{\varphi \in U^X\colon\ \varphi(x) = \varphi(y)\}$$

are known as *diagonal hyperplanes* in $U^X$. Put $\|F\| := \{\|f\|\colon\ f \in F\}$, where $F$ is the set of formulas of the language; the algebra

$$Cs(\boldsymbol{U}) := (\|F\|, \cup, \cap, -, C_x, D_{xy})_{x,y\in X}$$

is a version of cylindric set algebra [8, 9]. More precisely, according to Theorem 4.3.5 of [9], it is a regular and locally finite cylindric set algebra. We shall call it the *cylindric algebra of* $\boldsymbol{U}$. Two $\mathcal{L}^{\approx}$-structures have isomorphic cylindric algebras if and only if they are elementarily equivalent—this follows from Remark 4.3.68(7) in [9].

If the alphabet of $\mathcal{L}^{\approx}$ contains any operation symbols, then we may construct even a richer derived structure. Consider the term algebra $\boldsymbol{T} := (T, \ldots)$ and set

$$D_{st} := \{\varphi \in U^X\colon\ \tilde\varphi(s) = \tilde\varphi(t)\},$$

where $\tilde\varphi$ is the homomorphism $\boldsymbol{T} \to \boldsymbol{U}$ induced by $\varphi$. Now $\|s \approx t\| = D_{st}$. In terms of [2], the algebra

$$Cs_{\boldsymbol{T}}(\boldsymbol{U}) := (\|F\|, \cup, \cap, -, C_x, D_{st})_{x\in X,\ s,t\in T}$$

is a $\boldsymbol{T}$-*cylindric* set algebra, and the function $D\colon\ T\times T \to \mathcal{P}(U^X)$ defined by $D(s,t) := D_{st}$ is a $\boldsymbol{T}$-*diagonal* on it.

**1.2** In the case when $\approx$ is the single predicate symbol in $\mathcal{L}^{\approx}$ and, correspondingly, $\boldsymbol{U}$ is merely an algebra, $Cs_{\boldsymbol{T}}(\boldsymbol{U})$ is generated by the "$\boldsymbol{T}$-diagonal planes" $D_{st}$. Hence, the $\boldsymbol{T}$-diagonal $D$ carries then all information on $\boldsymbol{U}$ available in $Cs_{\boldsymbol{T}}(\boldsymbol{U})$, and we may concentrate on $\boldsymbol{T}$-diagonals rather than deal with whole $\boldsymbol{T}$-cylindric algebras. Actually, even more general situation was studied in [3], where $\boldsymbol{T}$ was an algebra free in some variety $\mathcal{K}$. It was shown there that every $\mathcal{K}$-algebra can be restored from its $\boldsymbol{T}$-diagonal and that homomorphisms between $\mathcal{K}$-algebras can also be characterized in terms of $\boldsymbol{T}$-diagonals. Moreover, the class of those functions $T^2 \to \mathcal{P}(U^X)$ that are $\boldsymbol{T}$-diagonals of algebras from $\mathcal{K}$ was given an axiomatic description. Axioms of $\boldsymbol{T}$-diagonals were used in [2] to introduce the concept of an abstract cylindric algebras with terms. For another approach to such algebras, involving substitutions along with diagonals, see [5].

Consequently, from the point of view of algebraic logic, algebras from $\mathcal{K}$ are well-presented by their $\boldsymbol{T}$-diagonals. Some relevant information on an algebra $\boldsymbol{U}$ may be read directly from $D$. For example, $D_{st}$ may be considered as the set of solutions of the equation $s \approx t$ in $\boldsymbol{U}$, and the algebra satisfies this equation iff $D_{st} = U^X$. Given a relation $\theta \subset T \times T$, let $D_\theta$ be the intersection $\bigcap(D_{st}\colon\ (s,t) \in \theta)$. In the sense of universal algebraic geometry as it is developed in [12, 13], $D_\theta$ is essentially the algebraic variety in the space $U^X$ described by the set of $\boldsymbol{T}$-equations $\theta$.

**1.3** Our aim in this paper is to extend the approach of [3] to multi-algebras. A minor trouble is that, for multi-algebras, there are several possible ways how to interpret the equality symbol $\approx$. Probably, the most popular one is the reading of the equation $s \approx t$ as '$s$ and $t$ have the same (sets of) values'. Such equations are discussed, for example, in [17]; seemingly, this interpretation of $\approx$ is suggested by tradition of complex, or powerset, algebras—see [7, 6]. On the other hand, the weak commutativity or weak distributivity laws for certain ring-like multi-algebras (see, e.g., [16]) can be written as equations, where $\approx$ expresses overlapping of values sets of both terms; then $s \approx t$ means '$s$ and $t$ have a common value'. A possible substituend for equality and overlapping is inclusion. In ordinary algebras all of these concepts reduce to identity of elements of the base set.

Following [14], instead of any of the above relations, we choose the relation of singular inclusion $\varepsilon$ to be the basic one: the atomic formula $s \;\varepsilon\; t$ is informally read as 'the term $s$ has a single value, and it is also a value of $t$'. For partial algebras, the formula reduces to the so called existential equation $s \stackrel{\mathrm{e}}{=} t$ (see, e.g., [1]), while for ordinary algebras $\varepsilon$ has the same meaning as $\approx$. Note that the identity relation on the base set is presented by formulas of type $s \;\varepsilon\; t \wedge t \;\varepsilon\; s$, and that overlapping, inclusion and equality relations for values sets of $s$ and $t$ are definable by formulas $\exists x(x \;\varepsilon\; s \wedge x \;\varepsilon\; t)$, $\forall x(x \;\varepsilon\; s \to x \;\varepsilon\; t)$ and $\forall x(x \;\varepsilon\; s \leftrightarrow x \;\varepsilon\; t)$, respectively (where $x$ is free neither in $s$ nor $t$). At last, $t \varepsilon t$ means that the term $t$ is single-valued.

Since singular inclusion models some appropriate aspects of the set-theoretical 'element_of' relation, we consider singular inclusion as the most natural primitive for the language of multi-algebras. Inclusion has also been preferred to equality in some papers on logic of multi-algebras; see, e.g., [11, 10], where equality was shown to be a concept too weak for certain purposes. In fact, aside from inclusion, neither overlapping nor singular inclusion can be expressed in terms of equality.

## 2. Multi-algebras, valuations and resolvents

In this section we recall the notion of a multi-algebra and introduce the notion of an $\varepsilon$-resolvent of a multi-algebra, which is the $\varepsilon$-analogue of its $\boldsymbol{T}$-diagonal (the latter could also be termed its $\approx$-resolvent). Let $\Omega$ be some signature, and let now $\boldsymbol{T}$ be an $\Omega$-algebra relatively free on an infinite set of variables $X$. We consider elements of $T$ as "squeezed" terms.

**2.1** Let us first recall some constructions and facts from [15] concerning algebras of squeezed terms. Given $Y \subset X$, we say that $Y$ *supports* the term $t$ if $t$ belongs to the subalgebra of $\boldsymbol{T}$ generated by $Y$, and that $t$ is *independent* of a variable $x$ if $t$ is supported by some $Y$ not containing $x$. According to [15, Theorem 2.1], $Y$ supports $t$ iff $\sigma(t) = t$ for every endomorphism $\sigma$ of $\boldsymbol{T}$ that coincides with the identity map on $Y$.

The set $\triangle t := \bigcap(Y\colon\ Y \text{ supports } t)$ of all those variables $t$ depends on is always finite and supports $t$. If $\boldsymbol{T}$ is the absolutely free word algebra (as in Sect. 1), then $\triangle t$ consists just of the variables occurring in $t$. In any case,

$$\triangle \omega t_1 t_2 \dots t_m \subset \triangle t_1 \cup \triangle t_2 \cup \dots \cup \triangle t_m \tag{2}$$

and, if $[s/x]$ stands for the endomorphism of $\boldsymbol{T}$ that takes $x$ into $s$ and coincides with the identity map on $X \setminus \{x\}$, then

$$\triangle [s/x] t \subset \triangle s \cup (\triangle t \setminus \{x\}). \tag{3}$$

Note that $t$ depends on $x$ iff $x \in \triangle t$, and that $[s/x]t = t$ iff $t$ is independent of $x$.

We further isolate, for each variable $x$, the subset $L_x$ of terms *linear in* $x$. It is defined to be the smallest set containing $x$ as well as all terms $\omega t_1 t_2 \dots t_m$ with $t_i \in L_x$ for some $i$ and $x \notin \triangle t_j$ for $j \neq i$. An ordinary term is linear in $x$ if and only if $x$ occurs in it just once; this is the meaning in which the attribute 'linear' has been used, say, in [6].

**2.2** An $m$-ary *multi-operation* on $U$ is any function $o$ of type $U^m \to \mathcal{P}(U)$. We shall identify singletons from $\mathcal{P}(U)$ with respective elements of $U$; therefore, any operation on $U$ may be treated as a multi-operation. The *extension* of $o$ is the operation $\bar{o}$ on $\mathcal{P}(U)$ defined by

$$\bar{o}(A_1, A_2, \ldots, A_m) := \bigcup(o(u_1, u_2, \ldots, u_m)\colon \\ u_1 \in A_1, u_2 \in A_2, \ldots, u_m \in A_m).$$

*Definition* 1. A *multi-algebra* is a system $\boldsymbol{U} := (U, \omega_{\boldsymbol{U}})_{\omega\in\Omega}$, where each $\omega_{\boldsymbol{U}}$ is a multi-operation on $U$ whose arity is determined by $\omega$. A mapping $\mu\colon T \to \mathcal{P}(U)$ is said to be a *valuation* in $\boldsymbol{U}$ if

$$\mu(x) \in U, \quad \mu(\omega t_1 t_2 \ldots t_m) = \bar{\omega}_{\boldsymbol{U}}(\mu(t_1), \mu(t_2), \ldots, \mu(t_m)).$$

for $x \in X$, $\omega \in \Omega$ and $t_1, t_2, \ldots, t_m \in T$.

Thus every valuation in $\boldsymbol{U}$ is an extension of some assignment from $U^X$, and may be regarded as a kind of multihomomorphism from $\boldsymbol{T}$ to $\boldsymbol{U}$. In particular, valuations in an ordinary algebra $\boldsymbol{U}$ are just homomorphisms from $\boldsymbol{T}$ to $\boldsymbol{U}$. Let $Val(\boldsymbol{U})$ stand for the set of all valuations in $\boldsymbol{U}$. Note that $Val(\boldsymbol{T}) = End(\boldsymbol{T})$.

A multi-algebra $\boldsymbol{U}$ is said to be $\boldsymbol{T}$*-shaped* if $Val(\boldsymbol{U})$ is maximally rich, i.e. if every assignment $\varphi$ can be extended to a valuation $\tilde{\varphi}$ (necessarily unique) in $U$. Then elements of $\tilde{\varphi}(t)$ are thought of as *values* of the term $t$ on $\varphi$. According to our convention on singletons, a term $t$ has a single value on $\varphi$ iff $\tilde{\varphi}(t) \in U$. We denote by $\mathcal{V}(\boldsymbol{T})$ the class of all $\boldsymbol{T}$-shaped multi-algebras. Clearly, $\mathcal{V}(\boldsymbol{T})$ includes the variety of ordinary algebras generated by $\boldsymbol{T}$, and contains all multi-algebras when $\boldsymbol{T}$ is absolutely free. Furthermore, for $\boldsymbol{U} \in \mathcal{V}(\boldsymbol{T})$,

$$\varphi|\triangle t = \psi|\triangle t \Rightarrow \tilde{\varphi}(t) = \tilde{\psi}(t) \tag{4}$$

and, if $t$ is linear in $x$,

$$\tilde{\varphi}([s/x]t) = \{v\colon \exists u(v \in \widetilde{\varphi^x_u}(t) \text{ and } u \in \tilde{\varphi}(s))\}. \tag{5}$$

The routine proof of (5) is by induction on $L_x$, using (2) and (3).

It is easily seen that every $\boldsymbol{T}$-shaped multi-algebra is completely determined by its valuations. Indeed, assume that $\boldsymbol{U}$ and $\boldsymbol{U}'$ are two different multi-algebras with a common carrier $U$. Then there is an operation symbol $\omega \in \Omega$ such that $\omega_{\boldsymbol{U}}(u_1, u_2, \ldots, u_m) \neq \omega_{\boldsymbol{U}'}(u_1, u_2, \ldots, u_m)$ for some $u_1, u_2, \ldots, u_m \in U$. For sake of definiteness, suppose that $u \in \omega_{\boldsymbol{U}}(u_1, u_2, \ldots, u_m)$ and $u \notin \omega_{\boldsymbol{U}'}(u_1, u_2, \ldots, u_m)$. Furthermore, choose distinct variables $x_1, x_2, \ldots, x_m$ and a valuation $\mu$ such that $\mu(x_i) = u_i$ for all $i$. Now, if $t$ is the term $\omega x_1 x_2 \cdots x_m$, then $u$ is a value of $t$ on $\mu$ in $\boldsymbol{U}$, but not in $\boldsymbol{U}'$. So, the sets of valuations are also distinct.

In what follows, we shall consider only $\boldsymbol{T}$-shaped multi-algebras.

**2.3** Let us introduce the notion of a resolvent—the multi-algebra equivalent of a $\boldsymbol{T}$-diagonal of an ordinary algebra (see Introduction). Recall that the formula $s \mathrel{\varepsilon} t$ can also be considered as a kind of equation, and then the resolvent provides us with solutions of these "$\varepsilon$-equations"; this motivates the suggested term.

*Definition* 2. The *$\varepsilon$-resolvent*, or just *resolvent of a multi-algebra* $\boldsymbol{U}$ is the function $Res(\boldsymbol{U})\colon\ T \times T \to \mathcal{P}(U^X)$ defined as follows:

$$Res(\boldsymbol{U})(s,t) := \{\varphi \in U^X\colon\ \tilde\varphi(s) \in \tilde\varphi(t)\}. \tag{6}$$

Therefore, $\|s \mathrel{\varepsilon} t\| = Res(\boldsymbol{U})(s,t)$. Note that the set algebra

$$Cs_{\boldsymbol{T}}(\boldsymbol{U}) := (\|F\|, \cup, \cap, -, C_x, R_{st})_{x\in X,\ s,t\in T},$$

where $R_{st}$ stands for $Res(\boldsymbol{U})(s,t)$, is an ordinary algebra generated by these elements.

A multi-algebra is completely determined even by a "half" of its resolvent, the first argument being a variable which the second one does not depend on. Namely, we can restore the operation $\omega_{\boldsymbol{U}}$ of $\boldsymbol{U}$ corresponding to an operation symbol $\omega \in \Omega$ as follows:

$$v \in \omega_{\boldsymbol{U}}(u_1, u_2, \ldots, u_m) \Leftrightarrow \varphi \in R_{yt},$$

where $t$ is $\omega x_1 x_2, \ldots, x_m$ and $y \notin \triangle t$ for distinct variables $x_1, x_2, \ldots, x_m, y$, while $\varphi$ is selected so that $\varphi(y) = v$ and $\varphi(x_i) = u_i$.

Thus, different algebras from $\mathcal{V}(\boldsymbol{T})$ have different resolvents.

By a *support* of a set $A \subset U^X$ we shall mean any subset $Y \subset X$ such that, for all $\varphi, \psi \in U^X$,

$$\varphi \in A,\ \varphi|Y = \psi|Y \ \Rightarrow\ \psi \in A.$$

This concept comes from the theory of polyadic algebras. By analogy with standard cylindric algebras (see [8, 9]), the set algebra $Cs_{\boldsymbol{T}}$ could be called *regular* if every its element $A$ is regular in the sense that the subset $\{x \in X\colon C_x(A) \neq A\}$ is a support of $A$. However, apart from the note just after Theorem 2 below, we shall not concern with regularity property in this paper.

**Theorem 1.** *If a function $R\colon\ T \times T \to \mathcal{P}(U^X)$ is a resolvent of a $\boldsymbol{T}$-shaped multi-algebra, then it satisfies the conditions*

(R0): $R(x,y) = D_{xy}$,

(R1a): $R(r,s) \cap R(s,t) \subset R(s,r)$,

(R1b): $R(r,s) \cap R(s,t) \subset R(r,t)$,

(R2): $R(s,[r/x]t) = C_x(R(x,r) \cap R(s,t))$ *if* $t \in L_x$ *and* $x \notin \triangle r \cup \triangle s$,

(R3): *every* $R(s,t)$ *has a finite support.*

*Proof.* (R0) and (R1b) are obvious, while (R1a) is true because the left hand side assures that the value set of $s$ is a singleton. We shall check only (R2) and (R3) here. By (6), (5), (4), again (6), and (1),

$$\begin{aligned}\varphi \in R(s,[r/x]t) &\Leftrightarrow \tilde{\varphi}(s) \in \tilde{\varphi}([r/x]t)\\ &\Leftrightarrow \exists u(\tilde{\varphi}(s) \in (\widetilde{\varphi^x_u})(t) \text{ and } u \in \tilde{\varphi}(r))\\ &\Leftrightarrow \exists u(\varphi^x_u(s) \in (\widetilde{\varphi^x_u})(t) \text{ and } u \in \tilde{\varphi}(r))\\ &\Leftrightarrow \exists u(\varphi^x_u \in R(s,t) \text{ and } \varphi^x_u \in R(x,r))\\ &\Leftrightarrow \varphi \in C_x(R(x,r) \cap R(s,t)),\end{aligned}$$

i.e. (R2) holds. By (2) and (4), the finite set $\triangle s \cup \triangle t$ is a support of $R(s,t)$, and (R3) also holds. □

Note that these conditions are, in fact, properties of singular inclusion written algebraically. Thus, (R1b) fixes transitivity of $\varepsilon$, while (R2) says that $s\varepsilon[r/x]t$ holds iff $x\varepsilon r$ and $s\varepsilon t$ hold for some value of $x$. We shall need only the following two particular cases of (R2):

$$R(s,r) = C_x(R(s,x) \cap R(x,r)) \tag{7}$$

with $x \notin \triangle s \cup \triangle t$, and

$$R(y,[r/x]t) = C_x(R(x,r) \cap R(y,t)) \tag{8}$$

with $t \in L_x$ and $x \neq y \notin \triangle s$, $y \notin \triangle t$. (In fact, (R2) is a consequence of them.)

*Definition* 3. A ***T**-resolvent on a set $U$* is any function $R\colon\ T \times T \to \mathcal{P}(U^X)$ satisfying the conditions (R0)–(R2). The resolvent is said to be *finitary* iff it satisfies also (R3).

According to the preceding theorem, the resolvent of any multi-algebra is a finitary resolvent in this abstract sense on its base set. The following representation theorem, which is the main result of the paper, states the converse.

**Theorem 2.** *Every finitary **T**-resolvent is a resolvent of some multi-algebra from $\mathcal{V}(\boldsymbol{T})$.*

This theorem is a close analogue of Theorem 3 in [3] and Theorem 4.3 in [2] on superdiagonals of **T**-cylindric algebras, with the exception that in the latter one the superdiagonal was required to be regular rather than just finitary. This difference is not essential: as all sets $\triangle t$ are finite, both conditions turn out to be equivalent in our context. The theorem will be proved in the next section.

We already observed just after Definition 2 that different algebras with the same base set still have different resolvents. So we come to a corollary which shows that, for algebraic logic, every multi-algebra $\boldsymbol{U}$ is adequately presented by some resolvent, and conversely.

**Theorem 3.** *The transformation Res: $\boldsymbol{U} \mapsto Res(\boldsymbol{U})$ provides a one-to-one correspondence between $\boldsymbol{T}$-shaped multi-algebras with the base set $U$ and finitary $\boldsymbol{T}$-resolvents on $U$.*

We remind that the set algebra $Cs_{\boldsymbol{T}}(\boldsymbol{U})$, being generated by the resolvent of $\boldsymbol{U}$, is completely determined by it. Hence, Theorem 2 could serve as a basis for a representation of an appropriate class of "$\varepsilon$-cylindric" algebras (cf. a similar situation with $\boldsymbol{T}$-diagonals and $\boldsymbol{T}$-cylindric algebras in Sect. 4 of [2]) and, further, for an algebraic proof of completeness of a logic with multivalued terms (see [14] for such a logic).

## 3. Proof of Theorem 2

The proof consists of a sequence of technical lemmas.

**3.1** First we derive some additional properties of $\boldsymbol{T}$-resolvents.

**Lemma 4.** *Suppose that $R$ is a $\boldsymbol{T}$-resolvent on $U$. If a term $t$ does not depend on the distinct variables $y$ and $z$, then, for all assignments $\varphi$ and elements $u \in U$*
(a) $\varphi \in R(y,t)$ *if and only if* $\varphi^z_u \in R(y,t)$,
(b) $\varphi^y_u \in R(y,t)$ *if and only if* $\varphi^z_u \in R(z,t)$.
*If, furthermore, assignments $\varphi$ and $\psi$ agree on $\triangle t$, and $R(y,t)$ has a finite support, then*
(c) $\varphi^y_u \in R(y,t)$ *if and only if* $\psi^y_u \in R(y,t)$
*for all $u \in U$.*

*Proof.* Assume that $t, y$ and $z$ are as indicated. We first note that, by (7),

$$C_z(R(y,t)) = C_z(C_z(R(y,z) \cap R(z,t))) = C_z(R(y,z) \cap R(z,t)) = R(y,t). \tag{9}$$

Now, if $\varphi \in R(y,t)$, then $\varphi^z_u \in C_zR(y,t) = R(y,t)$, but if $\varphi^z_u \in R(y,t)$, then $\varphi \in C_zR(z,t) = R(y,t)$. Therefore, (a) holds.

Once again referring to (7), and using (1), (R0), (a), we arrive at (b):

$$\begin{aligned}\varphi_u^y \in R(y,t) &\Leftrightarrow \varphi_u^y \in C_z(R(y,z)\cap R(z,t))\\ &\Leftrightarrow \exists v(\varphi_{uv}^{yz} \in R(y,z) \text{ and } \varphi_{uv}^{yz} \in R(z,t))\\ &\Leftrightarrow \exists v(u=v \text{ and } \varphi_v^z \in C_y(R(z,t)))\\ &\Leftrightarrow \varphi_u^z \in C_y(R(z,t)) = R(z,t).\end{aligned}$$

To prove (c), assume that $\varphi|\triangle t = \psi|\triangle t$. Then also $\varphi_u^y|\{y\}\cup\triangle t = \psi_u^y|\{y\}\cup\triangle t$ for any $u\in U$. If $Y$ is a finite support of $R(y,t)$, then we do not loss generality assuming that $\varphi$ and $\psi$ agree everywhere outside $Y$. Hence, $\varphi_u^y$ and $\psi_u^y$ may differ only on the set $\{x_1,x_2,\dots,x_n\} := Y-(\triangle t\cup\{y\})$; we are only interested in the case $n>0$. Now let $v_i := \psi(x_i)$ for all $i$; then

$$\varphi_u^y \in R(y,t) \;\Leftrightarrow\; \varphi_{uv_1v_2\cdots v_n}^{yx_1x_2\cdots x_n} \in R(y,t) \;\Leftrightarrow\; \psi_u^y \in R(y,t)$$

by multiple use of (a). □

**Corollary 5.** *Let $R$ be a $\boldsymbol{T}$-resolvent on $U$, and let $\varphi^*\colon\, T\to\mathcal{P}(U)$ be the extension of an assignment $\varphi$ in $U$ defined by the condition*

$$\varphi^*(t) := \{u\in U\colon\; \varphi_u^y \in R(y,t)\}, \tag{10}$$

*where $y\notin\triangle t$. Then $\varphi^*$ does not depend on the choice of $y$, and, if $z\notin\triangle t$,*

$$R(z,t) = \{\varphi\in U^X\colon\; \varphi(z)\in\varphi^*(t)\}. \tag{11}$$

*Moreover, if $R$ is finitary, then*

$$\varphi|\triangle t = \psi|\triangle t \Rightarrow \varphi^*(t) = \psi^*(t). \tag{12}$$

*Proof.* By (R0), $\varphi^*(x) = \varphi x$; so the function $\varphi^*$ is indeed an extension of $\varphi$. The fact that $\varphi^*$ does not depend on the choice of $y$ immediately follows from Lemma 4(b), and (12) is then another form of Lemma 4(c). By (10) and Lemma 4(b),

$$\varphi(z)\in\varphi^*(t) \Leftrightarrow \varphi_{\varphi(z)}^y \in R(y,t) \Leftrightarrow \varphi_{\varphi(z)}^z \in R(z,t) \Leftrightarrow \varphi\in R(z,t);$$

so (11) also holds. □

**Lemma 6.** *If $R$ is a finitary $\boldsymbol{T}$-resolvent on $U$, then*

$$R(s,t) = \{\varphi\in U^X\colon\; \varphi^*(s)\in\varphi^*(t)\}. \tag{13}$$

*Proof.* We first prove that

$$z \notin \triangle s,\ \psi \in R(s,z) \Rightarrow \psi^*(s) = \psi(z). \tag{14}$$

Suppose that $z \notin \triangle s$. If $\psi \in R(s,z)$, then $\psi \in R(z,s)$ by (R0) and (R1a), for (11) implies that $\psi \in R(z,s)$. Consequently, $\psi(z) \in \psi^*(s)$ by (11). Let, furthermore, $u$ be any element from $\psi^*(s)$. Choose one more variable $y \notin \triangle s$; in view of (12), we may assume that $\psi(y) = u$. Then $\psi \in R(y,s)$ according to (11); so, by (R1b), $\psi \in R(y,z)$, wherefrom $u = \psi(y) = \psi(z)$—see (R0). So, $\psi^*(s)$ is a singletone and must coincide with $\psi(z)$. Now (13) follows by (7) and (1), (14) and (10), and (12):

$$\begin{aligned} \varphi \in R(s,t) &\Leftrightarrow \exists u(\varphi^z_u \in R(s,z) \text{ and } \varphi^z_u \in R(z,t)) \\ &\Leftrightarrow \exists u((\varphi^z_u)^*(s) = u \text{ and } u \in \varphi^*(t)) \\ &\Leftrightarrow \exists u(\varphi^*(s) = u \text{ and } u \in \varphi^*(t)) \\ &\Leftrightarrow \varphi^*(s) \in \varphi^*(t), \end{aligned}$$

as needed. □

In view of this lemma, it remains to show that there is a $\boldsymbol{T}$-shaped multi-algebra such that the set of all extensions $\varphi^*$ turns out to be its set of valuations. This will be done in the next subsection. We need one more simple lemma.

**Lemma 7.** *Suppose that $t$ is linear in $x$ and that $s$ does not depend on $x$. Then*

$$\varphi^*([s/x]t) = \bigcup((\varphi^x_v)^*(t)\colon\ v \in \varphi^*(s)). \tag{15}$$

*Proof.* By (10), (8) and (1), Lemma 4(a), and (11),

$$\begin{aligned} u \in \varphi^*([s/x]t) &\Leftrightarrow \varphi^y_u \in R(y,[s/x]t) \\ &\Leftrightarrow \exists v(\varphi^{yx}_{uv} \in R(x,s) \text{ and } \varphi^{yx}_{uv} \in R(y,t)) \\ &\Leftrightarrow \exists v(\varphi^x_v \in R(x,s) \text{ and } \varphi^{yx}_{uv} \in R(y,t)) \\ &\Leftrightarrow \exists v(v \in \varphi^*(s) \text{ and } u \in (\varphi^x_v)^*(t)) \\ &\Leftrightarrow u \in \bigcup((\varphi^x_v)^*(t)\colon\ v \in \varphi^*(s)), \end{aligned}$$

where $y$ is appropriately chosen. □

Using the lemma repeatedly, we now obtain the following equality for every assignment $\varphi$, every term $t := \omega t_1 t_2 \cdots t_m$ and mutually distinct variables $x_1, x_2, \ldots, x_m$:

$$\varphi^*(t) = \bigcup(\psi^*(\omega x_1 x_2 \cdots x_m)\colon\ \psi \in U^X,\ \psi(x_i) \in \varphi^*(t_i) \quad (i = 1,2,\ldots,m)). \tag{16}$$

**3.2** We now claim that, for any $m$-ary $\omega \in \Omega$, the operation $\omega^R$ on $U$ defined by

$$\omega^R(u_1, u_2, \ldots, u_m) := \varphi^*(t),$$

where $t := \omega x_1 x_2 \ldots x_m$ (for distinct variables $x_i$) and $\varphi$ is an assignment in $U$ such that $u_i = \varphi(x_i)$ for all $i$, does not depend on the choice of $x_1, x_2, \ldots, x_m$ and $\varphi$. Indeed, suppose that $t' = \omega y_1 y_2 \ldots y_m$ and that $\psi$ is an assignment such that $\psi(y_i) = u_i$ for all $i$. If $\sigma$ is any endomorphism of $\boldsymbol{T}$ that takes every $x_i$ into $y_i$, then $\psi^*\sigma$ is an assignment that coincides with $\varphi$ on $\{x_1, x_2, \ldots, x_m\}$. Since the later set supports $t$ (see (2)), we may apply (12):

$$\psi^*(t') = \psi^*(\omega(\sigma x_1)(\sigma x_2) \cdots (\sigma x_m)) = \psi^*(\sigma(t)) = \varphi^*(t).$$

Note that the definition of $\omega^R$ may be rewritten in the form

$$\omega^R(\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_m)) = \varphi^*(t), \tag{17}$$

where now $\varphi$ is arbitrary.

This way the set $U$ can be turned into $\Omega$-multi-algebra $(U, \omega^R)_{\omega\in\Omega}$, which we denote by $Alg(R)$. Our next claim is that every $\varphi^*$ is the valuation in $Alg(R)$ induced by the assignment $\varphi$, i.e. that $\varphi^*$ coincides with $\tilde{\varphi}$.

Given a term $t := \omega t_1 t_2 \ldots t_m$, select mutually distinct variables $x_1, x_2, \ldots, x_m$ outside $\triangle t$. Then, by (16) and (17) (with $\psi$ in the role of $\varphi$) and the definition of an extended operation (viz., $\bar{\omega}^R$),

$$\begin{aligned} \varphi^*(t) &= \bigcup((\omega x_1 x_m \cdots x_m)\colon \ \psi \in U^X,\ \psi(x_i) \in \varphi^*(t_i)) \\ &= \bigcup(\omega^R(\psi(x_1), \psi(x_2), \ldots, \psi(x_m))\colon \ \psi \in U^X,\ \psi(x_i) \in \varphi^*(t_i)) \\ &= \bar{\omega}^R(\mu(t_1), \mu(t_2), \ldots \mu(t_m)), \end{aligned}$$

as needed.

It now follows that $Alg(R) \in \mathcal{V}(\boldsymbol{T})$. Thus, the proof of Theorem 2 is completed. Note that the transformation $Alg\colon\ R \mapsto Alg(R)$ is converse to $Res$ mentioned in Theorem 3.

## Acknowledgment

The author is indebted to the anonymous referee for comments which helped to improve on the presentation.

## References

[1] Burmeister, P.: *A Model Theoretic Oriented Approach to Partial Algebras.* Akademie-Verlag, Berlin, 1986.

[2] Cīrulis, J.: *An algebraization of first order logic with terms.* Colloq. Math. Soc. J. Bolyai **54**, Algebraic logic, 1991, 125–146.

[3] Cīrulis, J.: *Superdiagonals of universal algebras*, Acta Univ. Latviensis **576** (1992), 29–36,

[4] Cīrulis, J.: *Corrections to my paper "An algebraization of first-order logic with terms"*, Acta Univ. Latviensis **595** (1994), 49–51.

[5] Feldman, N.: *Cylindric algebras with terms.* J. Symb. Log., **55** (1990), 854–866.

[6] Grätzer, G., Lakser, H.: *Identities for globals (complex algebras) of algebras.* Colloq. Math. **56** (1988), 19–29.

[7] Guatam, N.D.: *The validity of equations of a complex algebra.* Arch. Math. Logik Grundl. **3** (1957), 117–124.

[8] Henkin, L., Monk, J.D. e.a.: *Cylindric Set Algebras* Lect. Notes Math. **883**, Springer, 1981.

[9] Henkin, L., Monk, J.D., Tarski, A.: *Cylindric Algebras, Part II*, North-Holland, 1985.

[10] Pinus, A.G.: *On elementary theories of generic multi-algebras* (Russian). Izv. Vuzov. Matematika, 1999:4, 39–43. English translation: Russ. Math. **43** (1999), No. 4, 32–41.

[11] Pinus, A., Madarász, R.Sz.: *On generic multi-algebras* Novi Sad J. Math. **27** (1997), 77–82.

[12] Plotkin, B.: *Varieties of algebras and algebraic varieties.* Israel Math. J. **96** (1996), 511-522.

[13] Plotkin, B.: *Seven lectures on the universal algebraic geometry.* Preprint of Institute of Mathematics, Hebrew University, Jerusalem, 2000/2001, 135 pp.

[14] Tsirulis, Ya.: *A logical system admitting vacuous and plural terms* (Russian). Zeitshr. math. Logik Grundl. Math. **31** (1985), 263–274.

[15] Tsirulis, Ya.P.: *Two generalizations of the notion of a polyadic algebra* (Russian). Izv. Vuzov. Matematika, 1988:12, 39–50. English translation: Sov. Math. **32** (1988), No. 12, 61–78.

[16] Vougiouklis, T.: *Hyperstructures and their representations.* Hadronic Press, Palm Harbor, 1994.

[17] Vas, L., Madarász, R.Sz.: *A note about multi-algebras, power algebras and identities.* Proc. IX conf. on applied math. (Budva, May/June 1994), Univ. Novi-Sad, 1995, 147–153.

## Contact information

**J. Cīrulis**

Department of Computer Science, University of Latvia, Raiņa b., 19, LV–1586 Riga, Latvia
*E-Mail:* `jc@lanet.lv`

Received by the editors: 09.10.2002.

Algebra and Discrete Mathematics RESEARCH ARTICLE
Number 2. **(2003).** pp. 32–35


# A note on maximal ideals in ordered semigroups

**N. Kehayopulu, J. Ponizovskii, M. Tsingelis**

Abstract. In commutative rings having an identity element, every maximal ideal is a prime ideal, but the converse statement does not hold, in general. According to the present note, similar results for ordered semigroups and semigroups -without order- also hold. In fact, we prove that in commutative ordered semigroups with identity each maximal ideal is a prime ideal, the converse statement does not hold, in general.

There is an important class of ideals of rings which are prime, namely, the maximal ideals. In fact, in a commutative ring with identity every maximal ideal is a prime ideal. On the other hand, there are rings possesing a nontrivial prime ideal which is not maximal (cf. e.g. [1]). Similar results for ordered semigroups, also for semigroups -without order- also hold.

If $(S, ., \leq)$ is an ordered semigroup, a non-empty subset $I$ of $S$ is called a left (resp. right) ideal of $S$ if 1) $SI \subseteq I$ (resp. $IS \subseteq I$) and 2) $a \in I$, $S \ni b \leq a$ implies $b \in I$ [2]. If $(S, .)$ is a semigroup, a left (resp. right) ideal of $S$ is a non-empty subset $I$ of $S$ such that $SI \subseteq I$ (resp. $IS \subseteq I$). If $S$ is a semigroup or an ordered semigroup and $I$ both a left and a right ideal of $S$, then it is called an ideal of $S$. An ideal $I$ of a semigroup (resp. ordered semigroup) $S$ is called prime if $a, b \in S$ such that $ab \in I$ implies $a \in I$ or $b \in I$. Equivalent Definition: $A, B \subseteq S$ such that $AB \subseteq I$ implies $A \subseteq I$ or $B \subseteq I$ [2]. An ideal $M$ of a semigroup or an ordered semigroup $S$ is called proper if $M \neq S$ [3]. A proper ideal $M$ of a semigroup or an ordered semigroup $S$ is called maximal if there exists no ideal $T$ of $S$ such that $M \subset T \subset S$, equivalently, if for each ideal $T$ of $S$ such that $M \subseteq T$, we have $T = M$ or $T = S$ (cf. also [2]). If $S$ is an ordered semigroup and $H \subseteq S$, we denote

$$(H] := \{t \in S \mid t \leq h \text{ for some } h \in H\}.$$

**2000 Mathematics Subject Classification:** *06F05.*
**Key words and phrases:** *maximal ideal, prime ideal in ordered semigroups.*

If $S$ is an ordered semigroup (resp. semigroup) and $\emptyset \neq A \subseteq S$, we denote by $\mathcal{I}(A)$ the ideal of $S$ generated by $A$ i.e. the smallest -under inclusion relation- ideal of $S$ containing $A$. For an ordered semigroup $S$, we have $\mathcal{I}(A) = (A \cup SA \cup AS \cup SAS]$ (cf. [2]). For a semigroup $S$, we have $\mathcal{I}(A) = A \cup SA \cup AS \cup SAS$.

Let $\{(S_i, \circ_i, \leq_i) \mid i \in I\}$ be a non-empty family of ordered semigroups. The cartesian product $\prod\limits_{i\in I} S_i$ with the multiplication "$*$" and the order "$\preceq$" on $\prod\limits_{i\in I} S_i$ defined by

$$* : \prod_{i\in I} S_i \times \prod_{i\in I} S_i \to \prod_{i\in I} S_i \mid ((x_i)_{i\in I}, (y_i)_{i\in I}) \to (x_i)_{i\in I} * (y_i)_{i\in I} \quad \text{where}$$

$$(x_i)_{i\in I} * (y_i)_{i\in I} := (x_i \circ_i y_i)_{i\in I}$$

$$\preceq := \left\{((x_i)_{i\in I}, (y_i)_{i\in I}) \in \prod_{i\in I} S_i \times \prod_{i\in I} S_i \mid x_i \leq_i y_i \ \forall\, i \in I\right\}$$

is an ordered semigroup.
In the following we consider the $\prod\limits_{i\in I} S_i$ as the ordered semigroup with the multiplication and the order defined above.

**Lemma 1.** *Let $\{(S_i, \circ_i, \leq_i) \mid i \in I\}$ be a family of ordered semigroups. If $J_i$ is an ideal of $S_i$ for every $i \in I$, then the set $\prod\limits_{i\in I} J_i$ is an ideal of $\prod\limits_{i\in I} S_i$.*

*Proof.* 1) $\emptyset \neq \prod\limits_{i\in I} J_i \subseteq \prod\limits_{i\in I} S_i$ (since $J_i \neq \emptyset \ \forall\, i \in I$).
2) $\prod\limits_{i\in I} S_i * \prod\limits_{i\in I} J_i \subseteq \prod\limits_{i\in I} J_i$. In fact:
Let $(x_i)_{i\in I} \in \prod\limits_{i\in I} S_i$ and $(y_i)_{i\in I} \in \prod\limits_{i\in I} J_i$. Since $x_i \in S_i$ and $y_i \in J_i$ for every $i \in I$, we have $x_i \circ_i y_i \in S_i \circ_i J_i \subseteq J_i$ for every $i \in I$. Then we have

$$(x_i)_{i\in I} * (y_i)_{i\in I} := (x_i \circ_i y_i)_{i\in I} \in \prod_{i\in I} J_i$$

.
3) Let $(y_i)_{i\in I} \in \prod\limits_{i\in I} J_i$ and $\prod\limits_{i\in I} S_i \ni (x_i)_{i\in I} \preceq (y_i)_{i\in I}$. Then $(x_i)_{i\in I} \in$

$\prod_{i\in I} J_i$.

Indeed: Since $y_i \in J_i$, $S_i \ni x_i \leq_i y_i$ and $J_i$ is an ideal of $S_i$ for every $i \in I$, we have $x_i \in J_i$ for every $i \in I$. Then $(x_i)_{i\in I} \in \prod_{i\in I} J_i$. Similarly, the set of $\prod_{i\in I} J_i$ is a right ideal of $\prod_{i\in I} S_i$. □

In the following, we denote by $S$ the closed interval $[0,1]$ of real numbers. The set $S := [0,1]$ with the usual multiplication- order "." and "$\leq$" is an ordered semigroup.

**Lemma 2.** *If $a \in S$, then the set $I_a := [0,a]$ is an ideal of $S$.*

*Proof.* First of all $\emptyset \neq I_a \subseteq S$ (since $a \in [0,a]$). Let $x \in S$, $y \in I_a$. Since $0 \leq x \leq 1$, $0 \leq y \leq a$, we have $0 \leq xy \leq 1a = a$. Then $xy \in I_a$. Let $y \in I_a$ and $S \ni x \leq y$. Since $0 \leq x$, $y \leq a$ and $x \leq y$, we have $0 \leq x \leq a$. Then $x \in I_a$. Similarly, the set $I_a$ is a right ideal of $S$. □

**Theorem.** *Let $(S, ., \leq)$ be a commutative ordered semigroup with identity. If $M$ is a maximal ideal of $S$, then $M$ is a prime ideal of $S$. The converse statement does not hold, in general.*

*Proof.* Let $e$ be the identity of $S$, and $M$ a maximal ideal of $S$.
Let $a,b \in S$, $ab \in M$, $a \notin M$. Then $b \in M$. In fact:
Since $S$ is commutative, we have

$$\begin{aligned}\mathcal{I}(M \cup \{a\}) = \quad & ((M \cup \{a\}) \cup S(M \cup \{a\}) \cup (M \cup \{a\})S \cup S(M \cup \{a\})S] \\ = \quad & ((M \cup \{a\}) \cup S(M \cup \{a\}) \cup S^2(M \cup \{a\})].\end{aligned}$$

Since $M \cup \{a\} = e(M \cup \{a\}) \subsetneq S(M \cup \{a\})$, we have

$$S(M \cup \{a\}) \subseteq S^2(M \cup \{a\}) \subseteq S(M \cup \{a\}),$$

then $S(M \cup \{a\}) = S^2(M \cup \{a\})$.

Hence we have

$$\mathcal{I}(M \cup \{a\}) = (S(M \cup \{a\})] \text{..........} (*)$$

On the other hand, $M \subset M \cup \{a\} \subseteq \mathcal{I}(M \cup \{a\})$ (since $a \notin M$). Since $\mathcal{I}(M \cup \{a\})$ is an ideal and $M$ a maximal ideal of $S$, we have $\mathcal{I}(M \cup \{a\}) = S$, and $e \in (S(M \cup \{a\})]$ by (*). Then there exist $x \in S$ and $y \in M \cup \{a\}$ such that $e \leq xy$. Then $b = eb \leq xyb$. If $y \in M$, then $xyb \in SMS \subseteq M$, and $b \in M$. If $y = a$, then $b \leq x(ab) \in SM \subseteq M$, and $b \in M$.

For the converse statement, we consider the ordered semigroup $S := [0, 1]$ and the ordered semigroup $(S \times S, * \preceq)$ constructed above. The set $(S \times S, *, \preceq)$ is a commutative ordered semigroup and the element $(1, 1)$ is the identity element of $S \times S$. Let

$$T := S \times \{0\}(= [0, 1] \times \{0\}).$$

Clearly $S$ is an ideal of $S$. By Lemma 2, the set $I_0(= \{0\})$ is an ideal of $S$. Then, by Lemma 1, the set $T := S \times \{0\}$ is an ideal of $S \times S$.
The set $T$ is a prime ideal of $S \times S$. In fact:
Let $(x, y), (z, w) \in S \times S, (x, y) * (z, w) \in T$. Since $(x, y) * (z, w) := (xz, yw) \in T := S \times \{0\}$, we have $yw = 0$, then $y = 0$ or $w = 0$. Then $(x, y) \in S \times \{0\} := T$ or $(z, w) \in S \times \{0\} := T$.
The set $T$ is not a maximal ideal of $S \times S$. Indeed: By Lemma 2, the set $[0, 1/2] := I_{1/2}$ is an ideal of $S$. By Lemma 1, the set $S \times [0, 1/2]$ is an ideal of $S \times S$. On the other hand, $T := S \times \{0\} \subset S \times [0, 1/2] \subset S \times S$. □

This is part of our research work supported by the Ministry of Development, General Secretariat of Research and Technology -International Cooperation Division (Greece-Russia), Grant No. 70/3/4967.

## References

[1] D. M. Burton, *A First Course in Rings and Ideals*, Addison-Wesley, 1970.

[2] N. Kehayopulu, *On weakly prime ideals of ordered semigroups*, Mathematica Japonica, **35** (No. 6) (1990), 1051-1056.

[3] N.Kehayopulu, *Note on Green's relations in ordered semigroups*, Mathematica Japonica, **36** (No. 2) (1991), 211-214.

Contact information

**N. Kehayopulu,**
**M. Tsingelis**
University of Athens, Department of Mathematics Section of algebra and geometry, Panepistemiopolis, Athens 157 84, GREECE
*E-Mail:* nkehayp@cc.uoa.gr

**J. Ponizovskii**
Russian State Hydrometeorological University Department of Mathematics Malookhtinsky pr. 98 195196, Saint-Petersburg, Russia

Received by the editors: 06.12.2002.

Algebra and Discrete Mathematics RESEARCH ARTICLE
Number 2. **(2003).** pp. 36–67


# On intersections of normal subgroups in free groups

**O. V. Kulikova**

Communicated by A. Yu. Olshanskii

Abstract. Let $N_1$ (respectively $N_2$) be a normal closure of a set $R_1 = \{u_i\}$ (respectively $R_2 = \{v_j\}$) of cyclically reduced words of the free group $F(A)$. In the paper we consider geometric conditions on $R_1$ and $R_2$ for $N_1 \cap N_2 = [N_1, N_2]$. In particular, it turns out that if a presentation $< A \mid R_1, R_2 >$ is aspherical (for example, it satisfies small cancellation conditions $C(p)\&T(q)$ with $1/p + 1/q = 1/2$), then the equality $N_1 \cap N_2 = [N_1, N_2]$ holds.

## Introduction

Let $F = F(A)$ be a free group generated by an alphabet $A$. Let $N_1$ (respectively $N_2$) be the normal closure of a set $R_1 = \{u_i\}$ (respectively $R_2 = \{v_j\}$) of cyclically reduced words of $F$. We will consider non-intersecting symmetrized $R_1$ and $R_2$.

It is evident that the inclusion $[N_1, N_2] \subset N_1 \cap N_2$ always holds. But the reverse inclusion does not always hold (the simplest example is $R_1 = \{a_1, {a_1}^{-1}\}$ and $R_2 = \{a_1, a_2, {a_1}^{-1}, {a_2}^{-1}\}$). The aim of this paper is to find necessary and sufficient conditions on $R_1$ and $R_2$ for

$$N_1 \cap N_2 = [N_1, N_2]. \qquad (1)$$

These conditions are expressed in terms of certain geometric objects called pictures (see, for example [4]).

**2000 Mathematics Subject Classification:** *20F05, 20F06.*
**Key words and phrases:** *normal closure of words in free groups, presentations of groups, pictures, mutual commutants, intersection of groups, aspherisity, small cancellation conditions.*

In particular, it turns out that if a presentation $< A \mid R_1, R_2 >$ is aspherical (for example, it satisfies small cancellation conditions $C(p)\&T(q)$ with $1/p + 1/q = 1/2$ ([4], theorem 2.2)), then the equality $N_1 \cap N_2 = [N_1, N_2]$ holds.

The paper is divided into three sections, each of which is further subdivided. In the first section we give main definitions, prove some results concerning relations between them, formulate the main result of the paper (Theorem 1) and prove corollaries of it. The second section is devoted to the proof of Theorem 1. In the third section we prove some simple corollaries of Theorem 1 in the case of free products.

It should be noted that a geometric test of the equality (1) obtained in Theorem 1 is hard to verify, but its corollaries are useful. The following question seems to be open: if the equality (1) holds if and only if there exist sets of words $\tilde{R}_1$ and $\tilde{R}_2$ such that
(i) $\tilde{R}_1^F = N_1$ and $\tilde{R}_2^F = N_2$;
(ii) the presentation $G =< A \mid \tilde{R}_1 \cup \tilde{R}_2 >$ is strictly $(\tilde{R}_1, \tilde{R}_2)$-separable (see Definition 1 below).

I am very grateful to my scientific advisor professor A. Yu. Ol'shanskii for his constant attention to my work. Also I am grateful to A. A. Klyachko and D. V. Osin for conversations during preparation of this paper.

## 1. Formulation and corollaries of Theorem 1.

In the beginning we give some definitions and recall the definition of pictures.

### 1.1 Main definitions. Relations between definitions.

Let $N$ be the normal closure of a set $R$ of cyclically reduced words of the free group $F(A)$.

*A picture* $P$ over presentation $G =< A \mid R >$ on a surface $S$ ( see details in [4] ) is a finite collection of "vertices" $V_1, ..., V_n \in S$, together with a finite collection of simple pairwise disjoint connected oriented "edges" $E_1, ..., E_m \in S \setminus \{\{V_1, ..., V_n\} \cup \partial S\}$, labelled by letters of $A$ (in [4], vertices are called discs and edges are called arcs). But these edges need not all connect two vertices. An edge may connect a vertex to a vertex (possibly coincident), a vertex to $\partial S$, or $\partial S$ to $\partial S$. Moreover, some edges need have no endpoints at all, but be circles disjoint from the rest of $P$, such edges are called edges-circles. If an edge connects two vertices, then one is the start of the edge, the other one is the end of the edge.

For each vertex $V$ of $P$ consider a circle $C$ of a small radius with center at $V$ and a point $p$ on $C$ not lying on any edge of $P$. The labels of edges intersected by $C$ starting at $p$ form a word $r \in R$. Changing of the disposition of $p$ on $C$ and the direction of moving around $C$, we can read any cyclic permutation of $r$ and $r^{-1}$.

Below we will consider pictures on $S$, where $S$ is a sphere (spherical pictures) or a disk (planar pictures).

In the case of a planar picture the labels of edges, intersected by a circle $\bar{C}$ near the boundary of the disk $\partial S$, starting at a point $\bar{p}$ on $\bar{C}$, form a word $W$, which will be called *the boundary label* of the picture.

The following result is well-known (use Theorem 11.1 of [1] and dualise):

**Lemma 1.** *Let $W$ be a non-empty word on the alphabet $A$. Then $W$ represents the identity of the group $G = F/N$ if and only if there is a planar picture over the presentation $G =< A \mid R >$ with the boundary label $W$.*

Let $P$ be a picture over $G =< A \mid R >$ and $\gamma$ be a path on $S$ not passing through any vertex of $P$. If we travel around $\gamma$ we encounter a succession of edges. Reading the labels on these edges gives a word called *the word along the path* $\gamma$ and denoted by $Lab(\gamma)$.

If $\gamma$ is closed, consider a point $p$ of $\gamma$ not lying on any edge of $P$. The word along $\gamma$ read from $p$ will be denoted by ${Lab_p}^+(\gamma)$ or by ${Lab_p}^-(\gamma)$ (depending as the direction of reading is counterclockwise or not). It is clear that ${Lab_p}^+(\gamma)^{-1} = {Lab_p}^-(\gamma)$. Changing the disposition of $p$ we obtain the same word up to cyclic permutation. Changing the direction of reading we obtain the inverse word. We will denote the word along the path $\gamma$ by $Lab(\gamma)$ when the disposition of $p$ and the direction of reading are not essential.

By $\mathbf{1}$ denote the identity element of the free group.

A *dipole* in a picture $P$ over $G =< A \mid R >$ is two vertices $V_1$ and $V_2$ of $P$ if there is a simple path $\psi$ connecting points $p_1$ and $p_2$ lying on circles $C_1$ and $C_2$ around these vertices such that the following conditions hold:

(i) $Lab(\psi) = \mathbf{1}$;

(ii) ${Lab_{p_1}}^+(C_1) = {Lab_{p_2}}^-(C_2)$.

A picture over $G =< A \mid R >$ is *reduced* if it does not contain a dipole.

A presentation $G =< A \mid R >$ is *aspherical* if every connected spherical picture over $G =< A \mid R >$ contains a dipole.

**Definition 1.** *Let $R_1$ and $R_2$ be two sets of words in the free group $F(A)$. We say that a presentation $G =< A \mid R_1 \cup R_2 >$ is strictly $(R_1, R_2)$-separable or satisfies the condition of strict $(R_1, R_2)$-separability if for every reduced spherical picture $P$ containing both $R_1$-vertices and $R_2$-vertices there is a simple closed path $\gamma$ dividing the sphere into two disks such that the following conditions hold:*

1) *the both disks contain vertices;*

2) *$Lab(\gamma) = \mathbf{1}$.*

**Assertion 1.** *If in Definition 1 "every reduced spherical picture" is replaced by "every spherical picture", then the class of presentations satisfying strict $(R_1, R_2)$-separability is not changed.*

*Proof.* Let $P$ be a non-reduced spherical picture over $G =< A \mid R_1 \cup R_2 >$ containing both $R_1$-vertices and $R_2$-vertices. Since $P$ is not reduced, there is a dipole, i.e., there are two vertices $V_1$ and $V_2$ that inverse words from $R_1 \cup R_2$ correspond to, and $V_1$ and $V_2$ can be connected by a simple path $\psi$ such that $Lab(\psi) = \mathbf{1}$. It is easily seen that the simple closed path $\gamma$ from Definition 1 may be obtained going around $V_1$ and $V_2$ and by-passing near $\psi$ in the both directions. □

**Definition 2.** *Let $R_1$ and $R_2$ be two sets of words in $F(A)$. We say that a presentation $G =< A \mid R_1 \cup R_2 >$ is weakly $(R_1, R_2)$-separable or satisfies the condition of weak $(R_1, R_2)$-separability if for every reduced spherical picture $P$ containing both $R_1$-vertices and $R_2$-vertices there is a simple closed path $\gamma$ dividing the sphere into two disks such that the following three conditions hold:*

1) *the both disks contain vertices;*

2) *$Lab(\gamma) \in [N_1, N_2]$;*

3) *if one of the disks contains only $R_1$-vertices, then the other one contains only $R_2$-vertices.*

**Assertion 2.** *If the condition* 3) *in Definition* 2 *is omitted, then the class of presentations satisfying only* 1) *and* 2) *of Definition* 2 *will be wider.*

*Proof.* As a counterexample proving the assertion, one can consider a presentation $G =< A \mid R_1 \cup R_2 >$, where $R_1$ is a symmetrized set of words $\{a_1, a_3, [a_1, a_2]\}$ and $R_2$ is a symmetrized set of words $\{a_2, a_3[a_1, a_2]\}$, where $\{a_i\} \in A$. This presentation satisfies the conditions $1), 2)$ of Definition 2, but it does not satisfy the condition 3).

Indeed, let $P$ denote a reduced spherical picture containing both both $R_1$-vertices and $R_2$-vertices.

If there is an $R_1$-vertex labelled by $[a_1, a_2] \in [N_1, N_2]$ in $P$, then the path $\gamma$ is obtained going around this vertex.

If, in $P$, there is not any $R_1$-vertex labelled by $[a_1, a_2]$ and there is an $R_1$-vertex labelled by $a_3$, then the edge starting at this $a_3$-vertex must have the end at an $R_2$-vertex labelled by $a_3[a_1, a_2]$, since $P$ is reduced. Thus the path $\gamma$ is obtained going around this $(a_3)$-vertex, the edge starting at this $a_3$-vertex and the $(a_3[a_1, a_2])$-vertex.

If in $P$ there are neither $R_1$-vertices labelled by $[a_1, a_2]$ nor $R_1$-vertices labelled by $a_3$, then $P$ is non-reduced. Hence the conditions $1), 2)$ of Definition 2 hold.

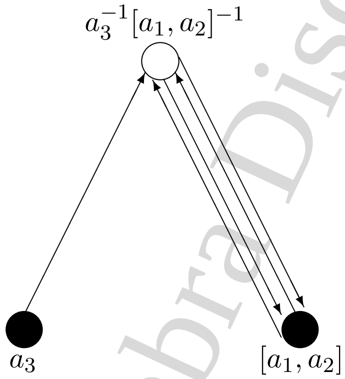



Fig. 1

But the presentation does not satisfy the condition 3), which is shown by the picture containing only three vertices: an $(a_3)$-vertex, an $([a_1, a_2])$-vertex and an $(a_3[a_1, a_2])$-vertex (see Fig.1), since $a_3$, $a_3[a_1, a_2] \notin [N_1, N_2]$. □

**Assertion 3.** *If a presentation $G =< A \mid R_1 \cup R_2 >$ is strictly $(R_1, R_2)$-separable, then it is weakly $(R_1, R_2)$-separable.*

*Proof.* Let $P$ be a reduced spherical picture over strictly $(R_1, R_2)$-separable $G =< A \mid R_1 \cup R_2 >$ containing both $R_1$-vertices and $R_2$-vertices. It is sufficient to find a simple closed path $\gamma$ dividing the sphere into two disks such that

1) the both disks contain vertices;

2) $Lab(\gamma) = \mathbf{1}$;

3) one of the disks contains only $R_1$-vertices and the other one contains only $R_2$-vertices.

It follows from the condition of strict $(R_1, R_2)$-separability that there is a simple closed path $\gamma_1$ in $P$ dividing the sphere into two disks such that the following conditions hold:

1) the both disks contain vertices;

2) $Lab(\gamma_1) = \mathbf{1}$.

If at least one of these disks both contains $R_1$-vertices and $R_2$-vertices, then it follows from the condition of strict $(R_1, R_2)$-separability that there is another simple closed path $\gamma_2$ in $P$ non-crossing the path $\gamma_1$ and dividing such disk into two subdisks such that the following conditions hold:

1) the both subdisks contain vertices;

2) $Lab(\gamma_2) = \mathbf{1}$.

We continue in this fashion to obtain a finite set of simple closed paths $\Gamma = \{\gamma_i\}$ satisfying the following conditions:

1) these paths are pairwise disjoint;

2) $Lab(\gamma) = \mathbf{1}$ for each $\gamma_i \in \Gamma$;

3) the union of the paths $\gamma_i \in \Gamma$ divides the sphere into parts $\{D_k\}$ each of which contains only $R_1$-vertices or only $R_2$-vertices.

Each path of $\Gamma$ separates one part of $\{D_k\}$ from another one. If any path $\gamma_i \in \Gamma$ separates one part of $\{D_k\}$ from another one so that the both parts contain either only $R_1$-vertices or only $R_2$-vertices, then we will remove this path $\gamma_i$ from $\Gamma$. We thus get that each path of $\Gamma$ separates a part of $\{D_k\}$ containing only $R_1$-vertices from another part containing only $R_2$-vertices.

Below we will transform each not simply connected part $D_i$ of $\{D_k\}$ in order to decrease the number of paths in $\Gamma$. Since $D_i$ is not simply connected, there are at least two paths $\gamma_i$ and $\gamma_j$ of $\Gamma$ bounding the part $D_i$. Join the paths as follows. Fix a point $a_i^+$ on the path $\gamma_i$ and a point $a_j^+$ on $\gamma_j$ so that $a_i^+$ and $a_j^+$ do not belong to any edge of the picture $P$. It is clear that the points $a_i^+$ and $a_j^+$ may be joined by a simple path $\psi_{(i,j)}^+$ so that

1) the whole path $\psi_{(i,j)}^+$ lies in $D_i$;

2) $\psi_{(i,j)}^+$ does not pass through any vertex of $P$;

3) $\psi^+_{(i,j)}$ does not intersect the paths of $\Gamma$ except $\gamma_i$ and $\gamma_j$ at the points $a_i^+$ and $a_j^+$.

It is possible to draw a path $\psi^-_{(i,j)}$ through points $a_i^- \in \gamma_i$ and $a_j^- \in \gamma_j$ close to the path $\psi^+_{(i,j)}$, in a parallel way with the properties similar to $\psi^+_{(i,j)}$ so that $\psi^-_{(i,j)}$ intersects the same edges of $P$ as $\psi^+_{(i,j)}$ does, and that $\psi^+_{(i,j)}$ and $\psi^-_{(i,j)}$ are disjoint. Besides $\psi^-_{(i,j)}$ can be drawn so that the segment $[a_i^+, a_i^-]$ of $\gamma_i$ and the segment $[a_j^+, a_j^-]$ of $\gamma_j$ do not intersect the edges of $P$. Removing these segments gives a new simple closed path $\gamma_{ij} = (\gamma_i/[a_i^+, a_i^-]) * \psi^+_{(i,j)} * (\gamma_j/[a_j^+, a_j^-]) * \psi^-_{(i,j)}$ such that $Lab(\gamma_{ij}) = \mathbf{1}$. Replacing $\gamma_i$ and $\gamma_j$ by $\gamma_{ij}$ in $\Gamma$ gives a set of paths satisfying the same properties $1), 2), 3)$. Besides the number of paths in the resulting set of paths becomes less than in the original set.

Consequently after a finite number similar transformations, all parts from $\{D_k\}$ become simply connected. Therefore the resulting set $\Gamma$ contains just one path, which is the desired path $\gamma$. ☐

**Assertion 4.** *The condition of weak $(R_1, R_2)$-separability is not equivalent to the condition of strict $(R_1, R_2)$-separability.*

*Proof.* As an example proving the assertion one can consider a presentation $G =< A \mid R_1 \cup R_2 >$, where $R_1$ is a symmetrized set of words $\{a_1, [a_1, a_2]\}$ and $R_2 = \{a_2, {a_2}^{-1}\}$, where $\{a_i\} \in A$.

It follows from Theorem 1 and Corollary 5 (see below) that this presentation is weakly $(R_1, R_2)$-separable since $N_1 \cap N_2 = [N_1, N_2]$.

Let us show that the existence of a spherical picture $P$ containing only an $[a_1, a_2]$-vertex and two $a_2$-vertices contradicts the condition of strict $(R_1, R_2)$-separability.

Indeed, suppose that there is a simple closed path $\gamma$ dividing the sphere into two disks. Then one of the disks must contain only one vertex. Consequently $Lab(\gamma)$ must be equal to the label of this vertex, which is not equal to the identity element in the free group. This contradicts the definition of strict $(R_1, R_2)$-separability. ☐

**Assertion 5.** *If every spherical picture over a presentation $G =< A \mid R_1 \cup R_2 >$ containing both $R_1$-vertices and $R_2$-vertices is not reduced (this condition will be called $(R_1, R_2)$-asphericity), then the presentation is strictly $(R_1, R_2)$-separable.*

*Proof.* is similar to the proof of Assertion 1. ☐

**Definition 3.** *Let $R_1$ and $R_2$ be two sets of words in $F(A)$. We say that a presentation $G =< A \mid R_1 \cup R_2 >$ is $(R_1, R_2)$-separable or satisfies the condition of $(R_1, R_2)$-separability if for every spherical picture*

*$P$ containing both $R_1$-vertices and $R_2$-vertices there is a simple closed path $\gamma$ dividing the sphere into two disks such that the following three conditions hold:*

1) *the both disks contain vertices;*

2) *$Lab(\gamma) \in [N_1, N_2]$;*

3) *one of the disks contains only $R_1$-vertices and the other one contains only $R_2$-vertices.*

## 1.2 Formulation of Theorem 1 and corollaries from it.

**Theorem 1.** *A presentation $G =< A \mid R_1 \cup R_2 >$ is weakly $(R_1, R_2)$-separable if and only if $N_1 \cap N_2 = [N_1, N_2]$.*

**Corollary 1.** *The conditions of weak $(R_1, R_2)$-separability and $(R_1, R_2)$-separability are equivalent.*

*Proof.* It is easy to see that if a presentation is $(R_1, R_2)$-separable, then it is weakly $(R_1, R_2)$-separable.

It remains to prove the converse statement. Let $P$ be a spherical picture over a weakly $(R_1, R_2)$-separable presentation $G =< A \mid R_1 \cup R_2 >$ containing both $R_1$-vertices and $R_2$-vertices. It is evident that there is a simple closed path $\gamma$ not passing through any vertex of $P$ and dividing the sphere into two disks one of which contains only $R_1$-vertices and the other one contains only $R_2$-vertices. Hence $Lab(\gamma) \in N_1 \cap N_2$. Since $G =< A \mid R_1 \cup R_2 >$ is weakly $(R_1, R_2)$-separable, Theorem 1 leads to $Lab(\gamma) \in [N_1, N_2]$. Consequently $\gamma$ is desired. □

**Corollary 2.** *The conditions of weak $(R_1, R_2)$-separability and weak $(R_2, R_1)$-separability are equivalent. Moreover we get the equivalent condition if in Definition 2 the item 3) is replaced by*
$3'$) *if one disk contains both $R_1$- both $R_1$- and $R_2$-vertices, then the other one both $R_1$- and $R_2$-vertices.*

**Corollary 3.** *Let a presentation $G =< A \mid R_1 \cup R_2 >$ satisfy one of the following conditions:*

*(i)* *strict $(R_1, R_2)$-separability;*

*(ii)* *$(R_1, R_2)$-asphericity;*

*(iii)* *asphericity.*

*Then $N_1 \cap N_2 = [N_1, N_2]$.*

*Proof.* (i) It follows directly from Theorem 1 and Assertion 3.
(ii) It follows directly from (i) and Assertion 4.
(iii) It follows from (ii). □

**Corollary 4.** *Let $\{R_1, R_2\}$ be a set of words satisfying one of the following small cancellation conditions: either $C(6)$, or $C(4)\&T(4)$, or $C(3)\&T(6)$. Then $N_1 \cap N_2 = [N_1, N_2]$.*

*Proof.* According to [4] (see also [3]), every small cancellation condition described above is sufficient to asphericity of $G =< A \mid R_1 \cup R_2 >$. □

**Corollary 5.** *Suppose that:*

1) *an alphabet $A = X \sqcup Y \sqcup Z$;*

2) *$R_1$ is an arbitrary set of words on $X$ and $R_2$ is an arbitrary set of words on $Y$.*

*Then $N_1 \cap N_2 = [N_1, N_2]$.*

*Proof.* Without loss of generality, one can assume that $R_1$ and $R_2$ are symmetrized.

By Corollary 3 it is sufficient to show that $G =< A \mid R_1 \cup R_2 >$ is strictly $(R_1, R_2)$-separable.

Let $P$ be a reduced spherical picture containing both both $R_1$-vertices and $R_2$-vertices.

If there is an edge-circle in $P$ dividing the sphere into two disks each of which contains vertices, then Corollary 5 is proved (the simple closed path can be drawn near this edge-circle).

If there is an edge-circle in $P$ dividing the sphere into two disks one of which does not contain vertices, then this edge-circle can be removed from $P$.

Therefore we can suppose that there is no edge-circle, hence each edge connects a vertex to a vertex.

Since $R_1$ and $R_2$ are written on the disjoint alphabets $X$ and $Y$, we conclude that if any edge starts at an $R_1$-vertex, then it ends at an $R_1$-vertex (similarly for $R_2$-vertices).

For an $R_1$-vertex, consider a connected component of $P$ containing this vertex. All vertices of this component are $R_1$-vertices. Since $P$ also contains $R_2$-vertices, the complement of this component also contains vertices and falls into connected components. Among all these connected components there is at least one, which can be covered by a domain homeomorphic to a disk such that this domain does not intersect the other connected components. Then the boundary of the domain forms the desired simple closed path $\gamma$. $Lab(\gamma) = \mathbf{1}$, since no edge of $P$ intersects $\gamma$, and the statement follows. □

## 2. Proof of Theorem 1.

In the beginning we give several definitions, which we will use in the proof of Theorem 1.

### 2.1 Additional definitions.

1) *Picture $P$ with equator $Equ$. Subpictures of $P$.*
Let $P$ be a picture on the sphere $S^2$ over a presentation $G =< A \mid R_1 \cup R_2 >$. Fix a simple closed path (denoted by $Equ$) on $S^2$ not passing through any vertex of $P$ and dividing $S^2$ into two parts so that one part does not contain $R_1$-vertices and the other one does not contain $R_2$-vertices. The path $Equ$ is called an *equator*.

$Lab(Equ)$ is denoted by $W$ or $W^{-1}$ (the sign depends on the direction of reading). By the choice of the equator, it follows that $W \in N_1 \cap N_2$.

**Remark 1.** The start point on $Equ$ is not fixed, i.e., the equatorial label $W$ is considered as a cyclic word (if $W$ belongs to some normal subgroup, then all its cyclic permutations belong to the same subgroup). For simplicity of notation all cyclic permutations of $W$ will be denoted again by $W$.

In the sequel, $P$ denotes a picture with a fixed equator $Equ$.

Let $P = P' \sqcup P''$ be a disjoint union of two spherical pictures. $P'$ (respectively, $P''$) will be called *a subpicture* of $P$. $P'$ (respectively, $P''$) may be both connected and disconnected.

*2) Boundaries of vertices. North and south vertices.*
Every vertex is labelled by a word of $R_1$ ($R_1$-word) or $R_2$ ($R_2$-word). Consider a small disk with centre at a vertex such that the word along its boundary coincides with the label of the vertex. The boundary of the disk will be called *the boundary of the vertex*.

$Equ$ divides the sphere into two hemispheres: so called north and south. A vertex is called *north* (respectively, *south*) if it lies in the north (respectively, south) hemisphere.

Suppose that $R_1$-words correspond to the north vertices (so called $R_1$-vertices), $R_2$-words correspond to the south vertices (so called $R_2$-vertices). Each word can be read along the boundary of the corresponding vertex starting at some point on the boundary and choosing the direction of reading.

*3) Admissible moves.*
In the proof of Theorem 1 we will transform a picture $P$ with an equator $Equ$ on $S^2$. A move (i.e., a transformation) is called *admissible* if, after the move, the word $W$ along $Equ$ is replaced by a word $W'$ equal to

$W$ to within an element of $[N_1, N_2]$ (for simplicity of notation, we will use the same letter $W$ for the notation $W'$). Moreover admissible moves preserve the subdivision of $P$ by $Equ$ into the north and south vertices.

*4) Map.*
Each domain $U \subset S^2$ homeomorphic to a square

$$\{(x, y) \in \mathbb{R}^2 \mid -1 < x < 1, -1 < y < 1\},$$

together with vertices and parts of edges lying in it is called *a map.*

*5) Components.*
Each spherical subpicture of $P$ is called *a component* if it contains vertices. A component is called *reduced* (respectively, *non-reduced*) if the corresponding subpicture is reduced (respectively, non-reduced). A component is called north (respectively, south) if the corresponding subpicture contains only north (respectively, only south) vertices.

South and north components are called *uniform.* A component is called *mixed* if the corresponding subpicture contains both south and north vertices. Note that $P$ is a component of itself.

*6) States. Territories of states. Boundaries of states.*
A component is called a *state* if it is uniform and can be covered by a closed domain homeomorphic to a disk so that the domain does not intersect the other components of $P$. The state is called *north* (respectively, *south*) if the corresponding uniform component is north (respectively, south).

The domain mentioned above is called a *territory* of the state. The boundary of the territory is called the *boundary of the state.* Note that the boundary of each state is homeomorphic to a circle. Without loss of generality we can assume that the boundary of each state intersects $Equ$ in a finite set of points the number of which can not be decreased by changing the territory.

*7) Pieces of equator Equ.*
Let $T$ be a state. $Equ$ intersects its boundary $2m$ times and is divided into $2m$ connected parts by these intersection points. Those of parts which lie on the territory of $T$ are called *pieces of the equator.*

*8) Regions of states.*
Let $T$ be a north state. $Equ$ divides the territory of $T$ into connected parts, which are called *regions north* or *south* depending on what hemisphere they belong to. A north region of a north state is called *regular* if its boundary consists exactly of two parts: a connected part belonging to the boundary of $T$ and a piece of the equator. Respectively, a north region of a north state is called *irregular* if it is not regular. Similarly, one can define regions and regular regions of south states by replacing the word "north" by "south".

9) *σ-state.*
A north state is called a north *σ-state* if all its north regions are regular. (South σ-states are defined similarly by replacing the word "north" by "south".)

10) *σ-picture.*
A picture with a fixed equator is called σ-picture if it can be reduced to a picture containing only σ-states by a finite number of admissible moves.

## 2.2 Proof of Theorem 1.

The proof of a statement that weak $(R_1, R_2)$-separability implies $N_1 \cap N_2 = [N_1, N_2]$ is similar to the proof of Corollary 1.

Therefore it remains to prove the converse statement. Let a presentation $G =< A \mid R_1 \cup R_2 >$ satisfy weak $(R_1, R_2)$-separability.

Since the inclusion $[N_1, N_2] \subset N_1 \cap N_2$ always holds, it is sufficient to prove the reverse inclusion.

Let $W$ be an arbitrary word of the intersection $N_1 \cap N_2$. Then there are two representations:

$$W = \prod_k g_k r_{1,k} {g_k}^{-1}, \qquad \text{since } W \in N_1; \tag{2}$$

and

$$W^{-1} = \prod_l h_l r_{2,l} {h_l}^{-1}, \qquad \text{since } W \in N_2, , \tag{3}$$

where $g_k, h_l$ are words of $F$, each $r_{1,k}$ belongs to $R_1$, and each $r_{2,l}$ belongs to $R_2$. To show that $W \in [N_1, N_2]$, let us construct two planar pictures. The word $W$ in the form (2) is written on the boundary of the first picture which contains only $R_1$-vertices. The word $W^{-1}$ in the form (3) is written on the boundary of the second picture which contains only $R_2$-vertices. Pasting together the planar pictures by their boundaries gives a picture $P$ on $S^2$ with a fixed equator $Equ$. $Lab(Equ)$ is equal to $W$ or $W^{-1}$ depending on the direction of moving along $Equ$.

Now the proof of Theorem 1 follows from the following Propositions 1 and 2, which will be proved in the subsections 2.4 and 2.5 below:

**Proposition 1.** *Let a picture $P$ with a fixed equator $Equ$ be over a presentation $G =< A \mid R_1 \cup R_2 >$. If the presentation satisfies weak $(R_1, R_2)$-separability, then $P$ is a σ-picture.*

**Proposition 2.** *Let a picture $P$ with a fixed equator $Equ$ be over a presentation $G =< A \mid R_1 \cup R_2 >$. If $P$ is a σ-picture, then the word $W$ along $Equ$ can be reduced to the identity element in the free group by a finite number of admissible moves.*

Indeed, by Propositions 1, 2, there is a finite sequence of admissible moves, which reduces the equatorial label $W$ in $P$ to the identity element in the free group. Since admissible moves replace $W$ by words equal to $W$ to within elements of $[N_1, N_2]$, we have that $W \in [N_1, N_2]$, as claimed. □

### 2.3 Some admissible moves. Auxiliary lemmas.

In the subsection 2.3 we describe admissible moves, which will be used in the proof of Propositions 1 and 2.

*1) Isotopy.*
*An isotopy of* $P$ is defined by replacing $P$ by a picture $F_1(P)$, where $F_t : S^2 \times [0,1] \to S^2 \times [0,1]$ is a continuous isotopy of the sphere $S^2$ such that

(i) $F_t$ leaves fixed all vertices, i.e., $F_t(V_i) = V_i$ for $t \in [0,1]$ and for each vertex $V_i$;

(ii) for each $t$ and each edge $E_j$ the intersection of an edge $F_t(E_j)$ and $Equ$ consists of a finite number of points, moreover the edge $F_1(E_j)$ intersects $Equ$ transversally at every intersection point.

It is evident that an isotopy of $P$ is an admissible move because either it corresponds to an insertion or a cancellation of pairs of inverse letters in the equatorial label $W$ (see Fig.2) or it does not change $W$ at all.

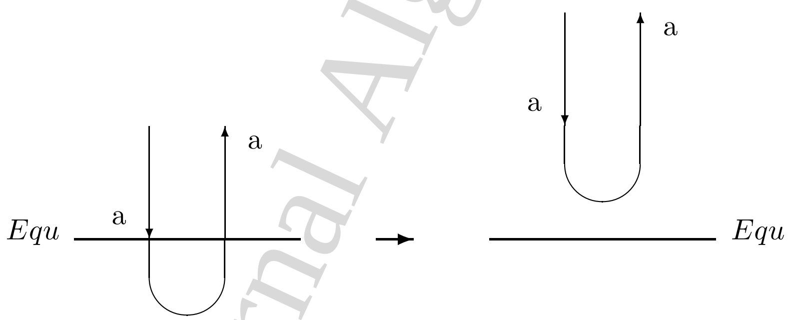



Fig. 2

2) *Bridge moves.*
Assume that a map $U$ contains only two edges

$$\{x = -1/2, -1 < y < 1\} \text{ and } \{x = 1/2, -1 < y < 1\},$$

which are contrariwise oriented and labelled by the same letter. A move of $P$ is called *a bridge move* (see also in [3]) if it does not change $P$ out of the map $U$ and change $P$ in $U$ as is shown on Fig 3.

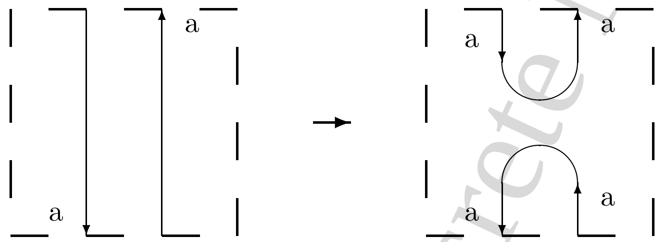



Fig. 3

The bridge move is an admissible move because either it corresponds to an insertion or a cancellation of pairs of inverse letters in the equatorial label $W$ or it does not change $W$ at all.

3) *Removing components and edge-circles of $P$ not intersecting Equ.* If a component (in particular, a state) of $P$ does not intersect $Equ$, then it does not contribute to the equatorial label $W$. Therefore such component can be removed. Similarly one can remove edge-circles not intersecting $Equ$.

4) *Removing superfluous loops.*
Assume that $Equ$ intersects any edge in two successive points which divide $Equ$ into two parts such that one of these two parts does not intersect edges. Such part of the edge is called *a superfluous loop.* It is evident that superfluous loops do not contribute to the equatorial label $W$ (considered as an element of the free group). Therefore superfluous loops can be removed (see Fig. 2).

( This move is a special case of isotopy (see the move 1)) or a composition of the admissible moves 2) and 3).)

5) *Uniting $\sigma$-states.*
Let $T_1$ and $T_2$ be two distinguished north $\sigma$-states (the move of south states is similar). Assume that there are points $p_1$ on the boundary of $T_1$ and $p_2$ on the boundary of $T_2$ so that

(i) the points $p_1$ and $p_2$ lie in the south hemisphere;

(ii) it is possible to connect the points $p_1$ and $p_2$ by a simple path $\eta$ which does not intersect the territory of any state and lies in the south hemisphere as a whole.

Then the territories of the $\sigma$-states $T_1$ and $T_2$ can be united in one by adding a small neighborhood of the path $\eta$ (see Fig.4).

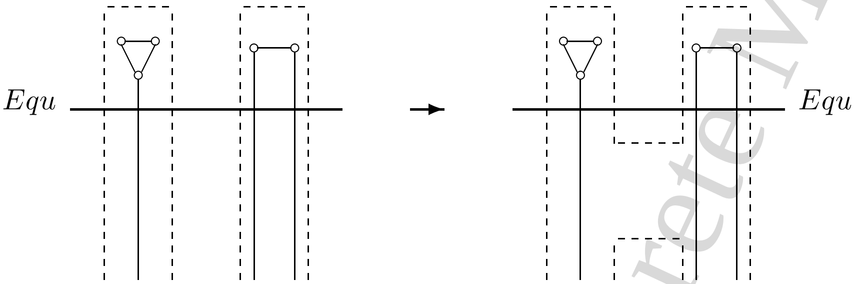



Fig. 4

It is clear that the united territory is homeomorphic to a disc again. The $\sigma$-states $T_1$ and $T_2$ lying in the united territory form one state, which is a $\sigma$-state again.

**Remark 2.** By a deformation of territories of $\sigma$-states, we can exclude the case when the territory of any $\sigma$-state successively intersects $Equ$ in two pieces $I_1$ and $I_2$ so that a part of $Equ$ between $I_1$ and $I_2$ does not intersect the territory of any state.

6) *Pasting a map with a commutator subpicture in $P$.*
In notation of the subsection 2.5, assume that there are two regular regions: one of them belongs to a north state $T_1$, the other one belongs to a south state $T_2$. By Lemma 2 below, the north region contains a picture with a word $w_1 \in N_1$ written along a piece $I_2$ of the equator, the south region contains a picture with a word $w_2 \in N_2$ written along a piece $I_3$ of the equator.

We choose a map $M_1$ so that the north region is contained in $\{0 \le y \le 1\}$, $I_2 \subset \{y = 0\}$, and the intersection of $T_1$ and $\{-1 < y < 0, -1 < x < 1\}$ contains only parts of edges given in coordinates $(x, y)$ by $\{x = x'_1, -1 < y < 0\}, ..., \{x = x'_{n'}, -1 < y < 0\}$.

Similarly we choose a map $M_2$ so that the south region is contained in $\{0 \ge y \ge -1\}$, $I_3 \subset \{y = 0\}$, and the intersection of $T_2$ and $\{0 < y < 1, -1 < x < 1\}$ contains only parts of edges given in coordinates $(x, y)$ by $\{x = x_1, 0 < y < 1\}, ..., \{x = x_n, 0 < y < 1\}$.

For the map $M_1$ ( respectively, for $M_2$) we construct a mirror-like map $M'_1$ (respectively, $M'_2$) by reflecting $M_1$ (respectively, $M_2$) with respect to the axis $\{x = 0\}$ and by changing the orientations of the edges. The map $M'_1$ (respectively, $M'_2$) contains the picture with the word ${w_1}^{-1}$

(respectively, ${w_2}^{-1}$) written along the piece of the equator. The piece of the equator in the map $M_1'$ (respectively, $M_2'$) will be denoted by $I_2'$ (respectively, $I_3'$).

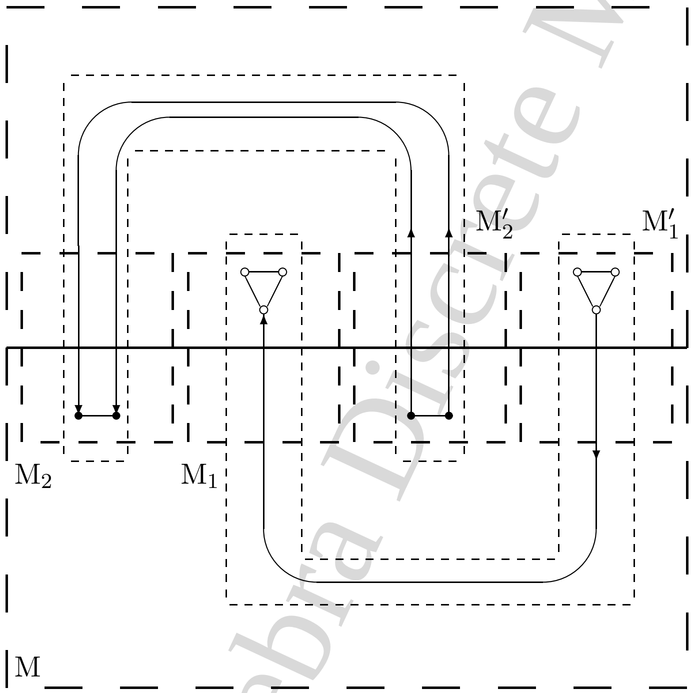



Fig. 5

A map $M$ containing a picture for the word $w_2 w_1 {w_2}^{-1} {w_1}^{-1}$ is constructed as follows:

(i) the equator passes through $\{y = 0\}$;

(ii) a part corresponding to the south hemisphere is $\{0 \geq y \geq -1\}$ and a part corresponding to the north hemisphere is $\{0 \leq y \leq 1\}$;

(iii) the map $M_2$ is disposed in the rectangle $\{-1/5 < y < 1/5, -4/5 < x < -3/5\}$; the map $M_1$ is disposed in $\{-1/5 < y < 1/5, -2/5 < x < -1/5\}$; the map $M_2'$ is in $\{-1/5 < y < 1/5, 1/5 < x < 2/5\}$; the map $M_1'$ is in $\{-1/5 < y < 1/5, 3/5 < x < 4/5\}$;

(iv) in the part corresponding to the south hemisphere, we extend the edges of the map $M_1$ to join them with the corresponding edges of $M_1'$; the union of the pictures of these maps and the joining edges forms a north $\sigma$-state which will be denoted by $T_1^{comm}$;

(v) in the part corresponding to the north hemisphere, we extend the edges of the map $M_2$ to join them with the corresponding edges of $M_2'$; the union of the pictures of these maps and the joining edges forms a south $\sigma$-state which will be denoted by $T_2^{comm}$. (See Fig. 5)

It is clear that the word along the equator of the map $M$ is $w_2 w_1 {w_2}^{-1} {w_1}^{-1}$.

A small map $M_s$ in $P$ containing nothing but a part $\{y = 0\}$ of $Equ$ is replaced by the constructed map $M$. This move is admissible because it corresponds to an insertion of the commutator $w_2 w_1 {w_2}^{-1} {w_1}^{-1}$ of the elements from $N_1$ and $N_2$ in the equatorial label $W$.

**Lemma 2.** *Let $T$ be a state in $P$. Let $I$ be an arbitrary piece of the equator belonging to the territory of $T$. Then $Lab(I) \in N_1$ if $T$ is north, and $Lab(I) \in N_2$ if $T$ is south.*

*Proof.* Let $T$ be north (the proof in the case of a south state is similar). The piece $I$ divides the territory of $T$ into two parts $T'$ and $T''$. We consider one of them denoted by $T'$. $T'$ may contain only north vertices (i.e., $R_1$-vertices) and edges labelled by letters of the alphabet $A$. Hence $T'$ contains a planar picture over $< A \mid R_1 >$. By Lemma 1, a word along the boundary of $T'$ belongs to $N_1$. Since the edges intersect the boundary of $T'$ only in a part which coincides with $I$, Lemma 2 follows. □

**Lemma 3.** *Let $T$ be a state in $P$. If $Equ$ intersects the boundary of $T$ exactly two times, then the word along the piece of the equator lying inside the territory of $T$ is equal to the identity element in the free group.*

*Proof.* Assume that $T$ is north (the proof in the case of a south state is similar). The given piece of the equator divides the territory of $T$ into two regions: north and south. Moreover all vertices lie in the north region. Hence each edge intersects $Equ$ even times. Clearly, there is an edge a part of which forms a superfluous loop in the south hemisphere. Removing superfluous loops (see the admissible move 4)) gives rise to the case when no edge of $T$ intersects $Equ$. Therefore the original word along the piece of the equator lying inside the territory of $T$ was equal to the identity element in the free group. □

## 2.4 The proof of Proposition 1.

The proof of Proposition 1 will be divided into several steps (lemmas). In Step 1 we will show that the picture $P$ with the fixed equator $Equ$ can be divided into a finite number of uniform components. In Step 2 the uniform components will be transformed to states and edges-circles not belonging to the states. In Step 3 we will get rid of the edges-circles not belonging to the states. In Step 4 the states will be divided into $\sigma$-states. In all steps we will use only a finite number of admissible moves. Therefore we will get that $P$ is a $\sigma$-picture.

All pictures obtained from $P$ will be denoted by $P$ again for simplicity of notation.

**Step 1.** *Reducing $P$ to a picture containing only uniform components.*
In Step 1 we will use the following two admissible moves.
*Operation A: Transformations of reduced mixed components.*
Let the presentation $G =< A \mid R_1 \cup R_2 >$ be weakly $(R_1, R_2)$-separable. By $K$ denote a reduced mixed component.

Since $K$ is a reduced spherical picture, the condition of weak $(R_1, R_2)$-separability leads to the existence of a simple closed path $\gamma$ dividing the sphere into two parts so that

1) the both parts contains vertices;

2) $U = Lab(\gamma) \in [N_1, N_2]$;

3) if one part contains only north vertices then the other one contains only south vertices.

By the property 3) of $\gamma$, the following three cases are possible.

The first case: the path $\gamma$ divides $K$ into two parts one of which contains only south vertices, the other one contains only north vertices. The second case: $\gamma$ divides $K$ into two parts one of which contains only south vertices, the other one contains both south and north vertices. In these two cases we can assume that some segment $\psi$ of the path $\gamma$ lies on $Equ$. The complement of $\psi$ to $\gamma$ will be denoted by $\neg\psi$. One of the endpoints of $\psi$ will be denoted by $p$.

The third case: the path $\gamma$ divides $K$ into two parts each of which contains both south and north vertices. Consequently, $\gamma$ is intersected by $Equ$ and divided by $Equ$ into segments among which there are segments lying in the north hemisphere wholly. Fix one of them. By $\psi$ denote its connected part not intersecting $Equ$. By $p$ denote one of the endpoints of $\psi$. By $\neg\psi$ denote the complement of $\psi$ to $\gamma$.

In each of these three cases one can assume that all edges intersecting the path $\gamma$ intersect it in the segment $\psi$, because otherwise all edges intersecting $\neg\psi$ can be moved by isotopy ( the admissible move 1) ) to $\psi$ along the path $\gamma$ in the direction of the point $p$ starting successively at the nearest to $p$ edge.

In each of these three cases we select a map $M$ on $S^2$ with the following properties: the map $M$ contains the segment $\psi$ of $\gamma$ (coinciding with the part of $Equ$ in the first two cases) and parts of edges intersecting $\psi$: more precisely,

1) $\{y = 0\}$ is the segment $\psi$ of the path $\gamma$;

2) $\{x = x_1\}, ..., \{x = x_n\}$ correspond to the edges, which intersect the segment $\psi$;

3) in the first two cases the rectangle $\{-1 < x < 1, y < 0\}$ belongs to the south hemisphere and the rectangle $\{-1 < x < 1, y > 0\}$ belongs to the north hemisphere; in the third case the both rectangles belong to the north hemisphere. (An example for the third case is shown on Fig. 6 )

A new map $M'$ will be constructed as follows. Since $[N_1, N_2] \subset N_1 \cap N_2$, the word $U$ along $\psi$ belongs to the both groups $N_1$ and $N_2$. We construct planar pictures $P_1$ and $P_2$ with the boundary labels respectively $U$ and $U^{-1}$. Moreover in the first two cases $P_1$ is constructed over $< A \mid R_2 >$ (using south vertices) and $P_2$ is constructed over $< A \mid R_1 >$ (using north vertices); in the third case the both pictures $P_1$ and $P_2$ are constructed over $< A \mid R_1 >$ (using north vertices). Then these pictures are disposed on the new map $M'$ as follows:
1) $P_1$ lies in the rectangle $\{-1 < y < -1/2, -1 < x < 1\}$; and the edges corresponding to the boundary of $P_1$ start at $(x_1, -1)$,...,$(x_n, -1)$;
2) $P_2$ lies in the rectangle $\{1/2 < y < 1, -1 < x < 1\}$; and the edges corresponding to the boundary of $P_2$ start at $(x_1, 1)$,...,$(x_n, 1)$;
3) $\{y = 0\}$ is the segment $\psi$ of $\gamma$ (coinciding with the part of $Equ$ in the first two cases).

The old map $M$ is cut out from $P$ and replaced by the new one $M'$. (An example for the third case is shown on Fig. 7.) By this replacing of maps the equatorial label $W$ is changed by the commutator from $[N_1, N_2]$ in the first two cases and it is not changed in the third case. After such replacing all $R_1$-vertices lie in the north hemisphere and all $R_2$-vertices lie in the south one. Therefore this move of $P$ is admissible.

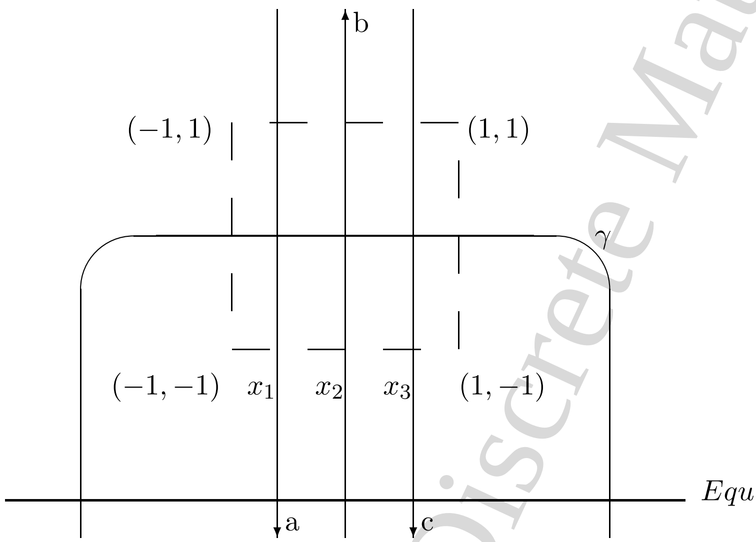



Fig. 6

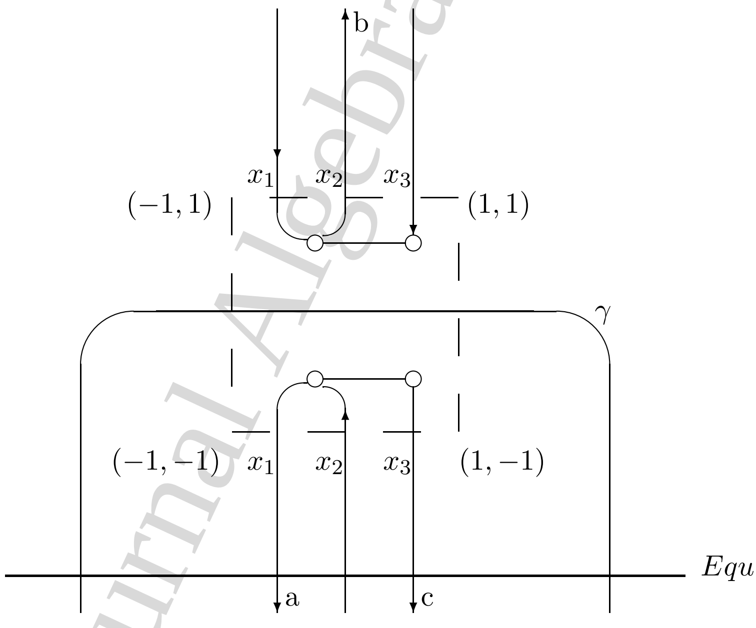



Fig. 7



By such replacing of maps the component $K$ falls into two components $K_1$ and $K_2$ separated from each other by the path $\gamma$. In the first case each of the components $K_i$ is uniform; in the second case one of the components (let it be denoted by $K_1$) is mixed, the other one ($K_2$) is south uniform; in the third case each of the components $K_i$ is mixed.

**Remark 3.** The components $K_i$ can be non-reduced. The number of south vertices in each mixed component $K_i$ ($K_1$ in the second case; $K_1$ and $K_2$ in the third case) is strictly less than the number of south vertices in the original component $K$, since by Operation $A$ we added only north vertices to obtain the mixed components.

*Operation B: Transformations of non-reduced mixed components.*
Let $K$ be a non-reduced component. Then there is a dipole in $K$, i.e., there are two vertices $V'$ and $V''$ satisfying the following conditions:

(i) there is a simple path $\psi$ joining some points $p_1$ and $p_2$, which lie on the boundaries $C_1$ and $C_2$ of these vertices, so that $Lab(\psi) = \mathbf{1}$;

(ii) ${Lab_{p_1}}^+(C_1) = {Lab_{p_2}}^-(C_2)$.

**Remark 4.** The both vertices of a dipole can be either north or south, since the sets $R_1$ and $R_2$ are mutually disjoint.

Evidently, it is possible to surround $V'$ and $V''$ by a simple closed path $\gamma\prime$ passing along $\psi$ such that $Lab(\gamma\prime) = \mathbf{1}$. Therefore if $\gamma\prime$ intersects edges then among them there are two edges intersecting $\gamma\prime$ successively and labelled by inverse letters. These edges can be removed from $\gamma\prime$ by bridge moves. It is easily seen that no edge intersects $\gamma\prime$ after a finite number of bridge moves.

Thus the component $K$ falls into two components $K_1$ and $K_2$ not connected with each other. The component $K_1$ contains only the dipole of $V'$ and $V''$ and the edges joining them together with some edges-circles. Hence $K_1$ is uniform. The component $K_2$ may be either uniform or mixed, either reduced or non-reduced.

**Remark 5.** Operation $B$ does not increase the number of vertices. Therefore the number of vertices in each $K_i$ is strictly less than the number of vertices in the original component $K$. In particular, the number of south vertices in $K_2$ is not more than the corresponding number in the original component $K$.

**Lemma 4.** *Let a presentation $G = < A \mid R_1 \cup R_2 >$ be weakly $(R_1, R_2)$-separable. Then a picture $P$ with a fixed equator $Equ$ falls into a finite number of uniform components by a finite number of Operations A and B (being admissible moves).*

*Proof.* Let $P$ consist of $n$ components $\{K_i\}$. In particular, $n$ may be equal to one. Using the admissible move 3) we have that all components $\{K_i\}$ intersect $Equ$.

To each mixed component $K_l$ of $P$ assign a number $s_l$ equal to the number of south vertices in $K_l$. Let $s$ be the maximum of $\{s_l\}$.

The proof of Lemma 4 will be by induction on $s$.

If $s = 0$, then all components $\{K_i\}$ are uniform, hence there is nothing to prove.

Let $s > 0$. In this case we transform each mixed component $K \in \{K_i\}$ containing $s$ south vertices as follows.

a) If $K$ is not reduced, then by Operation $B$, the component $K$ falls into two components $K'_1$ and $K'_2$, at least one of which (say $K'_1$) contains only a dipole and possibly edges-circles. The component $K'_1$ is uniform. The component $K'_2$ contains $s'_2$ south vertices, where $s'_2 \leq s$. If $K'_2$ is mixed and non-reduced, then Operation $B$ applies to it again. By Remark 5, after applying a finite number of Operations $B$, the component $K$ falls into uniform components $\{K'_i\}$ containing dipoles and a reduced component $K''$ containing $s''$ south vertices, where $s'' \leq s$. If $K''$ is mixed and $s'' = s$, then we transform it in the way described in the item b) below.

b) If $K$ is reduced, then it follows from weak $(R_1, R_2)$-separability that Operation $A$ can be applied to $K$. By Operation $A$, the component $K$ falls into two components $K''_1$ and $K''_2$. If any of $K''_j$ is mixed, then by Remark 3 it contains $s''_j$ south vertices, where $s''_j < s$ .

Thus after a finite number of Operations $A$ and $B$ the picture $P$ falls into a finite number of uniform components, since $s$ and $n$ are finite. □

**Step 2.** *Reducing a picture $P$ containing only uniform components to a picture containing only states and edges-circles not belonging to the states.*

Let $\{K_i\}$ be a subdivision of $P$ into uniform components.

Fix a point on $S^2$ lying neither on any edge and any vertex of $P$ nor on $Equ$. Assume that this point lies in the north hemisphere. We will call it *the north pole*.

Assume that a component $K \in \{K_i\}$ is such that some edges of $K$ form a simple closed loop $\eta$. The loop $\eta$ divides the sphere into two parts homeomorphic to disks: a disk containing the north pole will be called *exterior* and the other disk will be called *interior*.

If there is another component of $\{K_i\}$ lying in the interior disk with respect to the loop $\eta$, then this component is called *interior* for $K$. If the interior disk of $\eta$ contains at least one interior component for $K$ but it does not contain other closed loops of $K$ with the same property, then $\eta$ is called *minimal.*

The intersection of all exterior disks for all components contains the north pole and is called *the absolute exterior.*

In the step 2 we will use the following two admissible moves.

*Operation C: Uniting a component and its interior components belonging to the same hemisphere.*

Let $\eta$ be a minimal loop of a component $K$. Assume that all interior components being interior with respect to $\eta$ belong to the same hemisphere as the component $K$ does. Then we unite these interior components and $K$ in one component which will be denoted again by $K$.

*Operation D: Separating a component and its interior components belonging to the other hemisphere.*

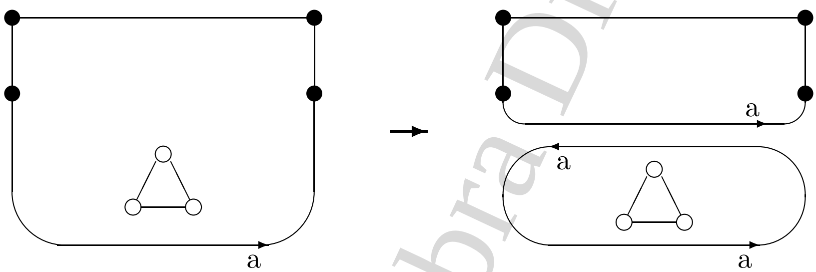



Fig. 8

Let $\eta$ be a minimal loop of a component $K$. If at least one of the interior components being interior with respect to $\eta$ does not belong to the same hemisphere as $K$ does, then we move all interior components of $K$ into the absolute exterior by bridge moves (see Fig. 8).

**Remark 6.** Operation $D$ strictly decreases the summary number of interior components in $P$ and adds a finite number of edges-circles to $P$.

**Lemma 5.** *Let $\{K_i\}$ be a finite subdivision of a picture $P$ into uniform components. Then a finite number of Operation $C$ and $D$ (being admissible moves) gives a subdivision of $P$ into states and edges-circles not belonging to the states.*

*Proof.* To each component $K_i$ of $P$ assign a number $s_i$ equal to the number of minimal loops in the component $K_i$. Let $s$ be the maximum of $\{s_i\}$.

For $s = 0$, there is nothing to prove.

Let $s > 0$. We transform each component $K \in \{K_i\}$ containing $s$ minimal loops as follows.

Let $\eta$ be one of $s$ minimal loops in $K$. If all interior components interior with respect to $\eta$ belong to the same hemisphere as $K$ does, then by Operation $C$ the loop $\eta$ ceases to be minimal, hence the number of minimal loops in $K$ becomes strictly less.

If there are some interior components interior with respect to $\eta$ such that they belong to the other hemisphere, then by Operation $D$ the number of minimal loops in $K$ becomes strictly less. Operation $D$ increases the number of minimal loops for none of components $\{K_i\}$, but it increases the number of edges-circles (see Remark 6).

Decreasing the maximum number $s$ of minimal loops by admissible moves, we obtain that each component lies in exterior disks for the other components, i.e., $P$ consists only of the states and edges-circles not belonging to the states. This proves Lemma 5. □

**Step 3.** *Getting rid of edges-circles.*
Assume that $P$ consists not only of states but of some edges-circles not belonging to the states. Each such edge-circle divides the sphere into two parts homeomorphic to disks: an exterior disk containing the north pole (see Step 2) and an interior disk not containing it.

Applying the admissible moves 4) and 3), we can assume that both interior and exterior disks of each edge-circle contain states.

An edge-circle $C$ is called *minimal*, if the interior disk of $C$ does not contain the other edges-circles.

In Step 3 we will use the following two admissible moves.
*Operation E: Uniting an edge-circle and its interior states belonging to the same hemisphere.*
Let $C$ be a minimal edge-circle such that its interior disk contains states belonging to the same hemisphere only. Then we unite $C$ and the states interior to $C$ in one state of the same hemisphere.

**Remark 7.** Operation $E$ decreases the number of edges-circles in $P$. An edge-circle, which was not minimal before Operation $E$, can become minimal after Operation $E$.

*Operation F: Cutting an edge-circle if its interior states belong to the different hemispheres.*
Let $C$ be a minimal edge-circle such that its interior disk contains states belonging to the different hemispheres. By bridge moves, $C$ can be cut into some edges-circles $\{C_i\}$ so that for each $C_i$ its interior states belong only to the same hemisphere (see Fig. 9).

**Remark 8.** Operation $F$ is a composition of a finite number of bridge moves. New edges-circles $\{C_i\}$ are minimal again. An edge-circle, which was not minimal before Operation $F$, remains not minimal after Operation $F$.

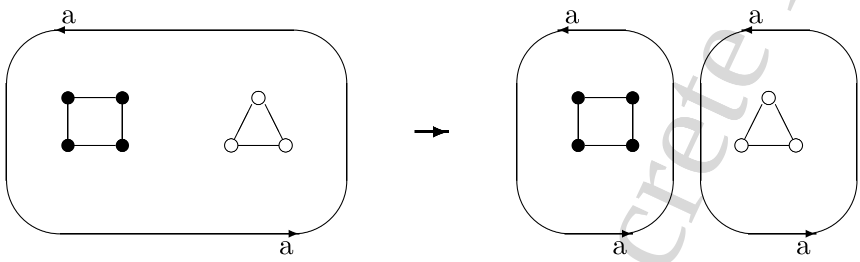



Fig. 9

**Lemma 6.** *Assume that there are a finite subdivision of $P$ into states $\{T_i\}$ and edges-circles not belonging to the states. Then after a finite number of bridge moves, $P$ will consist of states only.*

*Proof.* Assume that $P$ contains $s$ edge-circles not belonging to the states.

The proof is by induction on $s$. For $s = 0$, there is nothing to prove. Let $s > 0$.

Let $m$ be the number of minimal edges-circles ($m > 0$) and $n$ be the number of edges-circles not being minimal among all $s$ edges-circles ($n = s - m$).

If there are states from the different hemispheres interior to any minimal edge-circle $C$, then by Operation $F$ the edge-circle $C$ can be cut into a finite number of minimal edges-circles $\{C_i\}$ such that for each $C_i$ its interior states belong only to the same hemisphere. Note that Operation $F$ does not change the number $n$ of not minimal edges-circles.

Thus we can obtain that for each minimal edge-circle its interior states are from the same hemisphere. By Operation $E$, we can unite each minimal edges-circle and its interior states in one state. In addition, the number of the edge-circles not belonging to the states becomes at most $n$, where $n < s$. □

**Step 4.** *Reducing a picture $P$ containing only states to a picture containing only $\sigma$-states.*
Let $P$ contain only states.

On this step we will apply the following admissible move.
*Operation $G$. Crushing a state into $\sigma$-states.*
Let $\mathcal{R}$ be an irregular north region of a north state $T$ ( the move of an irregular south region of a south state is similar). The boundary of the

region $\mathcal{R}$ contains at least two connected pieces of the equator. We fix any of them and denote it by $I$. The following transformations are performed near $I$. A map $M$ is chosen in $\mathcal{R}$ so that it contains $I$ together with edges intersecting $Equ$ at points of $I$: more precisely, $\{y = -1\}$ coincides with $I$; $\{x = x_1\}, ..., \{x = x_n\}$ correspond to the edges intersecting $Equ$ in $I$; $\{x = -1\}$ and $\{x = 1\}$ coincide with parts of the boundary of $T$. ( See an example on Fig. 10. )

The construction of a new map $M'$ is similar to Step 1. Namely, by Lemma 2, $U = Lab(I) \in N_1$. Two planar pictures $P_1$ and $P_2$ with the boundary labels respectively $U$ and $U^{-1}$ are constructed and disposed on $M'$ as follows:
1) $P_1$ lies in the rectangle $\{-1 < y < -1/2, -1 < x < 1\}$ and all edges intersecting the boundary of $P_1$ start at $(x_1, -1)$,...,$(x_n, -1)$;
2) $P_2$ lies in the rectangle $\{1/2 < y < 1, -1 < x < 1\}$ and all edges intersecting the boundary of $P_2$ start at $(x_1, 1)$,...,$(x_n, 1)$;
3) the territory of $T$ divides into parts $\{y \leq -1/2\}$ and $\{y \geq 1/2\}$;
4) $\{y = -1\}$ is the considered piece $I$ of the equator.

The map $M$ is cut out from $P$ and replaced by the new map $M'$ (see Fig.11 ).

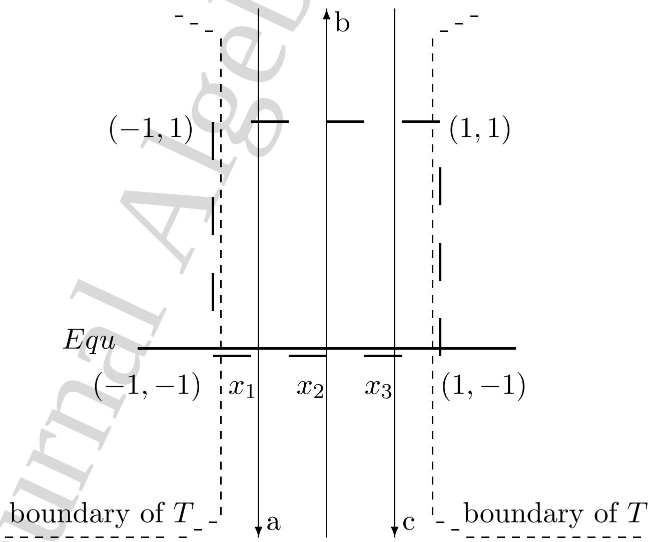



Fig. 10

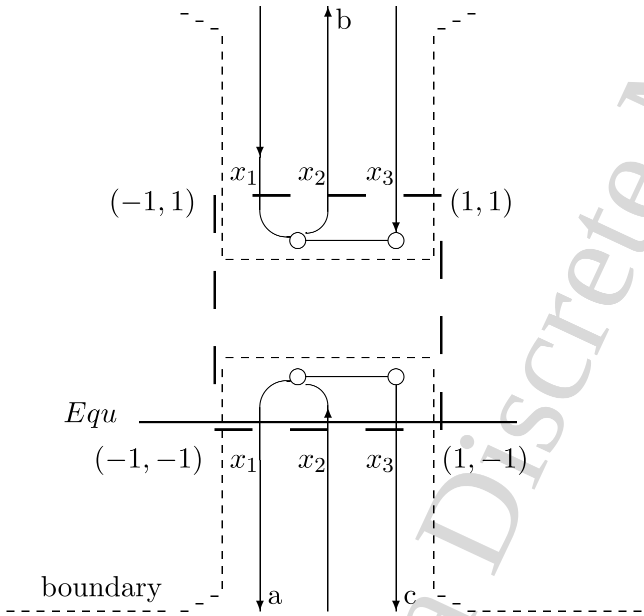



Fig. 11

This move does not change the equatorial label $W$, hence it is admissible.

**Remark 9.** By this move the number of states increases by one and the irregular region $\mathcal{R}$ divides into two parts: one is a regular region and the other one is an irregular region such that the number of pieces of the equator belonging to its boundary is strictly less than the number of pieces belonging to the boundary of the original $\mathcal{R}$.

**Lemma 7.** *Let a picture $P$ contain only a finite number of states. Then $P$ can be reduced to a picture containing only $\sigma$-states by a finite number of Operations $G$ (being admissible moves).*

*Proof.* If all north regions of all north states and all south regions of all south states are regular, then there is nothing to prove.

Otherwise, to each irregular region $\mathcal{R}_i$ assign the number $s_i$ of pieces of the equator lying on the boundary of $\mathcal{R}_i$. Let $s$ be the maximum of $\{s_i\}$ $(s > 1)$.

Operation $G$ applied to each irregular region decreases the number $s$ by one. Therefore all states become $\sigma$-states after a finite number of Operations $G$. □

Thus Proposition 1 follows from Lemmas 4 - 7. □

## 2.5 Proof of Proposition 2.

By admissible moves, $P$ is reduced to a picture containing only $\sigma$-states.

We assume that $P$ consists only of two $\sigma$-states: a north $\sigma$-state $T_1$ and a south $\sigma$-state $T_2$, since we can always unite all north $\sigma$-states with each other and all south $\sigma$-states with each other (see the admissible move 5)).

$Equ$ can be divided into arcs so that each of these arcs is intersected by one $\sigma$-state only. We fix the direction of moving along $Equ$ (let it be from the west to the east) and renumber these arcs $J_1, J_2, ..., J_k$ successively. We may assume that an arc $J_i$ is intersected by the north $\sigma$-state for even $i$ and by the south $\sigma$-state for odd $i$. An arc will be called south (respectively, north), if it is intersected only by the south $\sigma$-state ( respectively only by the north $\sigma$-state ). According to Remark 2 (see the admissible move 5)), each arc $J_i$ contains precisely one piece $I_i$ of the equator belonging to the territory of any $\sigma$-state (this piece $I_i$ is a part of the boundary of a regular region of the corresponding $\sigma$-state)

The proof of Proposition 2 is by induction on $k$ equal to the number of arcs.

If $k \leq 2$, then Proposition 2 follows from Lemma 3.

Let $k > 2$. Then we do the following admissible move.

We consider the second north arc $J_2$ and the third south arc $J_3$. By Lemma 2, a word $w_1$ along the piece $I_2$ of the north arc $J_2$ belongs to $N_1$ and a word $w_2$ along the piece $I_3$ of the south arc $J_3$ belongs to $N_2$.

Construct a map $M$ containing a subpicture with the part $\{y = 0\}$ of the equator corresponding to the word $w_2 w_1 {w_2}^{-1} {w_1}^{-1}$. Cut out a small map $M_s$ between $I_1$ and $I_2$ such that $M_s$ contains nothing but a small part $\{y = 0\}$ of $Equ$. Paste $M$ in place of $M_s$ (see the admissible move 6)). This adds two new $\sigma$-states: north one $T_1^{comm}$ and south one $T_2^{comm}$. (See an example on Fig. 12) This move corresponds to the insertion of the commutator $w_2 w_1 {w_2}^{-1} {w_1}^{-1}$ of $w_1 \in N_1$ and $w_2 \in N_2$ in the word $W = ...w_1 w_2...$, and replacing it by $W' = ...(w_2 w_1 {w_2}^{-1} {w_1}^{-1}) w_1 w_2....$

By the admissible move 5), we unite the north $\sigma$-states $T_1$ and $T_1^{comm}$ in one $\sigma$-state $T_1^n$, after that we use Remark 2. Let $I_2^n$ be a piece of the equator in the north $\sigma$-state $T_1^n$ containing the piece of the equator $I_2$ of the original $\sigma$-state $T_1$ and the piece of the equator $I_2'$ of the original $\sigma$-state $T_1^{comm}$ ( the notation are from the description of the admissible move 6)). Since $Lab(I_2^n) = \mathbf{1}$, after a finite number of bridge moves, no edge of $T_1^n$ intersects $I_2^n$ (see the admissible move 2)).

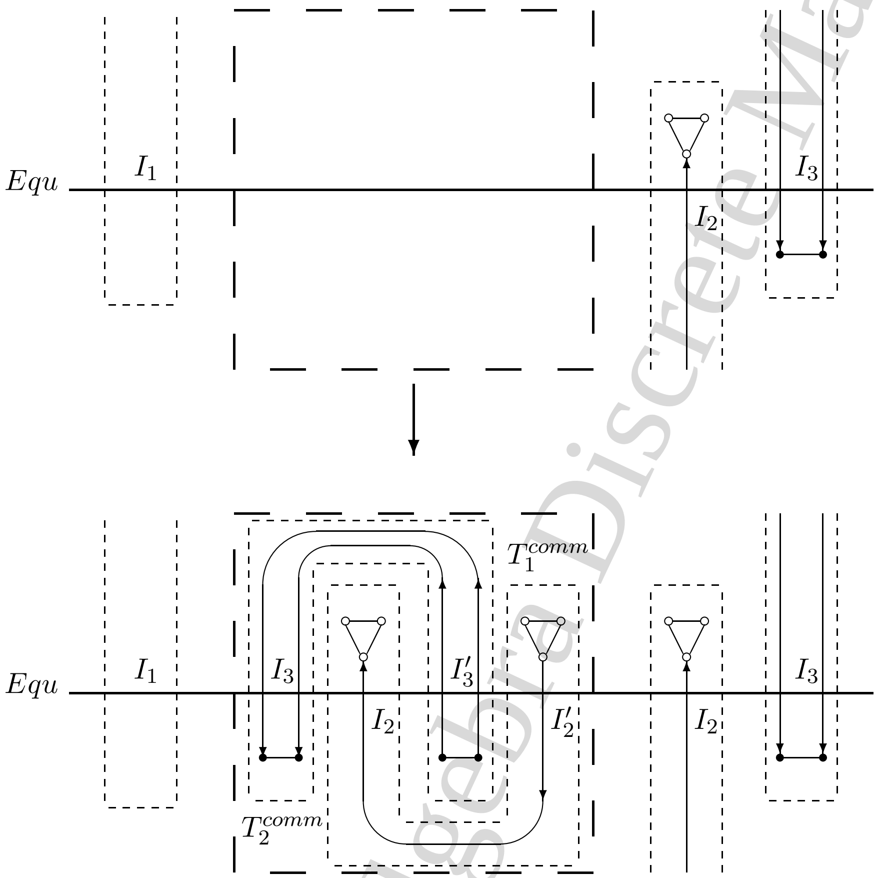



Fig. 12

If a regular region (the boundary of which contains the piece $I_2^n$) contains a subpicture not intersecting $Equ$, then this subpicture is removed (see the admissible move 3)). Deformation of the territory of $T_1^n$ gives that the territory of $T_1^n$ does not intersect $Equ$ by the piece $I_2^n$.

By the admissible move 5), we unite the south $\sigma$-states $T_2$ and $T_2^{comm}$ in one $\sigma$-state $T_2^n$. By Remark 2, the piece $I_3^n$ of the south $\sigma$-state $T_2^n$ contains the piece $I_3$ of the $\sigma$-state $T_2$ and the piece $I_3'$ of the $\sigma$-state $T_2^{comm}$. By similar transformations as above, we obtain that the $\sigma$-state $T_2^n$ does not intersect $Equ$ by the piece $I_3^n$.

Note that as a result of the transformations described above the equatorial label $W' = ...(w_2 w_1 {w_2}^{-1} {w_1}^{-1}) w_1 w_2...$ is reduced to the form $W' = ...w_2 w_1....$

Thus again $P$ consists of the two $\sigma$-states $T_1^n$ and $T_2^n$, which intersect $Equ$ by the same pieces as the $\sigma$-states $T_1$ and $T_2$ did, but $T_1^n$ and $T_2^n$

intersect $Equ$ in other order. More precisely, the south pieces of the equator $I_1$ and $I_2$ turn out to be side by side (similarly the north pieces $I_2$ and $I_4$ turn out to be side by side). Therefore by Remark 2, the number of arcs in the subdivision of $Equ$ becomes strictly less and the number of states and vertices does not increase.

By induction, we get a subdivision of $Equ$ into two arcs, i.e., the case when the equatorial label $W$ is equal to the identity element in the free group, which is the desired conclusion. □

## 3. Some corollaries in the case of free products.

Let $G_1 =< X \mid S_1 >$ and $G_2 =< Y \mid S_2 >$ represent some groups and $G = G_1 * G_2$ be a free product of them. We will assume that the sets $S_1$ and $S_2$ are symmetrized. We consider the normal closure $N_1$ of a set of elements $R_1 = \{u_i\}$ and the normal closure $N_2$ of a set of elements $R_2 = \{v_j\}$ in $G$. We assume that $R_1$ and $R_2$ are mutually disjoint and symmetrized.

**Assertion 6.** *Let a presentation $< X \cup Y \mid S_1 \cup S_2 \cup R_1 \cup R_2 >$ be weakly $(T_1, T_2)$-separable, where $T_1$ and $T_2$ are defined in one of the following ways:*
*(i) $T_1 = R_1 \cup S_1$, $T_2 = R_2 \cup S_2$;*
*(ii) $T_1 = R_1 \cup S_1 \cup S_2$, $T_2 = R_2$.*
*Then $N_1 \cap N_2 = [N_1, N_2]$.*

*Proof.* Let us show that Assertion 6 follows from the similar statement in the case of free groups.

Let $N_1{}' =< T_1 >^F$ and $N_2{}' =< T_2 >^F$ denote the normal closures of the sets of words respectively $T_1$ and $T_2$ in the free group $F = F(X) * F(Y)$. By Theorem 1, weak $(T_1, T_2)$-separability leads to

$$N_1{}' \cap N_2{}' = [N_1{}', N_2{}']. \qquad (*)$$

Consider the canonical homomorphism $\psi : F \longrightarrow G = F / < S_1 \cup S_2 >^F$. Let us show that the following equalities hold:
1) $\psi(N_1{}') = N_1$;
2) $\psi(N_2{}') = N_2$;
3) $\psi([N_1{}', N_2{}']) = [N_1, N_2]$;
4) $\psi(N_1{}' \cap N_2{}') = N_1 \cap N_2$.

The first two equalities are immediate. It follows from them that the third equality and the inclusion $\psi(N_1{}' \cap N_2{}') \subset N_1 \cap N_2$ hold. Therefore it remains to prove only the inclusion $N_1 \cap N_2 \subset \psi(N_1{}' \cap N_2{}')$.

Let $c \in N_1 \cap N_2$ be an arbitrary element and $\bar{n}_1 \in N_1{}'$ and $\bar{n}_2 \in N_2{}'$ be some of its preimages. We have $\bar{n}_1 = n * \bar{n}_2$, where $n \in Ker\,\psi$. If $n$

is equal to the identity element in the free group, then there is nothing to prove. Otherwise, we will look for such preimages of $c$ that $n$ will be equal to the identity element in the free group.

In the case (ii) the element $n^{-1} * \bar{n}_1$ belongs to $N_1{}'$. Thus in this case preimages of $c$ can be chosen so that $n$ will be equal to the identity element in the free group.

Consider the case (i). The element $n$ can be represented as a product of elements $s_{1,i} \in< S_1 >^F$ and $s_{2,j} \in< S_2 >^F$. There are two possibilities: the first factor belongs either to $< S_1 >^F$ or to $< S_2 >^F$. We will show that it is possible to replace $\bar{n}_1$ by $\tilde{n}_1$ and $\bar{n}_2$ by $\tilde{n}_2$ in each of these cases so that $\psi(\bar{n}_1) = \psi(\tilde{n}_1)$, $\psi(\bar{n}_2) = \psi(\tilde{n}_2)$, and the number of factors in the factorization of the product $\tilde{n}_1 * \tilde{n}_2^{-1}$ will be strictly less than one in the factorization of $\bar{n}_1 * \bar{n}_2^{-1}$.

If $s_{2,1} \in< S_2 >^F$ is the first factor of $n$, that is $n = \bar{n}_1 * \bar{n}_2^{-1} = s_{2,1} * g$, then we conjugate it by the element $s_{2,1} \in Ker\psi$: $(s_{2,1}{}^{-1} * \bar{n}_1 * s_{2,1}) * (s_{2,1}{}^{-1} * \bar{n}_2^{-1} * s_{2,1}) = g * s_{2,1}$ and denote the conjugated elements $\bar{n}_1$ and $\bar{n}_2$ by $\bar{n}_1$ and $\bar{n}_2$ again. We have that the number of factors in the factorization of $n$ does not increase.

Therefore we can assume that $s_{1,1} \in< S_1 >^F \subset \ker\psi$ is the first factor of $n$. We multiply $n$ by $s_{1,1}{}^{-1}$ on the left and denote the product $s_{1,1}{}^{-1} * \bar{n}_1$ again by $\bar{n}_1$. Note that the new $\bar{n}_1$ also belongs to $N_1{}'$ and the number of factors in the factorization of $n$ decreases.

If the last factor of $n$ is $s_{2,k} \in< S_2 >^F \subset \ker\psi$, then we multiply $n$ by $s_{2,k}{}^{-1}$ on the right and denote the product $\bar{n}_2 * s_{2,k}{}^{-1}$ again by $\bar{n}_2$. We have that the new $\bar{n}_2 \in N_2{}'$ and the number of factors in the factorization of $n$ decreases.

Repeating as above, we can reduce $n$ to the identity element in the free group.

It follows now from (3), (4), $(*)$ that $N_1 \cap N_2 = [N_1, N_2]$. □

**Corollary 6.** *If $R_1$ belongs to $G_1$, and $R_2$ belongs to $G_2$, then $N_1 \cap N_2 = [N_1, N_2]$.*

*Proof.* It follows from Corollary 5, Theorem 1, and (i) of Assertion 6. □

**Corollary 7.** *Under the notation of Assertion 6, let a presentation $< X \cup Y \mid T_1 \cup T_2 >$ satisfy one of the following conditions:*
*(i) strict $(T_1, T_2)$-separability;*
*(ii)$(T_1, T_2)$-asphericity;*
*(iii) asphericity.*

*Then $N_1 \cap N_2 = [N_1, N_2]$.*

*Proof.* It follows from Corollary 3, Theorem 1, and Assertion 6. □

**Corollary 8.** *Let a set* $\{S_1, S_2, R_1, R_2\}$ *satisfy one of the following small cancellation conditions: either* $C(6)$*, or* $C(4)\&T(4)$*, or* $C(3)\&T(6)$*.*

*Then* $N_1 \cap N_2 = [N_1, N_2]$*.*

*Proof.* It follows from Corollary 4, Theorem 1, and Assertion 6. □

## References


[1] A.Yu. Ol'shanskij. *Geometry of defining relations in groups*, Moscow, "Nauka", 1989 (in Russian).

[2] Ol'shanskij A.Yu. *Geometry of defining relations in groups*, Mathematics and Its Applications. Soviet Series, 70, Dordrecht atc.: Kluwer Academic Publishers, 1991.

[3] R.S.Lindon, P.E.Schupp. *Combinatorial group theory,* Springer-Verlag, Berlin - Heidelberg - NewYork, 1977.

[4] W.A. Bogley, S.J. Pride. *Aspherical relative presentations,* Proc. Edinburgh Math. Soc., vol. **35** ( ser. II ), part 1, 1-40, 1992.


Contact information


**O. V. Kulikova**

Department of Mechanics and Mathematics
Moscow State University,Vorobievy Gory 1
119992 Moscow, Russia
*E-Mail:* olga.kulikova@mail.ru


Received by the editors: 09.12.2002.

Algebra and Discrete Mathematics RESEARCH ARTICLE
Number 2. **(2003).** pp. 68–76


# Direct decompositions of artinian modules related to formations of groups

**Leonid A. Kurdachenko and Javier Otal**

Abstract. We survey direct decompositions of artinian modules over group rings into two summands where all the chief factors of the first are $\mathfrak{X}$–central and all the chief factors of the other is $\mathfrak{X}$–eccentric, where $\mathfrak{X}$ is a certain formation of finite groups.

## 1. Introduction

Artinian modules are one of oldest objects of study in Algebra. The study of artinian modules over group rings began in the second half of the past century and was mainly stimulated by questions of the theory of (soluble) groups. Although the underlying group of the group ring considered can have a very complicated structure, the study of the action of the group on the module play a major role and provides in a natural way the existence of some direct decomposition of modules, which gives a good information for the group itself. Probably, a celebrated result due to H. Fitting, known as Fitting's lemma, is one of the first result on which one of these direct decompositions appear. We formulate it in the following form.

**Theorem** (Fitting) *Let $A$ be an $RG$–module of finite composition length, where $R$ is a ring and $G$ is a finite nilpotent group. Then $A = C \oplus E$, where the $RG$–chief factors $U/V$ of $E$ (respectively of $E$) satisfy $G = C_G(U/V)$ (respectively, $G \neq C_C(U/V)$). In particular, $A = C_A(G) \oplus A(\omega RG)$ provided $\Pi(G) \cap \Pi(A) = \emptyset$.*

---

*Supported by Proyecto BFM2001-2452 of CICYT (Spain) and Proyecto 100/2001 of Gobierno de Aragón (Spain)*

**2000 Mathematics Subject Classification:** *20C07, 20D10, 20F24.*

**Key words and phrases:** *decompositions, artinian modules, formation of groups.*

This raises the subordinated questions both of finding out complements of the upper $RG$–hypercenter and that of studying certain extensions of modules, which are near to modules of finite composition length. These problem are found very important applications in the study both of groups and modules with finiteness conditions. Moreover they are also connected with the question of the existence of complements for some residuals in groups.

Now we formulate the necessary concepts in their more general form.

A class $\mathfrak{X}$ of groups is said to be *a formation* if and only if it satisfies the following conditions:

(F1) If $G \in \mathfrak{X}$ and $H$ is a normal subgroup of $G$, then $G/H \in \mathfrak{X}$.

(F2) If $H_1$ and $H_2$ are normal subgroups of $G$ such that $G/H_1$ and $G/H_2 \in \mathfrak{X}$, then $G/H_1 \cap H_2\mathfrak{X}$.

Let $R$ be a ring, $G$ a group, $\mathfrak{X}$ a class of groups and $A$ an $RG$–module. As suggested above, if $B \leq C$ are $RG$–submodules of $A$, the factor $C/B$ is said to be $\mathfrak{X}$*–central* (respectively, $\mathfrak{X}$*–eccentric*) if $G/C_G(C/B) \in \mathfrak{X}$ (respectively, $G/C_G(C/B) \notin \mathfrak{X}$). To rule out these factors, we similarly define

$$\mathfrak{X}C_{RG}(A) = \{a \in A \mid G/C_G(aRG) \in \mathfrak{X}\}.$$

As $\mathfrak{X}$ is a formation of groups, then $\mathfrak{X}C_{RG}(A)$ is an $RG$–submodule of $A$ called the $\mathfrak{X}$*–center of* $A$ (more precisely, *the* $\mathfrak{X} - RG$*–center of* $A$). Proceeding in a similar way to that we did for groups, we construct the upper $\mathfrak{X}$–central series of the module $A$ as

$$\{0\} = A_0 \leq A_1 \leq \cdots A_\alpha \leq A_{\alpha+1} \leq \cdots \leq A_\gamma$$

where $A_1 = \mathfrak{X}C_{RG}(A)$, $A_{\alpha+1}/A_\alpha = \mathfrak{X}C_{RG}(A/A_\alpha)$, for all ordinals $\alpha < \gamma$ and $\mathfrak{X}C_{RG}(A/A_\gamma) = \{0\}$. The last term $A_\gamma = HZ_{\mathfrak{X}-RG}(A)$ of this series is called *the upper* $\mathfrak{X}$*–hypercenter of* $A$ (or *the* $\mathfrak{X} - RG$*–hypercenter*) and the other terms $A_\alpha$, *the* $\mathfrak{X}$*–hypercenters of* $A$. If $A = A_\gamma$, then $A$ is said to be $\mathfrak{X}$*–hypercentral* ($\mathfrak{X}$*–nilpotent*, if $\gamma$ is finite). If $\mathfrak{X} = \mathfrak{I}$ and $\mathfrak{X} = \mathfrak{F}$, we have the $RG$–center $\zeta_{RG}(A)$ of $A$, the upper $RG$–hypercenter $\zeta^{\infty}_{RG}(A)$ of $A$, the $FC$–center $FC_{RG}(A)$ of $A$ and the upper $FC$–hypercenter $FC^{\infty}_{RG}(A)$ of $A$.

On the other hand, an $RG$–submodule $C$ of $A$ is said to be $\mathfrak{X} - RG$*–hypereccentric* if it has an ascending series

$$\{0\} = C_0 \leq C_1 \leq \cdots C_\gamma \leq C_{\alpha+1} \leq \cdots C_\gamma = C$$

of $RG$–submodules of $A$ such that each factor $C_{\alpha+1}/C_\alpha$ is an $\mathfrak{X}$–eccentric simple $RG$–module, for every $\alpha < \gamma$.

Given a formation $\mathfrak{X}$ of groups, we say that *the $RG$–module $A$ has the $\mathfrak{X}$–decomposition* or, more precisely, *the $\mathfrak{X} - RG$–decomposition*) if the following equality holds

$$A = HZ_{\mathfrak{X}-RG}(A) \bigoplus HE_{\mathfrak{X}-RG}(A).$$

It is worth to mentioning that $HE_{\mathfrak{X}-RG}(A)$ is the unique maximal $\mathfrak{X} - RG$–hypereccentric $RG$–submodule of $A$. For, let $B$ an $\mathfrak{X} - RG-$hypereccentric $RG$–submodule of $A$, $E = HE_{\mathfrak{X}-RG}(A)$. If $(B+E)/E$ is non-zero, it includes a non-zero simple $RG$–submodule $U/E$. Since $(B + E)/E \cong B/(B \cap E)$, $U/E$ is $RG$–isomorphic to some simple $RG$–factor of $B$ and it follows that $G/C_G(U/E) \notin \mathfrak{X}$. On the other hand, $(B + E)/E \leq A/E \cong HZ_{\mathfrak{X}-RG}(A)$, that is $G/C_G(U/E) \in \mathfrak{X}$. This contradiction shows that $B \leq E$. Hence $HE_{\mathfrak{X}-RG}(A)$ includes every $\mathfrak{X} - RG-$hypereccentric $RG$–submodule and, in particular, it is unique, as claimed.

If $\mathfrak{X} = \mathfrak{I}$, the decomposition is simply called *the $\mathfrak{Z}$–decomposition* whereas if $\mathfrak{X} = \mathfrak{F}$, we called it *the $\mathfrak{F}$–decomposition*.

Since we are discussing the existence of the $\mathfrak{X}$–decomposition for modules over $\mathfrak{X}C$–hypercentral groups, we will need to define concepts in groups that are similar to the above ones.

If $G$ is a group and $x \in G$, we put $x^G = \{g^{-1}xg \mid g \in G\}$; clearly, $C_G(x^G)$ is normal in $G$. Let now $\mathfrak{X}$ be a class of groups, and define *the $\mathfrak{X}$–center* of $G$ as

$$\mathfrak{X}C(G) = \{x \in G \mid G/C_G(x^G) \in \mathfrak{X}\}.$$

If $\mathfrak{X}$ is a formation of groups, then $\mathfrak{X}C(G)$ is a characteristic subgroup of $G$ and $G$ is said to be *an $\mathfrak{X}C$–group* if the equality $G = \mathcal{X}C(G)$ holds. If $\mathfrak{X} = \mathfrak{I}$ is the class of all identity groups, then $\mathfrak{X}C(G) = \zeta(G)$ is the ordinary center of $G$ whereas, if $\mathfrak{X} = \mathfrak{F}$ is the class of all finite groups, then $\mathfrak{X}C(G) = FC(G)$ is precisely the $FC$–center of $G$. From this subgroup, we may construct *the upper $\mathfrak{X}$–central series* of $G$ as

$$\langle 1 \rangle = C_0 \leq C_1 \leq \cdots \leq C_\alpha \leq C_{\alpha+1} \leq \cdots C_\gamma,$$

where $C_1 = \mathfrak{X}C(G)$, $C_{\alpha+1}/C_\alpha = \mathfrak{X}C(G/C_\alpha)$, for an $\alpha < \gamma$, and $\mathfrak{X}C(G/C_\gamma) = \langle 1 \rangle$. The last term $C_\gamma$ of this series is called *the upper $\mathfrak{X}$–hypercenter of $G$* and denoted by $HZ_{\mathfrak{X}}(G)$. If $G = C_\gamma$, then $G$ is said to be *$\mathfrak{X}$–hypercentral* and, if $\gamma$ is finite, *$\mathfrak{X}$–nilpotent*. Once again, if $\mathfrak{X} = \mathfrak{I}$ or $\mathfrak{X} = \mathfrak{F}$, $HZ_{\mathfrak{X}}(G) = \zeta^\infty(G)$ or $= FC^\infty(G)$ are the upper hypercenter or upper $FC$–hypercenter of $G$.

## 2. Some results

The first result on the existence of the$\mathfrak{Z}$–decomposition for infinite modules was obtained by Hartley and Tomkinson [3].

**Theorem 1.** *Let $G$ be a locally nilpotent group and let $A$ be a $\mathbb{Z}G$–module. If $A$ is $\mathbb{Z}$–periodic and the $p$–component of $A$ has finite special rank for every prime $p$, then $A$ has the $\mathfrak{Z}$–decomposition.*

In this case, $G/C_G(A)$ is actually hypercentral and every $p$–component of $A$ is an artinian $\mathbb{Z}$–module. Artinian modules are the most classical extension of modules, having finite composition series. Thus, we naturally come to the question of the existence of the $\mathfrak{Z}$–decomposition for artinian modules. The question was solved by Zaitsev [9], which showed the following result.

**Theorem 2.** *Let $G$ be a hypercentral group and let $A$ be a $\mathbb{Z}G$–module. If $A$ is an artinian $\mathbb{Z}G$–module, then $A$ has the $\mathfrak{Z}$–decomposition.*

Note that this result can be easily extended to artinian $DG$–module, where $D$ is a Dedekind domain and $G$ is a hypercentral group.

The following natural step was the consideration of the question of the existence of the $\mathfrak{F}$–decomposition for artinian modules. We notice first that in this study the underlying group considered has to be $FC$–hypercentral instead of hypercentral. A first result in this direction was obtained by Zaitsev [10, 11] as follows.

**Theorem 3.** *Let $G$ be a hyperfinite locally soluble group and $A$ a $\mathbb{Z}G$–module. If $A$ is an artinian $\mathbb{Z}G$–module, then $A$ has the $\mathfrak{F}$–decomposition.*

**Theorem 4.** *Let $G$ be an $FC$–hypercentral group and let $A$ be a $\mathbb{Z}G$–module. If $A$ has finite composition $\mathbb{Z}G$–series, then $A$ has the $\mathfrak{F}$–decomposition.*

The next result is due to Duan [1], who showed a partial solution in considering a special type of $FC$–hypercentral groups.

**Theorem 5.** *Let $G$ be a locally soluble group having an ascending series of normal subgroups, every factor of which is finite or cyclic and let $A$ be a $\mathbb{Z}G$–module. If $A$ is an artinian $\mathbb{Z}G$–module, then $A$ has the $\mathfrak{F}$–decomposition.*

The final solution for the formation $\mathfrak{F}$ was obtained by Kurdachenko, Petrenko and Subbotin [5].

**Theorem 6.** *Let $G$ be a locally soluble $FC$–hypercentral group, $D$ a Dedekind domain and $A$ a $DG$–module. If $A$ is an artinian $DG$–module, then $A$ has the $\mathfrak{F}$–decomposition.*

In Module Theory, modules having a finite composition series are the algebraic objects more analogous to finite groups. For these modules, we can obtain the following result.

**Theorem 7.** *Let $\mathfrak{X}$ be a formation of finite groups, $G$ an $\mathfrak{X}$–hypercentral group, $D$ a Dedekind domain and $A$ a $DG$–module. If $A$ has a finite composition $DG$–series, then $A$ has the $\mathfrak{X}$–decomposition.*

Since $G$ is $FC$–hypercentral, $A$ has the $\mathfrak{F}$–decomposition, that is $A = B \oplus C$, where $B = HZ_{\mathfrak{F}-DG}(A)$ and $C = HE_{\mathfrak{F}-DG}(A)$. Since $B$ has a finite composition series every factor of which is $\mathfrak{F}$–central, $G/C_G(B)$ is finite. By [7], $B$ has the $\mathfrak{X}$–decomposition and then $A$ has the $\mathfrak{X}$–decomposition too.

Note that in this result, $\mathfrak{X}$ is an arbitrary formation of finite groups.

The question of the existence of the $\mathfrak{F}$–decomposition has been an important partial case of the general case and its solution has allowed to obtain the solution for many important formations $\mathfrak{X}$. At this point, it is needed to split the general study into two cases according to $\mathfrak{F} \leq \mathfrak{X}$ or $\mathfrak{X}$ is a proper formation of finite groups.

A formation $\mathfrak{X}$ is said to be *overfinite* if it satisfies the following conditions:

(i) if $G \in \mathfrak{X}$ and $H$ is a normal subgroup of $G$ of finite index, then $H \in \mathfrak{X}$.

(ii) if $G$ is a group, $H$ is a normal subgroup of finite index of $G$ and $H \in \mathfrak{X}$, then $G \in \mathfrak{X}$.

(iii) $\mathfrak{I} \leq \mathfrak{X}$.

Clearly, an overfinite formation always contains $\mathfrak{F}$. The most important examples of these formations are polycyclic groups, Chernikov groups, soluble minimax groups, soluble groups of finite special rank and soluble groups of finite section rank For locally soluble $FC$–hypercentral groups, the existence of the $\mathfrak{X}$–decomposition for an overfinite formation $\mathfrak{X}$ in an artinian $DG$–module $A$ was also showed by Kurdachenko, Petrenko and Subbotin in [6], who proved the next result.

**Theorem 8.** *Let $D$ be a Dedekind domain, $G$ a locally soluble $FC$–hypercentral group and $A$ an artinian $DG$–module. If $\mathfrak{X}$ is an overfinite formation of groups, then $A$ has the $\mathfrak{X}$–decomposition.*

**Corollary 1.** *If $G$ is a locable soluble $FC$–hypercentral group and $D$ is a Dedekind domain, then an artinian $DG$–module $A$ has the $\mathfrak{X}$-decomposition for the following formations $\mathfrak{X}$:*

*(1)* $\mathfrak{F}$*, formation of all finite groups;*
*(2)* $\mathfrak{P}$*, formation of all polycyclic groups;*
*(3)* $\mathfrak{C}$*, formation of all Chernikov groups;*
*(4)* $\mathfrak{S}_2$*, formation of all soluble minimax groups;*
*(5)* $\mathfrak{S}^{\wedge}$*, formation of all soluble groups of finite special rank; and*
*(6)* $\mathfrak{S}_0$*, formation of all soluble groups of finite section rank.*

Since every formation $\mathfrak{X}$ includes $\mathfrak{F}$, every $FC$–hypercentral group is likewise $\mathfrak{X}$–hypercental. Thus, the next natural step is to consider that the underlyin group is $\mathfrak{X}$–hypercentral and study artinian $DG$–modules.

A formation $\mathfrak{X}$ of finite groups is said to be *infinitely hereditary concerning a class of groups* $\mathfrak{Y}$ if it satisfies the following condition:

(IH) whenever an $\mathfrak{Y}$–group $G$ belongs to the class $\mathbf{R}\mathfrak{X}$, then every finite factor–group of $G$ belongs to $\mathfrak{X}$.

Many important formations of finite groups are infinitely hereditary concerning the class of $FC$–hypercentral groups, as for example:

(1) $\mathfrak{A} \cap \mathfrak{F}$, the formation of finite abelian groups,

(2) $\mathfrak{N}_c \cap \mathfrak{F}$, the formation of finite nilpotent groups of class at most $c$,

(3) $\mathfrak{S}_d \cap \mathfrak{F}$, the formation of finite soluble groups of derived length at most $d$,

(4) $\mathfrak{S} \cap \mathfrak{F}$, the formation of finite soluble groups,

(5) $\mathfrak{B}(n) \cap \mathfrak{F}$, the formation of finite groups of exponent dividing $n$ among others. Moreover all these five examples and

(6) $\mathfrak{N} \cap \mathfrak{F}$, the formation of finite nilpotent groups,

(7) $\mathfrak{U} \cap \mathfrak{F}$, finite supersoluble groups

are infinitely hereditary concerning both the classes of the $FC$–groups and hyperfinite groups.

Main results for these examples were obtained by Kurdachenko, Otal and Subbotin [4]

**Theorem 9.** *Let $D$ be a Dedekind domain, $\mathfrak{X}$ a formation of finite groups, $G$ an infinite $\mathfrak{X}$–hypercentral group, $A$ an artinian $DG$–module. If $\mathfrak{X}$ is infinitely hereditary concerning the class of $FC$–hypercentral groups, then $A$ has the $\mathfrak{X}$–decomposition.*

**Corollary 2.** *Let $D$ be a Dedekind domain, $\mathfrak{X}$ a formation of finite groups, $G$ an infinite $\mathfrak{X}$–hypercentral group, $A$ an artinian $DG$–module. Then $A$ has the $\mathfrak{X}$–decomposition for the formations $\mathfrak{X} = \mathfrak{A} \cap \mathfrak{F}$, $\mathfrak{N}_c \cap \mathfrak{F}$, $\mathfrak{S}_d \cap \mathfrak{F}$, $\mathfrak{S} \cap \mathfrak{F}$ and $\mathfrak{B}(n) \cap \mathfrak{F}$.*

**Corollary 3.** *Let $D$ be a Dedekind domain, $\mathfrak{X}$ a formation of finite groups, $G$ an infinite $\mathfrak{X}$–hypercentral group, $A$ an artinian $DG$–module.*

*If $G$ is an $FC$–group, then $A$ has the $\mathfrak{X}$–decomposition for the formations $\mathfrak{X} = \mathfrak{A} \cap \mathfrak{F}$, $\mathfrak{N}_c \cap \mathfrak{F}$, $\mathfrak{S}_d \cap \mathfrak{F}$, $\mathfrak{S} \cap \mathfrak{F}$, $\mathfrak{B}(n) \cap \mathfrak{F}$, $\mathfrak{N} \cap \mathfrak{F}$ and $\mathfrak{U} \cap \mathfrak{F}$.*

**Corollary 4.** *Let $D$ be a Dedekind domain, $\mathfrak{X}$ a formation of finite groups, $G$ an infinite $\mathfrak{X}$–hypercentral group, $A$ an artinian $DG$–module. If $G$ is a hyperfinite group, then $A$ has the $\mathfrak{X}$–decomposition for the formations $\mathfrak{X} = \mathfrak{A} \cap \mathfrak{F}$, $\mathfrak{N}_c \cap \mathfrak{F}$, $\mathfrak{S}_d \cap \mathfrak{F}$, $\mathfrak{S} \cap \mathfrak{F}$, $\mathfrak{B}(n) \cap \mathfrak{F}$, $\mathfrak{N} \cap \mathfrak{F}$ and $\mathfrak{U} \cap \mathfrak{F}$.*

**Corollary 5.** *Let $D$ be a Dedekind domain, $\mathfrak{X}$ a formation of finite groups, $G$ an infinite $\mathfrak{X}$–hypercentral group, $A$ an artinian $DG$–module. If $G$ is a Chernikov group, then $A$ has the $\mathfrak{X}$–decomposition.*

**Corollary 6.** *Let $D$ be a Dedekind domain, $\mathfrak{X}$ a formation of finite groups, $G$ an infinite $\mathfrak{X}$–hypercentral group, $A$ an artinian $DG$–module. If $G$ is finitely generated, then $A$ has the $\mathfrak{X}$–decomposition.*

The other classical generalization of modules with finite composition series are the noetherian modules. Here the situation is quite different as we cannot consider direct decomposition any longer for a noetherian module, as the following example shows. Let $A = \langle u \rangle \times \langle v \rangle$ be a free abelian group of rank 2. We construct the split extension $G$ of $A$ by a finite cyclic group $\langle g \rangle$ of order 3, where the action is given by: $u^g = v$ and $v^g = u^{-1}v^{-1}$. Then every non-identity $G$–invariant subgroup of $A$ has finite index. In particular, the $\mathbb{Z}\langle g \rangle$–module $A$ is noetherian. However $A$ is directly indecomposable and has central and non-central $G$–chief factors.

Despite of this, Robinson [8] was able to obtain the best result known, which gives some weak form of the $\mathbb{Z} - RG$–decomposition:

**Theorem (Robinson).** *If $R$ is a commutative ring, $G$ a nilpotent group, $W$ the augmentation ideal of the group ring $RG$ and $A$ a noetherian $RG$–module, then the lower $RG$–central series $\{A_\alpha \mid A_\alpha = AW^\alpha\}$ is stabilized at the first infinite ordinal $\omega$ and there is some positive integer $n$ such that $AW^n \cap \zeta_{RG}^{\infty}(A) = \langle 0 \rangle$.*

In this direction, the following result of Zaitsev [12] is worth to mentioning.

**Theorem 10.** *Let $A$ be a noetherian $\mathbb{Z}G$–module, where $G$ is a hypercentral group. Then $A$ has a non-zero $\mathbb{Z}G$–central image if and only if it has a non-zero $\mathbb{Z}G$–central factor.*

Zaitsev [13] also showed a similar result holds for $\mathfrak{F}$, a result that was deepened by Duan [2].

**Theorem 11.** *Let $A$ be a noetherian $\mathbb{Z}G$–module, where $G$ is a hyperfinite group. Then $A$ has a non-zero finite $\mathbb{Z}G$–central image if and only if it has a non-zero finite $ZG$–central factor.*

**Theorem 12.** *Let $A$ be a noetherian $\mathbb{Z}G$–module, where $G$ is a locally soluble hyperfinite group. Then $A = B \oplus C$, where $B$ is a $\mathbb{Z}G$–module with finite chief factors and $C$ is a $\mathbb{Z}G$–module with infinite chief factors.*

This last result raises the following questions.

**Question 1.** *Let $A$ be a noetherian $\mathbb{Z}G$–module, where $G$ is a locally soluble hyperfinite group. Find out formations $\mathfrak{X}$ (finite or infinite) for which $A$ has the $\mathfrak{X}$–decomposition.*

**Question 2.** *Let $A$ be a noetherian $\mathbb{Z}G$–module, where $G$ is a locally soluble $FC$–hypercentral group. Does have $A$ the $\mathfrak{F}$–decomposition?*

## References


[1] Z. Y. Duan. *Extensions of abelian–by–hyper (cyclic or finite) groups.* Comm. Algebra **20** (1992), 2305-2321.

[2] Z. Y. Duan. *The decompositions of noetherian modules over hyperfinite groups.* Ricerche Math. **44** (1995), 65-89.

[3] B. Hartley and M. J. Tomkinson. *Splitting over nilpotent and hyperfinite residuals.* Math. Proc. Cambridge Phil. Soc. **78** (1975), 215-226.

[4] L. A. Kurdachenko, J. Otal and I. Ya. Subbotin. *Baer-like decompositions of modules.* Illinois Math. J., to appear.

[5] L. A. Kurdachenko, B. V. Petrenko and I. Ya. Subbotin. *On generalized hypercenters in artinian modules.* Comm. Algebra **25** (1997), 1023-1046.

[6] L. A. Kurdachenko, B. V. Petrenko and I. Ya. Subbotin. *Direct decompositions in artinian modules over FC–hypercentral groups.* Mat. Contemporanea **14** (1998), 89-99.

[7] B. V. Petrenko. *One generalization of Fitting's theorem defined by the formations of finite groups.* In *The classes of groups with restrictions for subgroups.* Math. Inst. Kyiv 1997, 73-78.

[8] D. J. S. Robinson. *Hypercentral ideals, noetherian modules and theorem of Stroud.* J. Algebra **32** (1974), 234-239.

[9] D. I. Zaitsev. *On the existence of direct complements in groups with operators.* In *Investigations in Group Theory.* Math. Inst. Kyiv, Kyiv 1976, 26-44.

[10] D. I. Zaitsev. *The splitting extensions of abelian groups.* In *The structure of groups and the properties of their subgroups.* Math. Inst. Kyiv 1986, pp. 22-31.

[11] D. I. Zaitsev. *On direct decompositions of abelian groups with operators.* Ukrain. Math. J. **40** (1988), 303-309.

[12] D. I. Zaitsev. *On extensions of abelian groups.* In *The constructive description of groups with prescribed properties of subgroups.* Math. Inst. Kyiv 1980,pp. 16-40.

[13] D. I. Zaitsev. *The hyperfinite extensions of abelian groups*. In *The investigations of groups with restrictions on subgroups*, Math. Inst. Kyiv, Kyiv 1988, 17-26.

Contact information

**L. A. Kurdachenko** Department of Algebra, Facultet of mathematics and mechanik, University of Dnepropetrovsk, Vulycia Naukova 13, Dnepropetrovsk 50 49050 UKRAINE
*E-Mail:* lkurdachenko@ua.fm

**J. Otal** Departamento de Algebra Facultad de Ciencias Universidad de Zaragoza Pedro Cerbuna 12 50009 Zaragoza SPAIN
*E-Mail:* otal@posta.unizar.es

Received by the editors: 23.10.2002.

Algebra and Discrete Mathematics RESEARCH ARTICLE
Number 2. **(2003).** pp. 77–86


# Hyperbolic spaces from self-similar group actions

**Volodymyr Nekrashevych**

Communicated by V. I. Sushchansky

Abstract. We show that the limit space of a contracting self-similar group action is the boundary of a naturally defined Gromov hyperbolic space.

## 1. Introduction

Self-similar group actions (or self-similar groups) have proved to be interesting mathematical objects from the point of view of group theory and many other fields of mathematics (operator algebras, holomorphic dynamics, automata theory, etc). See the works [BGN02, GNS00, Gri00, Sid98, BG00, Nek02a], where different aspects of self-similar groups are studied.

An important class of self-similar group actions are *contracting actions*. Contracting groups have many nice properties. For example, the word problem is solvable in a contracting group in a polynomial time [Nek]. The author has shown (see [Nek02b]) that a naturally defined topological space, called the *limit space*, is associated with every contracting self-similar action. This topological space is metrizable and finite-dimensional.

It was discovered later (see [Nek02a]) that with many topological dynamical systems (like iterations of a rational function) a self-similar group

*The research was supported by the Swiss National Science Foundation and Alexander von Humboldt Foundation*

**2001 Mathematics Subject Classification:** *37B05, 37D99.*

**Key words and phrases:** *self-similar group actions, limit space, Gromov hyperbolic spaces, boundary of a hyperbolic space.*

is associated. This self-similar group (called the *iterated monodromy group*) is often contracting and in fact contains all the essential dynamics of the original dynamical system. In particular, if the dynamical system is expanding, then its iterated monodromy group is contracting and the limit space of the iterated monodromy group is homeomorphic to the Julia set of the dynamical system.

We show in this paper how to associate with every self-similar action a graph, which will be Gromov-hyperbolic if the group action is contracting. The boundary of this graph is then homeomorphic to the limit space of the group action.

In some sense this result (together with the notion of iterated monodromy group) can be interpreted as a new entry to the "Sullivan dictionary" [Sul85], which shows an analogy between the holomorphic dynamics and groups acting on hyperbolic spaces.

The author is grateful to L. Bartholdi, R. Grigorchuk, V. Kaimanovich for helpful discussions on the subject of the paper.

## 2. Self-similar actions

Let $X$ be a finite set (*alphabet*). By $X^*$ we denote the set of all finite words over the alphabet $X$, including the empty one $\emptyset$. By $X^{-\omega}$ we denote the space of all infinite sequences to the left of the form $\ldots x_2x_1$, with the topology of the infinite countable direct power of the discrete set $X$.

**Definition 2.1.** *A faithful action of a group $G$ on the set $X^*$ is said to be* self-similar *if for every $g \in G$ and every $x \in X$ there exist $h \in G$ such that*

$$g(xw) = g(x)h(w) \tag{1}$$

*for every $w \in X^*$.*

Iterating equation (1), we get that for every $v \in X^*$ there exists an element $h$ such that $g(vw) = g(v)h(w)$. The element $h$ is uniquely defined and is called *restriction of the element* $g$ in the word $v$ and is denoted $g|_v$.

We easily get the following properties of restrictions:

$$g|_{v_1v_2} = g|_{v_1}|_{v_2} \qquad (g_1g_2)|_v = \left(g_1|_{g_2(v)}\right)\left(g_2|_v\right). \tag{2}$$

**Definition 2.2.** *Let $P_n : (X^n)^* \longrightarrow X^*$ be the injective map, which carries the word*

$$(x_1, x_2, \ldots, x_n)(x_{n+1}, x_{n+2}, \ldots, x_{2n}) \ldots$$
$$\ldots (x_{(k-1)n+1}, x_{(k-1)n+2}, \ldots, x_{kn}) \in (X^n)^*$$

*to the word* $x_1x_2x_3\dots x_{kn} \in X^*$.

*The* $n$th tensor power *of a self-similar action is the action on the set* $(X^n)^*$ *obtained by conjugating the action on* $X^*$ *by the map* $P_n$.

The $n$th power of a self-similar action obviously is also self-similar.

**The adding machine.** One of the most classical examples of a self-similar action is the "*adding machine*" defined in the following way. We put $X = \{0, 1\}$ and define a transformation $a$ recurrently by the rules

$$\begin{aligned} a(0w) &= 1w \\ a(1w) &= 0a(w) \end{aligned}$$

for all $w \in X^*$.

It is easy to prove that if $a^n(x_1x_2\dots x_m) = y_1y_2\dots y_m$ then

$$y_1 + y_2 \cdot 2 + y_3 \cdot 2^2 + y_4 \cdot 2^3 + \dots y_m \cdot 2^{m-1} = \left(x_1 + x_2 \cdot 2 + x_3 \cdot 2^2 + x_4 \cdot 2^3 + \dots x_m \cdot 2^{m-1}\right) + n \pmod{2^m}.$$

## 3. Contracting actions and their limit spaces

**Definition 3.1.** *A self-similar action of a group* $G$ *is called* contracting *if there exists a finite set* $\mathcal{N} \subset G$ *such that for every* $g \in G$ *there exists* $k \in \mathbb{N}$ *such that* $g|_v \in \mathcal{N}$ *for all the words* $v \in X^*$ *of length* $\geq k$. *The minimal set* $\mathcal{N}$ *with this property is called the* nucleus *of the self-similar action.*

As an example of a contracting action, one can take the adding machine action. Other examples of contracting actions include the Grigorchuk group and the iterated monodromy groups of post-critically finite rational functions.

Let us fix some contracting self-similar action of a group $G$ on the set $X^*$.

**Definition 3.2.** *Two sequences* $\dots x_3x_2x_1, \dots y_3y_2y_1 \in X^{-\omega}$ *are said to be* asymptotically equivalent *with respect to the action of the group* $G$, *if there exist a finite set* $K \subset G$ *and a sequence* $g_k \in K, k \in \mathbb{N}$ *such that*

$$g_k(x_kx_{k-1}\dots x_2x_1) = y_ky_{k-1}\dots y_2y_1$$

*for every* $k \in \mathbb{N}$.

**Definition 3.3.** *The* limit space *of the self-similar action (denoted* $\mathcal{J}_G$*) is the quotient of the topological space* $X^{-\omega}$ *by the asymptotic equivalence relation.*

**Example.** In the case of the adding machine action of $\mathbb{Z}$ one can prove (see [Nek02b]) that two sequences are asymptotically equivalent if and only if

$$\sum_{n=1}^{\infty} x_n \cdot 2^{-n} = \sum_{n=1}^{\infty} y_n \cdot 2^{-n} \pmod 1.$$

Consequently, the limit space $\mathcal{J}_{\mathbb{Z}}$ is the circle $\mathbb{R}/\mathbb{Z}$.

We have the following properties of the limit spaces, which are proved in [Nek02b].

**Proposition 3.1.** *The limit space $\mathcal{J}_G$ is metrizable and has topological dimension $\leq |\mathcal{N}| - 1$, where $\mathcal{N}$ is the nucleus of the action.*

## 4. The limit space as a hyperbolic boundary

**Definition 4.1.** *Let $G$ be a finitely generated group with a self-similar action on $X^*$. For a given finite generating system $S$ of the group $G$ we define the* self-similarity graph $\Sigma(G, S)$ *as the graph with the set of vertices $X^*$ and two vertices $v_1, v_2 \in X^*$ belonging to a common edge if and only if either $v_i = xv_j$ for some $x \in X$ (the* vertical *edges) or $s(v_i) = v_j$ for some $s \in S$ (the* horizontal *edges), where $\{i, j\} = \{1, 2\}$.*

As an example, see a part of the self-similarity graph of the adding machine on Figure 1.

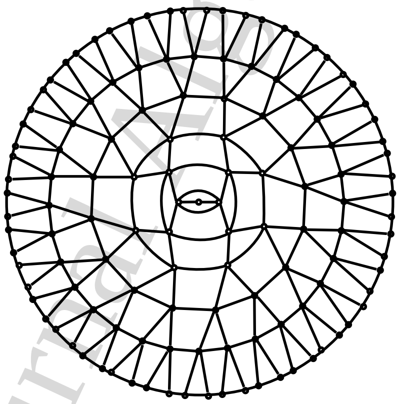


Figure 1: The self-similarity graph of the adding machine

If all restriction of the elements of the generating set $S$ also belong to $S$, then the self-similarity graph $\Sigma(G, S)$ is an *augmented tree* in sense of V. Kaimanovich (see [Kai03]).

We have a natural metric on the set of the vertices of any graph. The distance between two vertices in this metric is equal to the number of edges in the shortest path connecting them.

The definition of the self-similarity graph depends on the choice of the generating set $S$. We will use the classical notion of quasi-isometry in order to make it more canonical.

**Definition 4.2.** *Two metric spaces $\mathfrak{X}$ and $\mathfrak{Y}$ are said to be* quasi-isometric *if there exists a map (which is called then a* quasi-isometry*) $f : \mathfrak{X} \longrightarrow \mathfrak{Y}$ and positive constants $L, C$ such that*

*(i)*

$$L^{-1}d_{\mathfrak{X}}(x_1, x_2) - C < d_{\mathfrak{Y}}(f(x_1), f(x_2)) < Ld_{\mathfrak{X}}(x_1, x_2) + C,$$

*for all $x_1, x_2 \in \mathfrak{X}$ and*

*(ii) for every $y \in \mathfrak{Y}$ there exists $x \in \mathfrak{X}$ such that $d_{\mathfrak{Y}}(y, f(x)) < C$.*

**Lemma 4.1.** *Let $G$ be a group with a self-similar action, and let $\phi$ be the associated virtual endomorphism.*

1. *The self-similarity graphs $\Sigma(G, S_1)$ and $\Sigma(G, S_2)$, where $S_1, S_2$ are two different finite generating sets of the group $G$, are quasi-isometric.*

2. *The self-similarity graph of the $n$th tensor power of the self-similar action is quasi-isometric to the self-similarity graph of the original action.*

*Proof.* 1) The identical map on the set of vertices $\Sigma(G, S_1) \longrightarrow \Sigma(G, S_2)$ is a quasi-isometry. The constant $L$ is any number such that the length of every element of one of the generating sets has length less than $L$ with respect to the other generating set. The constant $C$ can be any positive number.

2) The set of vertices of the $n$th tensor power is equal to $\{\emptyset\} \cup X^n \cup X^{2n} \cup X^{3n} \cup \ldots$. Let $F : \Sigma_n(G, S) \longrightarrow \Sigma(G, S)$ be the natural inclusion of the vertex sets.

It is easy to see that $d(F(u), F(v)) \leq n \cdot d(u, v)$ and $d(F(u), F(v)) \geq d(u, v)$ for all $u, v \in \Sigma(G, S)$.

For every vertex $v = x_1x_2 \ldots x_m \in X^*$ of the graph $\Sigma(G, S)$ there exists a vertex $x_rx_{r+1} \ldots x_m$ belonging to the vertex set of the graph $\Sigma_n(G, S)$, which is at the distance less than $n$ from $v$ (one must take $r$ to be the minimal number, such that $m - r + 1$ is divisible by $n$). So the map $F$ satisfies both conditions of Definition 4.2. ☐

Let us recall the definition of Gromov-hyperbolic metric spaces [Gro87].

Let $\mathfrak{X}$ be a metric space with the metric $d(\cdot,\cdot)$. The *Gromov product* of two points $x, y \in \mathfrak{X}$ with respect to the basepoint $x_0 \in \mathfrak{X}$ is the number

$$\langle x \cdot y \rangle = \langle x \cdot y \rangle_{x_0} = \frac{1}{2}\left(d\left(x, x_0\right) + d\left(y, x_0\right) - d\left(x, y\right)\right).$$

**Definition 4.3.** *A metric space $\mathfrak{X}$ is said to be* Gromov-hyperbolic *(see [Gro87]) if there exists $\delta > 0$ such that the inequality*

$$\langle x \cdot y \rangle \geq \min\left(\langle x \cdot z \rangle, \langle y \cdot z \rangle\right) - \delta \tag{3}$$

*holds for all $x, y, z \in \mathfrak{X}$.*

If a proper geodesic metric space (for instance a graph) is quasi-isometric to a hyperbolic space, then it is also hyperbolic. For proofs of the mentioned facts and for other properties of the hyperbolic spaces and groups look one of the books [Gro87, CDP90, GH90].

**Theorem 4.2.** *If the action of a finitely-generated group $G$ is contracting then the self-similarity graph $\Sigma(G, S)$ is a Gromov-hyperbolic space.*

*Proof.* It is sufficient to prove that some quasi-isometric graph is hyperbolic. Therefore, we can change by statement (1) of Lemma 4.1 the set of generators $S$ so that it will contain all the restrictions of its elements and that there exists $N \in \mathbb{N}$ such that for every element $g \in G$ of length $\leq 4$ and any word $x_1x_2\ldots x_N \in X^*$, the restriction $g|_{x_1x_2\ldots x_N}$ belongs to $S$. Then the length of any restriction of an element $g \in G$ is not greater then the length of $g$.

It is sufficient, by statement (3) of Lemma 4.1, to prove that the self-similarity graph $\Sigma_N(G, S)$ of the $N$th tensor power of the action is Gromov-hyperbolic.

Let us prove the following lemma.

**Lemma 4.3.** *Any two vertices $w_1, w_2$ of the graph $\Sigma_N(G, S)$ can be written in the form $w_1 = a_1a_2\ldots a_nw, w_2 = b_1b_2\ldots b_mg(w)$, where $a_i, b_i \in X^N$, $w \in \left(X^N\right)^*$, $g \in G$, $l(g) \leq 4$ and $d(w_1, w_2) = n + m + l(g)$.*

*Then the Gromov product $\langle w_1 \cdot w_2 \rangle$ with respect to the basepoint $\emptyset$ is equal to $|w| - l(g)/2$.*

Here $l(g)$ denotes the length of the element $g$ with respect to some fixed finite generating set of the group.

*Proof.* Let $v_1 = w_1, v_2, \ldots v_k = w_2$ be the consecutive vertices of the shortest path connecting the vertices $w_1$ and $w_2$. Then every $v_{i+1}$ is obtained from $v_i$ by application of one of the following procedures:

1. deletion of the first letter $a \in X^N$ in $v_i$ (*descending edges*);
2. adding a letter $a \in X^N$ to the beginning of $v_i$ (*ascending edges*);
3. application of an element of $S$ to $v_i$ (*horizontal edges*);

If the path has three consecutive vertices $v_i, v_{i+1}, v_{i+2}$ such that $v_{i+1} = av_i$, $a \in X^N$ and $v_{i+2} = s(v_{i+1})$ for $s \in S$ then $v_{i+2} = bs'(v_i)$, where $b = s(a) \in X^N$ and $s' = s|_a \in S$. We replace the segment $\{v_i, v_{i+1}, v_{i+2}\}$ of the path by the segment $\{v_i, s'(v_i), bs'(v_i) = v_{i+2}\}$.

If the path has three consecutive vertices $v_i, v_{i+1}, v_{i+2}$ such that $v_{i+1} = s(v_i)$ for $s \in S$ and $v_{i+1} = av_{i+2}$ then $v_i = s^{-1}(av_{i+2}) = bs'(v_{i+2})$, where $b = s^{-1}(a) \in X^N$ and $s' = s^{-1}|_a \in S$. Then we replace the segment $\{v_i, v_{i+1}, v_{i+2}\}$ of the path by the segment $\{v_i = bs'(v_{i+2}), s'(v_{i+2}), v_{i+2}\}$.

Let us perform these replacements as many times as possible. Then we will not change the length of the path, so each time we will get a geodesic path connecting the vertices $w_1, w_2$. Note that a geodesic path can not have a descending edge next after an ascending one. Therefore, eventually after a finite number of replacements we will get a geodesic path in which we have at first only descending, then horizontal and then only ascending edges. Then $w_1 = a_1a_2\ldots a_nw, w_2 = b_1b_2\ldots b_mg(w)$, with $a_i, b_i \in X^N$, $w \in \left(X^N\right)^*$, $g \in G$, and $d(w_1, w_2) = n + m + l(g)$.

Suppose that $l(g) > 4$. Let $w = aw'$, $a \in X^N$ and denote $b = g(a)$ and $h = g|_a$. Then by the choice of the number $N$ we have $l\left(h\right) \le l(g) - 3$. Since $w_1 = a_1a_2\ldots a_naw'$ and $w_2 = b_1b_2\ldots b_mbh(w')$, we have $d(w_1, w_2) \le n + 1 + m + 1 + l(h) \le n + m + l(g) - 1$, which contradicts to the fact that the original path was the shortest one.

We have

$$\langle w_1 \cdot w_2\rangle = \frac{1}{2}\left(n + |w| + m + |w| - (n + m + l(g))\right) = |w| - \frac{l(g)}{2}.$$

□

Let us take three points $w_1, w_2, w_3$. We can write them by Lemma 4.3 as

$$w_1 = a_1a_2\ldots a_nw, \qquad w_2 = b_1b_2\ldots b_mg_1(w)$$

and

$$w_2 = b_1b_2\ldots b_pu, \qquad w_3 = c_1c_2\ldots c_qg_2(u),$$

where $a_i, b_i, c_i \in X^N$, $g_1, g_2 \in G$, $l(g_1), l(g_2) \le 4$ and

$$\langle w_1 \cdot w_2\rangle = |w| - l(g_1)/2, \qquad \langle w_2 \cdot w_3\rangle = |u| - l(g_2)/2.$$

We can assume that $p \le m$. Then $|u| \ge |w| = |g_1(w)|$, so we can write $u = vg_1(w)$ for some $v \in \left(X^N\right)^*$. Then $w_3 = c_1c_2\dots c_q g_2(v)hg_1(w)$, where $h = g_2|_v$. We have $l(h) \le l(g_2) \le 4$ and $d(w_1, w_3) \le n + l(h) + l(g_1) + q + |v|$, hence

$$\langle w_1 \cdot w_3 \rangle = \frac{1}{2}(n + |w| + q + |v| + |w| - d(w_1, w_3)) \ge$$
$$\ge |w| - (l(h) + l(g_1))/2 \ge |w| - 4.$$

Finally, $\min(\langle w_1 \cdot w_2\rangle, \langle w_2 \cdot w_3\rangle) \le \langle w_1 \cdot w_2 \rangle \le |w|$, so

$$\langle w_1 \cdot w_3 \rangle \ge \min(\langle w_1 \cdot w_2\rangle, \langle w_2 \cdot w_3\rangle) - 4,$$

and the graph $\Sigma_N(G, S)$ is 4-hyperbolic. □

Let $\mathfrak{X}$ be a hyperbolic space. We say that a sequence $\{x_n\}$ of points of $\mathfrak{X}$ *converges to the infinity* if the Gromov product $\langle x_n \cdot x_m\rangle$ trends to infinity when $m, n \to \infty$. This definition does not depend on the choice of the basepoint. We say that two sequences $\{x_n\}$ and $\{y_n\}$, convergent to the infinity, are *equivalent* if $\lim_{n,m\to\infty}\langle x_n \cdot y_m\rangle = \infty$.

The set of the equivalence classes of the sequences convergent to the infinity in the space $\mathfrak{X}$ is called the *hyperbolic boundary* of the space $\mathfrak{X}$ and is denoted $\partial\mathfrak{X}$. If a sequence $\{x_n\}$ converges to the infinity, then its *limit* is the equivalence class $a \in \partial\mathfrak{X}$, to which belongs $\{x_n\}$ and we say that $\{x_n\}$ *converges to* $a$.

If $a, b \in \partial\mathfrak{X}$ are two points of the boundary, then their *Gromov product* is defined as

$$\langle a \cdot b\rangle = \sup_{\{x_n\}\in a, \{y_m\}\in b} \liminf_{m,n\to\infty} \langle x_n \cdot y_m\rangle.$$

For every $r > 0$ define

$$V_r = \{(a, b) \in \partial\mathfrak{X} \times \partial\mathfrak{X} : \langle a \cdot b\rangle \ge r\}.$$

Then the set $\{V_r : r \ge 0\}$ is a fundamental neighborhood basis of a uniform structure on $\partial\mathfrak{X}$ (see [Bou71] for the definition of a uniform structure and [GH90] for proofs). We introduce on the boundary $\partial\mathfrak{X}$ the topology defined by this uniform structure.

**Theorem 4.4.** *The limit space $\mathcal{J}_G$ of a contracting action of a finitely generated group $G$ is homeomorphic to the hyperbolic boundary $\partial\Sigma(G, S)$ of the self-similarity graph $\Sigma(G, S)$. Moreover, there exists a homeomorphism $F : \mathcal{J}_G \longrightarrow \partial\Sigma(G, S)$, such that $D = F \circ \pi$, were $\pi : X^{-\omega} \longrightarrow \mathcal{J}_G$ is the canonical projection and $D : X^{-\omega} \longrightarrow \partial\Sigma(G, S)$ carries every sequence $\dots x_2x_1 \in X^{-\omega}$ to its limit*

$$\lim_{n\to\infty} x_n x_{n-1} \dots x_1 \in \partial\Sigma(G, S).$$

*Sketch of the proof.* We will need the following well known result (see [CDP90] Theorem 2.2).

**Lemma 4.5.** *Let $\mathfrak{X}_1, \mathfrak{X}_2$ be proper geodesic hyperbolic spaces and let $f_1 : \mathfrak{X}_1 \longrightarrow \mathfrak{X}_2$ be a quasi-isometry. Then a sequence $\{x_n\}$ of points of $\mathfrak{X}_1$ converges to infinity if and only if the sequence $\{f_1(x_n)\}$ does. The map $\partial f_1 : \{x_n\} \mapsto \{f_1(x_n)\}$ defines a homeomorphism $\partial f_1 : \partial\mathfrak{X}_1 \longrightarrow \partial\mathfrak{X}_2$ of the boundaries.*

We pass, using Lemma 4.5, to an $N$th tensor power of the self-similar action in the same way as in the proof of Theorem 4.2, so that for every $g \in G$ such that $l(g) \leq 4$ and for every $a \in X^N$ the restriction $g|_a$ belongs to the generating set.

Suppose that the sequence $\{w_n\}$ converges to the infinity. Choose its convergent subsequence in $X^{-\omega} \cup X^*$. Suppose its limit is $\ldots x_2x_1 \in X^{-\omega}$. The Gromov product $\langle w_i \cdot w_j \rangle$ with respect to the basepoint $\emptyset$ is equal to $|w| - l(g)/2$, where $w$ and $g$ are as in Lemma 4.3. It follows that all the accumulation points of $\{w_n\}$ in $X^{-\omega}$ are asymptotically equivalent to $\ldots x_2x_1$. We also have that the sequence $\{x_nx_{n-1}\ldots x_1\}_{n\geq 1}$ converges to the same point of the hyperbolic boundary as $\{w_n\}$. So the map $D : X^{-\omega} \longrightarrow \partial\Sigma(G, S)$ is surjective and the map $F : \mathcal{J}_G \longrightarrow \partial\Sigma(G, S)$, satisfying the conditions of the theorem, is uniquely defined.

Let $A = \{g \in G : l(g) \leq 4\}$ and for every $n \in \mathbb{N}$ define

$$U_n = \{(w_1v, w_2s(v)) : w_1, w_2 \in X^{-\omega}, v \in X^n, s \in A\} \subset X^{-\omega} \times X^{-\omega}.$$

By $\tilde{U}_n$ we denote the image of $U_n$ in $\mathcal{J}_G \times \mathcal{J}_G$.

It is easy to prove now that $\cap_{n\geq 1} U_n$ is equal to the asymptotic equivalence relation on $X^{-\omega}$ and that $\tilde{U}_n$ is the basis of neighborhoods of a uniform structure, defining the topology on $\mathcal{J}_G$.

It follows from Lemma 4.3 that

$$V_{n-2} \subseteq D \times D(U_n) \subseteq V_n.$$

This implies that map $F$ is a homeomorphism. □

## References

[BG00] Laurent Bartholdi and Rostislav I. Grigorchuk, *On the spectrum of Hecke type operators related to some fractal groups*, Proceedings of the Steklov Institute of Mathematics **231** (2000), 5–45.

[BGN02] Laurent Bartholdi, Rostislav I. Grigorchuk, and Volodymyr V. Nekrashevych, *From fractal groups to fractal sets*, Fractals in Graz 2001, Analysis–Dynamics–Geometry–Stochastics (Peter Grabner and Wolfgang Woess, eds.), Trends in Mathematics, vol. 19, Birkhäuser, 2003.

[Bou71] Nicolas Bourbaki, *Éléments de mathématique. Topologie générale. Chapitres 1 à 4*, Hermann, Paris, 1971.

[CDP90] Michel Coornaert, Thomas Delzant, and Athanase Papadopoulos, *Geometrie et theorie des groupes: Les groupes hyperboliques de Gromov*, Lectures Notes in Mathematics, vol. 1441, Springer Verlag, 1990.

[GH90] Étienne Ghys and Pierre de la Harpe, *Sur les groupes hyperboliques d'après Mikhael Gromov*, Progress in Mathematics, vol. 83, Birkhäuser Boston Inc., Boston, MA, 1990, Papers from the Swiss Seminar on Hyperbolic Group held in Bern, 1988.

[GNS00] Rostislav I. Grigorchuk, Volodymyr V. Nekrashevich, and Vitaliĭ I. Sushchanskii, *Automata, dynamical systems and groups*, Proceedings of the Steklov Institute of Mathematics **231** (2000), 128–203.

[Gri00] Rostislav I. Grigorchuk, *Just infinite branch groups*, New horizons in pro-$p$ groups (Aner Shalev, Marcus P. F. du Sautoy, and Dan Segal, eds.), Progress in Mathematics, vol. 184, Birkhäuser Verlag, Basel, etc., 2000, pp. 121–179.

[Gro87] Mikhael Gromov, *Hyperbolic groups*, Essays in Group Theory (S. M. Gersten, ed.), M.S.R.I. Pub., no. 8, Springer, 1987, pp. 75–263.

[Kai03] Vadim A. Kaimanovich, *Random walks on Sierpiński graphs: hyperbolicity and stochastic homogenization*, Fractals in Graz 2001, Analysis–Dynamics–Geometry–Stochastics (Peter Grabner and Wolfgang Woess, eds.), Trends in Mathematics, vol. 19, Birkhäuser, 2003.

[Nek02a] Volodymyr V. Nekrashevych, *Iterated monodromy groups*, preprint, Geneva University, 2002.

[Nek02b] ———, *Limit spaces of self-similar group actions*, preprint, Geneva University, 2002.

[Nek] ———, *Virtual endomorphisms of groups*, Algebra and Discrete Mathematics, Number 1, (2002), 96–136.

[Sid98] Said N. Sidki, *Regular trees and their automorphisms*, Monografias de Matematica, vol. 56, IMPA, Rio de Janeiro, 1998.

[Sul85] Dennis Sullivan, *Quasi-conformal homeomorphisms and dynamics I, solution of the Fatou-Julia problem on wandering domains*, Ann. Math. **122** (1985), 401–418.


## CONTACT INFORMATION


**V. Nekrashevych** Kyiv Taras Shevchenko University
Mechanics and Mathematics Faculty
Algebra Department
Volodymyrska, 60
Kyiv, Ukraine, 01033
*E-Mail:* nazaruk@ukrpack.net


Received by the editors: 30.01.2003.

Algebra and Discrete Mathematics RESEARCH ARTICLE
Number 2. **(2003).** pp. 87–92


# Principal quasi-ideals of cohomological dimension 1

**B. V. Novikov**

Abstract. We prove that a principal quasi-ideal of a non-commutative free semigroup has cohomological dimension 1 if and only if it is free.

In this note we continue to study semigroups of cohomological dimension 1 (c.d. 1). In [4] an analog of the Stallings—Swan theorem [1] was proved: a cancellative semigroup of c.d. 1 can be embedded into a free group. However, we have not complete description even for a free semigroup, because its subsemigroups can both have and not have c.d. 1.

In [4] following results about ideals of free semigroups were obtained:

– c.d. of every proper two-side ideal is greater than 1;

– c.d. of every left ideal equals 1 if and only if it is free;

– c.d. of every principal right ideal equals 1 if and only if it is free; this is not true for non-principal right ideals.

In [4] a problem was proposed: describe principal quasi-ideals of the free semigroup having c.d. 1. We solve this problem below. The answer is the same as for ideals: a principal quasi-ideal has c.d. 1 if and only if it is free. Nevertheless the proof of this assertion is carried out in different ways for two kinds of quasi-ideals.

In what follows we shall denote by $S$ a semigroup with an adjoined identity; by $F$ a free non-commutative semigroup; by $|a|$ the length of a word $a \in F$; by $\langle a \rangle$ (resp. $\langle a \rangle_q$) the subsemigroup (resp. quasi-ideal) generated by element $a$.

We recall that a subset $Q$ of a semigroup $S$ is called a *quasi-ideal* [6] if $QS \cap SQ \subset Q$. A *principal quasi-ideal* generated by element $w \in S$ is a subset $\langle w \rangle_q = S^1 w \cap w S^1$. We need to separate the case when $w$ is

**2001 Mathematics Subject Classification:** *20M05, 20M50.*

**Key words and phrases:** *cohomology of semigroups, cohomological dimension, quasi-ideal, free semigroup.*

not a power of any element different from $w$; such element $w$ is called *primitive*.

The next properties of free semigroups will be used mostly without reference to them:

1) If $ab = cd$ for some $a, b, c, d \in F$ and $|a| \leq |c|$ then $c \in aF^1$.

2) A subsemigroup $S \subset F$ is not free if and only if $\alpha S \cap S \neq \emptyset \neq S\alpha \cap S$ for some $\alpha \in F \setminus S$ ([3], Prop. 5.2.2).

3) If $uw = wv$ for some $u, v \in F$, $w \in F^1$ then there are $x, y \in F^1$ and an integer $k \geq 0$ such that $u = xy$, $v = yx$, $w = (xy)^k x$ ([3], Lemma 11.5.1).

In particular:

4) If $uv = uv$ $(u, v \in F^1$ then $u = x^m$, $v = x^n$ for some $x \in F$ and integers $m, n \geq 0$.

First we shall study the structure of principal quasi-ideals.

**Lemma 1.** *Let $x, w \in F$, $S = \langle w \rangle_q$. If $xS \cap S \neq \emptyset$ then $xS \subset S$; if $Sx \cap S \neq \emptyset$ then $Sx \subset S$.*

*Proof.* Let $y \in xS \cap S \subset xwF^1 \cap wF^1$, i.e. $y = xwf = wg$ for some $f, g \in F^1$. Then $|g| \geq |f|$ whence $g = hf$ $(h \in F^1)$ and $xw = wh$. So

$$xS = xF^1 w \cap xwF^1 = xF^1 w \cap whF^1 \subset F^1 w \cap wF^1 = S.$$

The second part of Lemma is proved analogously. □

**Theorem 1.** *A quasi-ideal $\langle w \rangle_q$ $(w \in F)$ is free if and only if $w$ is primitive.*

*Proof.* 1) Let $w = x^n$, $n \geq 2$. Since $|x| < |w|$, $x \notin \langle w \rangle_q$. On another hand

$$xw \in x\langle w \rangle_q \cap \langle w \rangle_q, \qquad wx \in \langle w \rangle_q x \cap \langle w \rangle_q,$$

i.e. $\langle w \rangle_q$ $(w \in F)$ is not free.

2) Let $w$ is primitive and $\langle w \rangle_q$ $(w \in F)$ is not free. Then

$$x\langle w \rangle_q \cap \langle w \rangle_q \neq \emptyset \neq \langle w \rangle_q x \cap \langle w \rangle_q$$

for some $x \in F \setminus \langle w \rangle_q$. By Lemma 1 $x\langle w \rangle_q \cup \langle w \rangle_q x \subset \langle w \rangle_q$. In particular

$$xw \in wF^1, \qquad wx \in F^1 w, \tag{1}$$

so if $|x| \geq |w|$ then $x \in \langle w \rangle_q$, what is impossible.

Therefore $|x| < |w|$. It follows from (1) that $w = ux = xv$ for some $u, v \in F$. Then there are $a, b \in F^1$ and $k \geq 0$ such that $u = ab$, $v = ba$, $x = (ab)^k a$ whence $w = (ab)^{k+1} a$.

The inclusions (1) imply too that $xw = wt$ for some $t \in F^1$. Substituting the values of $x$ and $w$ in this equation and cancelling, we obtain: $(ab)^{k+1}a = bat$. Then $ab = ba$ because $|ab| = |ba|$. Therefore $a = c^p$, $b = c^q$ for some $c \in F$ and $p, q \geq 0$. Then $w = c^{(p+q)(k+1)+p}$. The primitivity of $w$ implies $(p+q)(k+1) + p = 0$ whence $p = q = 0$ in contradiction with $w \in F$. □

Now we express arbitrary principal quasi-ideals by means of the free ones.

**Lemma 2.** *Let $a, b, w \in F$, $w$ is primitive, $n \geq 2$ and $aw^n = w^n b$. Then either $a = b \in \langle w \rangle$ or $a = w^{n-1}x$, $b = yw^{n-1}$ for such $x, y \in F^1$ that $xw = wy$.*

Proof uses induction on $n$. Suppose that $|a| < |w|$. Then $w = aw_1$ and $(aw_1)^n = w_1(aw_1)^{n-1}b$. Since $|aw_1| = |w_1 a|$, the last equation implies $aw_1 = w_1 a$. Hence $a = t^p$, $w_1 = t^q$ ($t \in F$, $p, q \geq 0$). But $a \neq 1$, so $w_1 = 1$ and $w = a$ in contradiction with $|a| < |w|$.

Thus $|a| = |b| \geq |w|$ whence $a = wa_1$, $b = b_1 w$, $a_1 w^{n-1} = w^{n-1} b_1$. If $a_1 = 1$ then $b_1 = 1$ and $a = b = w$. Otherwise we get by induction either $a_1 = b_1 \in \langle w \rangle$ (and then $a = b \in \langle w \rangle$) or $a_1 = w^{n-2}x$, $b_1 = yw^{n-2}$ (and then $a = w^{n-1}x$, $b = yw^{n-1}$). □

**Corollary 1.** *Let $w$ is primitive, $n \geq 2$. Then*

$$\langle w^n \rangle_q = w^{n-1} \langle w \rangle_q w^{n-1} \cup \{w^k \mid k \geq n\}. \quad \square$$

Now we pass to studying of cohomological dimension.

Recall that the $n$th cohomology group of semigroup $S$ with values in a left $S$-module $A$ is defined as $H^n(S, A) = \mathrm{Ext}^n_{\mathbf{Z}S}(\mathbf{Z}, A)$; another definition of semigroup cohomology in terms of cochains see, e.g. in [2] or [5]. The *cohomological dimension* of $S$ (c.d.$(S)$) is the smallest integer $n$ such that $H^k(S, A) = 0$ for every $S$-module $A$ and $k > n$.

The next assertion is a start point for the further consideration:

**Lemma 3.** ([5], Prop. 3.2) *Let c.d.$(S) = 1$ and $\alpha S \cap S \neq \emptyset \neq S\alpha \cap S$ for some $\alpha \in F \setminus S$ (so $S$ is not free). There exists $x \in \alpha S \cap S$ such that for every $u \in \alpha S \cap S$ one can choose $\lambda_1, \dots, \lambda_n \in F^1$ satisfying the next conditions:*

(i) $x\lambda_i \in \alpha S \cap S \qquad (1 \leq i \leq n)$,
(ii) $S^1 \cap \lambda_1 S^1 \neq \emptyset, \quad \lambda_i S^1 \cap \lambda_{i+1} S^1 \neq \emptyset \qquad (1 \leq i < n)$,
(iii) $u = x\lambda_n$. □

For quasi-ideals this lemma is modified as follows:

**Lemma 4.** *Let a quasi-ideal $S = \langle w\rangle_q \subset F$ is not free and c.d.$(S) = 1$. Then for every $\lambda \in F^1$ from $w\lambda \in F^1 w$ it follows $\lambda S \subset S$.*

*Proof.* First note that in situation when $S = \langle w\rangle_q$, one can set $n = 1$ in Lemma 3. Indeed, $\lambda_1 S \subset S$ (see Lemma 1), so it follows from $\lambda_1 S \cap \lambda_2 S \neq \emptyset$ that $\lambda_2 S \cap S \neq \emptyset$, i. e. $\lambda_2 S \subset S$. Repeating this reasoning we get at last $\lambda_n S \subset S$. But then the sequence $\lambda_1, \dots, \lambda_n$ can be replaced by the single element $\lambda = \lambda_n$ with the conditions (i) – (iii) be preserved (the condition (ii) turns into $\lambda S \subset S$).

Further, let $\alpha S \cap S \neq \emptyset \neq S\alpha \cap S$. Then an element $x$ from Lemma 3 has the least length in $\alpha S \cap S = \alpha S = \alpha F^1 w \cap \alpha w F^1$ accordingly to (iii). Hence $x = \alpha w$. Now setting $u = \alpha t$ $(t \in S)$, we can rewrite the conclusion of Lemma 3 in the form:

$$\begin{array}{c} \textit{for every } t \in S \textit{ there is } \lambda \in F^1 \textit{ such that} \\ \begin{array}{lrcl} \text{(a)} & \lambda S & \subset & S, \\ \text{(b)} & t & = & w\lambda. \end{array} \end{array} \tag{2}$$

Evidently, here $\lambda$ is defined uniquely by given $t$.

Now we can finish the proof of Lemma. Let $w\lambda \in F^1 w$. Then $w\lambda \in F^1 w \cap w F^1 = S$. Applying (2) to $t = w\lambda$, we obtain $\lambda S \subset S$. □

Every principal quasi-ideal can be written in the form $S = \langle w^n\rangle_q$ where $w$ is primitive and $n \geq 1$. If $n = 1$, $S$ is free (Theorem 1) and hence c.d.$(S) = 1$ (see, e. g. [2]). Therefore we suppose further on that $n \geq 2$. We shall show that c.d.$(S) > 1$, but the proof depends on if the word $w$ can be presented in the form $aba$ or not.

**Theorem 2.** *Let $S = \langle w^n\rangle_q \subset F$, $n \geq 2$, $w$ is primitive and $w = aba$ for some $a, b \in F$. Then c.d.$(S) > 1$.*

*Proof.* Set $\lambda = baw^{n-1}$. Then

$$w^n \lambda = w^{n-1} ababaw^{n-1} = w^{n-1} abw^n \in F^1 w^n$$

Show that $\lambda S \not\subset S$. Indeed, let $t \in S$ and $\lambda t \in S$. Then $\lambda t = w^n f$ for some $f \in F^1$, i. e. $baw^{n-1} t = abaw^{n-1} f$. From here $ba = ab$, so $a = c^p, b = c^q, w = c^{2p+q}$ $(c \in F)$ in contradiction with primitivity of $w$.

Therefore the conclusion of Lemma 4 is not valid and c.d.$(S) > 1$. □

Now consider the second kind of quasi-ideals.

**Lemma 5.** *Let a primitive word $w$ cannot be written in the form $w = aba$, $a, b \in F$. Then*

$$\langle w\rangle_q = wF^1 w \cup \{w\}.$$

*Proof.* Let $t \in \langle w\rangle_q \setminus (wF^1w \cup \{w\})$. Then $t = uw = wv$ and $u \neq 1 \neq v$ since $t \neq w$. Hence $u = xy$, $v = yx$, $w = (xy)^k x$ $(x, y \in F^1)$. Consider various values of $k$.

1) $\underline{k = 0}$. Then $w = x$ and $t = uw = wyw \in wF^1w$, what is impossible.

2) $\underline{k = 1}$. Then $w = xyx$ and $x = 1$ because of primitivity. Hence $t = uw = w^2 \in wF^1w$; contradiction.

3) $\underline{k > 1}$. Then $w = x\cdot y(xy)^{k-1}\cdot x$. Again $x = 1$ and $w = y^k$ contrary to primitivity. □

**Remark.** The converse is true too: if $w = aba$ then $ababa \in \langle w\rangle_q \setminus (wF^1w \cup \{w\})$ whence $\langle w\rangle_q \neq wF^1w \cup \{w\}$.

**Lemma 6.** *Let $w$ is the same as in Lemma 5. Then the semigroup $T_n = \langle w^n\rangle_q \cup \langle w\rangle$ is free for all $n \geq 1$.*

Proof is fulfilled by induction on $n$. For $n = 1$ the assertion follows from Theorem 1 since $T_1 = \langle w\rangle_q$.

Let $T_n$ is free. Accordingly to Corollary 1

$$\begin{aligned} T_{n+1} &= w^n\langle w\rangle_q w^n \cup \langle w\rangle = w(w^{n-1}\langle w\rangle_q w^{n-1} \cup \langle w\rangle)w \cup \{w, w^2\} \\ &= wT_nw \cup \{w, w^2\} = wT_n^1w \cup \{w\}. \end{aligned}$$

Since $T_n$ is free and $T_{n+1} \subset T_n$, this equality and Lemma 5 imply $T_{n+1}$ be coinciding with the quasi-ideal generating by $w$ in $T_n$. By Theorem 1 $T_{n+1}$ is free. □

**Theorem 3.** *Let a primitive word $w$ cannot be written in the form $w = aba$, $a, b \in F$. Then c.d.$\langle w^n\rangle_q > 1$ for $n \geq 2$.*

*Proof.* We use the fact that every proper subsemigroup $S \subset F$ of finite defect (i.e. $|F \setminus S| < \infty$) has c.d. $> 1$ ([5], Example 3.5). It follows immediately from here that c.d.$\langle w^n\rangle_q > 1$ $(n \geq 2)$ since $1 \leq |T_n \setminus \langle w^n\rangle_q| < n$. □

Joining Theorems 2 and 3 we obtain finally:

**Theorem 4.** *A principal quasi-ideal of a free non-commutative semigroup has c.d. 1 if and only if it is free.* □

## References

[1] K. S. Brown. *"Cohomology of Groups"*. Springer-Verlag, 1982.

[2] H. Cartan, S. Eilenberg. *"Homological Algebra"*. Princeton, 1956.

[3] G. Lallement. *"Semigroups and Combinatorial Applications"*. Wiley & Sons, NY, 1979.

[4] B. V. Novikov. *Semigroups of cohomological dimension one.* J. Algebra **204**(1998), no. 2, 386–393

[5] B. V. Novikov. *Ideals of free semigroups having cohomological dimension 1.* Semigroup Forum, **65**(2002), N2, 225–232.

[6] O. Steinfeld. *"Quasi-ideals in rings and semigroups".* Akadémiai Kiado, Budapest, 1978.

## Contact information

**B. V. Novikov** Saltovskoye shosse 258, apt. 20, 61178 Kharkov, Ukraine
*E-Mail:* boris.v.novikov@univer.kharkov.ua

Received by the editors: 10.02.2003.

Algebra and Discrete Mathematics RESEARCH ARTICLE
Number 2. **(2003).** pp. 93–102


# Uniform ball structures

**I. V. Protasov**

Communicated by V. M. Usenko

Abstract. A ball structure is a triple $\mathbb{B} = (X, P, B)$, where $X, P$ are nonempty sets and, for all $x \in X$, $\alpha \in P$, $B(x, \alpha)$ is a subset of $X, x \in B(x, \alpha)$, which is called a ball of radius $\alpha$ around $x$. We introduce the class of uniform ball structures as an asymptotic counterpart of the class of uniform topological spaces. We show that every uniform ball structure can be approximated by metrizable ball structures. We also define two types of ball structures closed to being metrizable, and describe the extremal elements in the classes of ball structures with fixed support $X$.

Following [2], by *ball structure* we mean a triple $\mathbb{B} = (X, P, B)$, where $X, P$ are nonempty sets and, for any $x \in X$, $\alpha \in P$, $B(x, \alpha)$ is a subset of $X$ which is called *a ball of radius* $\alpha$ around $x$. It is supposed that $x \in B(x, \alpha)$ for all $x \in X$, $\alpha \in P$. A set $X$ is called a *support* of $\mathbb{B}, P$ is called a *set of radiuses.*

Let $\mathbb{B}_1 = (X_1, P_1, B_1)$, $\mathbb{B}_2 = (X_2, P_2, B_2)$ be ball structures, $f : X_1 \longrightarrow X_2$ We say that $f$ is a $\succ$*-mapping* if, for every $\beta \in P_2$, there exists $\alpha \in P_1$ such that

$$B_2(f(x), \beta) \subseteq f(B_1(x, \alpha))$$

for every $x \in X_1$. If there exists a surjective $\succ$-mapping $f : X_1 \longrightarrow X_2$, we write $\mathbb{B}_1 \succ \mathbb{B}_2$.

A mapping $f : X_1 \longrightarrow X_2$ is called a $\prec$*-mapping* if, for every $\alpha \in P_1$,there exists $\beta \in P_2$ such that

$$f(B_1(x, \alpha)) \subseteq B_2(f(x), \beta)$$

**2001 Mathematics Subject Classification:** *03E99, 54A05, 54E15.*
**Key words and phrases:** *ball structure, metrizability.*

for every $x \in X$. If there exists an injective $\prec$-mapping $f : X_1 \longrightarrow X_2$, we write $\mathbb{B}_1 \prec \mathbb{B}_2$.

A bijection $f : X_1 \longrightarrow X_2$ is called an *isomorphism* between $\mathbb{B}_1$ and $\mathbb{B}_2$ if $f$ is a $\succ$-mapping and $f$ is a $\prec$-mapping.

Let $\mathbb{B}_1 = (X_1, P_1, B_1)$, $\mathbb{B}_2 = (X_2, P_2, B_2)$ be ball structures with common support $X$. We say that $\mathbb{B}_1 \subseteq \mathbb{B}_2$ if the identity mapping id: $X \longrightarrow X$ is a $\prec$-mapping from $\mathbb{B}_1$ to $\mathbb{B}_2$. If $\mathbb{B}_1 \subseteq \mathbb{B}_2$ and $\mathbb{B}_2 \subseteq \mathbb{B}_1$, we write $\mathbb{B}_1 = \mathbb{B}_2$.

A property $\mathcal{P}$ of ball structures is called a ball property if a ball structure $\mathbb{B}$ has a property $\mathcal{P}$ provided that $\mathbb{B}$ is isomorphic to some ball structure with property $\mathcal{P}$. Now we describe four basic ball properties.

Let $\mathbb{B} = (X, P, B)$ be a ball structure. For any $x \in X$, $\alpha \in P$ put

$$B^*(x, y) = \{y \in X : x \in B(y, \alpha)\}.$$

A ball structure $\mathbb{B}^* = (X, P, B)$ is called *dual* to $\mathbb{B}$. Note that $\mathbb{B}^{**} = \mathbb{B}$.

A ball structure $\mathbb{B} = (X, P, B)$ is called *symmetric* if $\mathbb{B} = \mathbb{B}^*$.

A ball structure $\mathbb{B} = (X, P, B)$ is called *multiplicative* if, for any $\alpha, \beta \in P$, there exists $\gamma(\alpha, \beta) \in P$ such that

$$B(B(x, \alpha), \beta) \subseteq B(x, \gamma(\alpha, \beta))$$

for every $x \in X$. Here

$$B(Y, \alpha) = \bigcup_{y \in Y} B(y, \alpha), \quad Y \subseteq X, \quad \alpha \in P.$$

Let $\mathbb{B} = (X, P, B)$ be a ball structure, $x, y \in X$, We say that $x, y$ are *connected* if there exists $\alpha \in P$ such that $x \in B(y, \alpha)$, $y \in B(x, \alpha)$. A subset $Y \subseteq X$ is called connected if any two elements from $Y$ are connected. A ball structure $\mathbb{B}$ is called *connected* if its support is connected. If $\mathbb{B}$ is symmetric and multiplicative, then connectivity is an equivalence on $X$, so $X$ disintegrates into connected components.

For an arbitrary ball structure $\mathbb{B} = (X, P, B)$ we define a preodering $\leq$ on the set $P$ by the rule $\alpha \leq \beta$ if and only if $B(x, \alpha) \subseteq B(x, \beta)$ for every $x \in X$. A subset $P' \subseteq P$ is called *cofinal* if, for every $\alpha \in P$, there exists $\beta \in P'$ such that $\alpha \leq \beta$. A *cofinality* $cf\mathbb{B}$ of $\mathbb{B}$ is the minimal cardinality of cofinal subsets of $P$.

Let $(X, d)$ be a metric space, $\mathbb{R}^+ = \{x \in \mathbb{R} : x \geq 0\}$. Given any $x \in X$, $r \in \mathbb{R}^+$, put

$$B_d(x, r) = \{y \in X : d(x, y) \leq r\}.$$

A ball structure $(X, \mathbb{R}^+, B_d)$ is denoted by $\mathbb{B}(X, d)$. We say that a ball structure $\mathbb{B}$ is *metrizable* if $\mathbb{B}$ is isomorphic to $\mathbb{B}(X, d)$ for some metric space $(X, d)$. We shall use the following metrizability criterion [2].

*A ball structure $\mathbb{B}$ is metrizable if and only if $\mathbb{B}$ is symmetric, multiplicative, connected and $cf \ \ \mathbb{B} \leq \aleph_0$.*

A ball structure is called *uniform* if it is symmetric and multiplicative.

In §1 we define a wide spectrum of examples of uniform ball structures related to groups and filters. In §2 we introduce some ball operations which give new uniform ball structures from a pregiven family of uniform ball structures. It is well known [1], that every uniform topological space can be approximated by pseudometrizable spaces. In §3 we prove a ball analogue of such an approximation. In §3 – 4 we introduce two types of ball structures (inductively metrizable and submetrizable) close to being metrizable. In §5 we describe extremal by inclusion elements in the classes of ball structures with fixed support.

## §1 Examples

Let $G$ be an infinite group with the identity $e, \gamma$ be an infinite cardinal, $\gamma < |G|$. Denote by $\Im_e(G, \gamma)$ the family of all subsets of $G$ of cardinality $< \gamma$ containing $e$. Given any $g \in G$, $F \in \Im_e(G)$, put

$$B_l(g, F) = Fg, \quad B_r(g, F) = gF.$$

The ball structures

$$(G, \Im_e(G, \gamma), B_l), \quad (G, \Im_e(G, \gamma), B_r)$$

will be denoted by $B_l(G, \gamma)$, $B_r(G, \gamma)$. Note that the mapping $g \longmapsto g^{-1}$ is an isomorphism between $B_l(G)$ and $B_r(G)$. In the case $\gamma = \aleph_0$ we write $B_l(G)$ and $B_r(G)$ instead of $B_l(G, \gamma)$ and $B_r(G, \gamma)$. It easy to see that $B_l(G)$=$B_r(G)$ if and only if the set $\{x^{-1}gx : x \in G\}$ is finite for every $g \in G$.

By metrizability criterion, $B_l(G, \gamma)$ is metrizable if and only if $\gamma = |G|$ and $cf \ \gamma = \aleph_0$. In particular, $B_l(G)$ is metrizable if and only if $G$ is countable.

Let $X$ be a set and let $\varphi$ be a filter on $X$. For any $x \in X$, $F \in \varphi$, put

$$B(x, F) = \begin{cases} X \backslash F, & \text{if } x \notin F; \\ \{x\}, & \text{if } x \in F; \end{cases}$$

and denote by $\mathbb{B}(X, \varphi)$ the ball structure $(X, \varphi, B)$. Note that $\mathbb{B}(X, \varphi)$ is connected if and only if either $\bigcap \varphi = \emptyset$ or $|X| = 1$. Hence, $\mathbb{B}(X, \varphi)$ is metrizable if and only if either $|X| = 1$ or $\bigcap \varphi = \emptyset$ and $\varphi$ has a countable base.

Now we define a wide class of ball structures containing all ball structures of filters and almost all ball structures of groups.

Let $X$ be a set and let $\mathcal{P}$ be a family of partitions of $X$. For any $x, y \in X$ and $P \in \mathcal{P}$, denote by $B(x, P)$ the set $\{y \in X : x, y$ are in the same cell of the partition $P\}$. A ball structure $(X, \mathcal{P}, B)$ is denoted by $\mathbb{B}(X, \mathcal{P})$. Clearly, $B(X, \mathcal{P})$ is symmetric. Given any $P_1, P_2 \in \mathcal{P}$, we say that $P_2$ is an enlargement of $P_1$ if $B(x, P_1) \subseteq B(x, P_2)$ for each $x \in X$. A ball structure $\mathbb{B}(X, \mathcal{P})$ is multiplicative if and only if, for any $P_1, P_2 \in \mathcal{P}$, there exists $P \in \mathcal{P}$ such that $P$ is an enlargement of $P_1$ and $P_2$.

A ball structure $\mathbb{B}$ is called *cellular* if $\mathbb{B}$ is isomorphic to $\mathbb{B}(X, \mathcal{P})$ for some set $X$ and some family $\mathcal{P}$ of partitions of $X$. Given any ball structure $\mathbb{B} = (X, P, B)$, $x, y \in X$ and $\alpha \in P$, we say that $x, y$ are *$\alpha$-path connected* if there exists a sequence $x_0, x_1, ..., x_n$, $x_0 = x$, $x_n = y$ such that

$$x_{i+1} \in B(x_i, \alpha), \quad x_i \in B(x_{i+1}, \alpha)$$

for every $i \in \{0, 1, ..., n-1\}$. For any $x \in X$, $\alpha \in P$, put

$$B^{\square}(x, \alpha) = \{y \in X : x, y \text{ } are \text{ } \alpha - path \text{ } connected\}.$$

A ball structure $\mathbb{B}^{\square}(X, P, B^{\square})$ is called a *cellularization* of $\mathbb{B}$. By [2], a ball structure $\mathbb{B}$ is cellular if and only if $\mathbb{B} = \mathbb{B}^{\square}$. A metrizable ball structure $\mathbb{B}$ is cellular if and only if $\mathbb{B}$ is isomorphic to $\mathbb{B}(X, d)$ for some non-Archimedian metric space.

Every ball structure $\mathbb{B}(X, \varphi)$ of a filter $\varphi$ on $X$ is cellular. A ball structure $\mathbb{B}(G, \gamma)$ of a group $G$ is cellular if and only if either $\gamma > \aleph_0$ or $\gamma = \aleph_0$ and every finite subsets of $G$ generates a finite subgroup.

## §2 Constructions

Let $\{\mathbb{B}_\lambda = (X_\lambda, P, B_\lambda) : \lambda \in I\}$ be a family of ball structures with pairwise disjoint supports and common set of radiuses, $X = \bigcup_{\lambda \in I} X_\lambda$. For every $x \in X$, $x \in X_\lambda$ and every $\alpha \in P$, put $B(x, \alpha) = B_\lambda(x, \alpha)$. A ball structure $\mathbb{B} = (X, P, B)$ is called a *disjoint union* of the family $\{\mathbb{B}_\lambda : \lambda \in I\}$. Every uniform ball structure is a disjoint union of its connected components.

Let $\{\mathbb{B}_\lambda = (X, P_\lambda, B_\lambda) : \lambda \in I\}$ be a family of ball structures with common support. Suppose that, for any $\lambda_1, \lambda_2 \in I$, there exists $\lambda \in I$ such that $\mathbb{B}_{\lambda_1} \subseteq \mathbb{B}_\lambda$, $\mathbb{B}_{\lambda_2} \subseteq \mathbb{B}_\lambda$. For every $\lambda \in I$, choose a copy $P'_\lambda = f_\lambda(P_\lambda)$ of $P_\lambda$ such that the family $\{P'_\lambda : \lambda \in I\}$ is disjoint. Put $P = \bigcup_{\lambda \in I} P'_\lambda$. For all $x \in X$, $\beta \in P$, $\beta \in P'_\lambda$, put $B(x, \beta) = B_\lambda(x, f_\lambda^{-1}(\beta))$.

A ball structure $\mathbb{B} = (X, P, B)$ is called an *inductive limit* of the family $\{\mathbb{B}_\lambda : \lambda \in I\}$. Clearly, $\mathbb{B}_\lambda \subseteq \mathbb{B}$ for every $\lambda \in I$. If every $\mathbb{B}_\lambda$ is uniform, $\mathbb{B}$ is uniform.

Let $\mathbb{B} = (X, P, B)$ be a ball structure, $Y \subseteq X$. For any $y \in Y$, $\alpha \in P$, put $B_Y(y, \alpha) = B(y, \alpha) \bigcap Y$. A ball structure $\mathbb{B}_Y = (Y, P, B_Y)$ is called a substructure of $\mathbb{B}$. If $\mathbb{B}$ is uniform, then $\mathbb{B}_Y$ is uniform.

Let $\{\mathbb{B}_\lambda = (X_\lambda, P_\lambda, B_\lambda) : \lambda \in I\}$ be an arbitrary family of ball structures. By *box product* of this family we mean a ball structure

$$\prod_{\lambda \in I} \mathbb{B}_\lambda = (\prod_{\lambda \in I} X_\lambda, \prod_{\lambda \in I} P_\lambda, B),$$

where

$$B(x, p) = \{y \in \prod_{\lambda \in I} X_\lambda : pr_\lambda(y) \in B_\lambda(pr_\lambda(x), pr_\lambda(p)), \lambda \in I\}$$

for all

$$x \in \prod_{\lambda \in I} X_\lambda, p \in \prod_{\lambda \in I} P_\lambda.$$

If every ball structure $\mathbb{B}_\lambda$ is uniform, then $\prod_{\lambda \in I} \mathbb{B}_\lambda$ is uniform. Note also that

$$\mathbb{B}_\gamma \prec \prod_{\lambda \in I} \mathbb{B}_\lambda, \quad \prod_{\lambda \in I} \mathbb{B}_\lambda \succ \mathbb{B}_\gamma$$

for every $\gamma \in I$.

Let $\mathbb{B} = (X, P, B)$ be a ball structure. A subset $Y \subseteq X$ is called *bounded* if there exist $x \in X$, $\alpha \in P$ such that $Y \subseteq B(x, \alpha)$. We say that $\mathbb{B}$ is *bounded* if its support is bounded. Let $\mathbb{B}$ be a connected uniform ball structure, $x_0 \in X, Y \subseteq X$. Then $Y$ is bounded if and only if there exists $\alpha \in P$ such that $Y \subseteq \mathbb{B}(x_0, \alpha)$. A box product of an arbitrary family of bounded ball structures is bounded. It is metrizable if and only if every $\mathbb{B}_\lambda, \lambda \in I$ is metrizable and all but finitely many of them are bounded.

We define also two modifications of box products. Let $\mathcal{F}$ be a family of all finite subsets of $I$. The first modification is

$$\overset{\vee}{\prod_{\lambda \in I}} \mathbb{B}_\lambda = (\prod_{\lambda \in I} X_\lambda, \Im \times \prod_{\lambda \in I} P_\lambda, \check{B}),$$

where

$$\check{B}(x, (F, p)) = \{y \in \prod_{\lambda \in I} X_\lambda : pr_\lambda(y) \in B_\lambda(pr_\lambda(x), pr_\lambda(p)) \quad \mathit{for\ \ every}\ \lambda \notin F\}.$$

The second modification is

$$\prod_{\lambda\in I}^{\wedge}\mathbb{B}_\lambda=(\prod_{\lambda\in I}X_\lambda,\Im\times\prod_{\lambda\in I}P_\lambda,\hat{B}),$$

where

$$\hat{B}(x,(F,p))=\{y\in\prod_{\lambda\in I}X_\lambda : pr_\lambda(y)\in B_\lambda(pr_\lambda(x),pr_\lambda(p)),$$
$$\lambda\in F \ \ and \ \ pr_\lambda(x)=pr_\lambda(y),\quad \lambda\notin F\}.$$

Clearly,

$$\prod_{\lambda\in I}^{\wedge}\mathbb{B}_\lambda\subseteq\prod_{\lambda\in I}\mathbb{B}_\lambda\subseteq\prod_{\lambda\in I}^{\vee}\mathbb{B}_\lambda.$$

## §3 Approximations

A ball structure $\mathbb{B}$ is called *pseudometrizable* if $\mathbb{B}$ is a disjoint union of metrizable ball structures.

**Theorem 3.1.** *Every uniform ball structure $\mathbb{B}=(X,P,B)$ is an inductive limit of some family of pseudometrizable ball structures.*

*Proof.* We may suppose that $B(X,\alpha)=B^*(X,\alpha)$ for all $x\in X$, $\alpha\in P$. Denote by $I$ the family of all subsets of $P$ of the form $\{\alpha_n\in P : n\in\omega\}$ such that $\alpha_n\le\alpha_{n+1}$, $n\in\omega$ and, for any $n,m\in\omega$, there exists $k(n,m)\in\omega$ such that

$$B(B(x,n),m)\subseteq B(x,k(n,m))$$

for every $x\in X$. For every $\lambda\in I$, $\lambda=\{\alpha_n : n\in\omega\}$, put $P_\lambda=\{\alpha_n : n\in\omega\}$. By metrizability criterion, every connected component of $\mathbb{B}_\lambda=(X,P_\lambda,B_\lambda)$, $B_\lambda(x,\alpha_n)=B(x,\alpha_n)$ is metrizable, so $\mathbb{B}_\lambda$ is pseudometrizable . It is easy to check that $\mathbb{B}$ is an inductive limit of the family $\{\mathbb{B}_\lambda : \lambda\in I\}$. $\square$

A ball structure $\mathbb{B}$ is called *inductively metrizable* if $\mathbb{B}$ is an inductive limit of metrizable ball structures.

**Theorem 3.2.** *For every uniform ball structure $\mathbb{B}=(X,P,B)$ the following statements are equivalent*

*(i) $\mathbb{B}$ is inductively metrizable;*

*(ii) there exists a metric space $(X,d)$ such that $\mathbb{B}(X,d)\subseteq\mathbb{B}$;*

*(iii) there exists a subset $P'\subseteq P, |P'|\le\aleph_0$ and $x_0\in X$ such that $\bigcup_{\alpha\in P'}B(x_0,\alpha)=X$.*

*Proof.* The implications $(i)\Longrightarrow(ii)\Longrightarrow(iii)$ are trivial.

$(iii) \Longrightarrow (i)$. We may suppose that $P' = \{\alpha_n : n \in \omega\}$, $\alpha_n \leq \alpha_{n+1}$, $n \in \omega$ and, for any $n, m \in \omega$ there exists $k(n,m) \in \omega$ such that

$$B(B(x,n),m) \subseteq B(x, k(n,m))$$

for every $x \in X$. Then $\mathbb{B}' = (X, P', B')$, $B'(x,\alpha) = B(x,\alpha)$, $\alpha \in P'$ is a metrizable ball structure. Consider the family $\Im$ of all metrizable ball structures on $X$ such that $\mathbb{B}' \subseteq \mathbb{B}''$ for every $\mathbb{B}'' \in \Im$. Clearly, $\mathbb{B}$ is an inductive limit of $\Im$. □

By Theorem 3.2, a ball structure $\mathbb{B}(X,\varphi)$ of a filter $\varphi$ on $X$, $|X| > 1$ is inductively metrizable if and only if there exists a countable subset $\psi$ of $\varphi$ such that $\bigcap \psi = \emptyset$. A ball structure $\mathbb{B}(G,\gamma)$ is inductively metrizable if and only if it is metrizable.

## §4 Submetrizability

A uniform ball structure $\mathbb{B} = (X, P, B)$ is called *submetrizable* if there exists an unbounded metrizable ball structure $\mathbb{B}' = (X, P', B')$ such that $\mathbb{B} \subseteq \mathbb{B}'$.

Let $\mathbb{B} = (X, P, B)$ be a ball structure, $f : X \longrightarrow \mathbb{R}$. We say that $f$ is a function of *bounded oscilation* (with respect to $\mathbb{B}$) if, for every $\alpha \in P$, there exists a natural number $n(\alpha)$ such that

$$diam \;\; f(B(x,\alpha)) \leq n(\alpha)$$

for every $x \in X$, where $diam \;\; A = sup\{|a-b| : a, b \in A\}$. Clearly, every bounded function is of bounded oscilation.

**Theorem 4.1.** *For every uniform ball structure* $\mathbb{B} = (X, P, B)$ *the following statements are equivalent*

*(i)* $\mathbb{B}$ *is submetrizable;*

*(ii) there exists an unbounded function* $f : X \longrightarrow \mathbb{R}$ *of bounded oscilation.*

*Proof.* $(i) \Longrightarrow (ii)$. Let $(X, d)$ be an unbounded metric space such that $\mathbb{B} \subseteq \mathbb{B}(X, d)$. Fix an arbitrary point $x_0 \in X$ and put $f(x) = d(x, x_0)$. Since $(X, d)$ is unbounded, $f$ is unbounded. If $x, y \in X$, then

$$|f(x) - f(y)| = |d(x_0, x) - d(x_0, y)| \leq d(x, y).$$

Hence, $f$ is of bounded oscilation on $(X, d)$. Since $\mathbb{B} \subseteq \mathbb{B}(X, d)$, $f$ is of bounded oscilation with respect to $\mathbb{B}$.

$(i) \Longrightarrow (ii)$. For all $x \in X$, $n \in \omega$, put

$$B'(x, n) = \{y \in X : |f(x) - f(y)| \leq n\}.$$

Clearly, the ball structure $\mathbb{B}' = (X, \omega, B')$ is symmetric, multiplicative, connected and $cf\mathbb{B}' = \aleph_0$. Hence, $\mathbb{B}$ is metrizable. To show that $\mathbb{B} \subseteq \mathbb{B}'$, fix an arbitrary $\alpha \in P$, choose $n(\alpha)$ such that diam $f(B(x,\alpha)) \leq n(\alpha)$ for every $x \in X$. Then $B(x,\alpha) \subseteq B'(x, n(\alpha))$ for every $x \in X$. □

A connected uniform ball structure $\mathbb{B} = (X, P, B)$ is called *ordinal* if there exists a cofinal well ordered by $\leq$ subset of $P$. Clearly, every metrizable ball structure is ordinal.

**Theorem 4.2.** *For every ordinal ball structure $\mathbb{B} = (X, P, B)$, the following statements are equivalent*

*(i) $\mathbb{B}$ is metrizable;*

*(ii) $\mathbb{B}$ is submetrizable;*

*Proof.* The implication $(i) \Longrightarrow (ii)$ is trivial.

$(ii) \Longrightarrow (i)$. We may suppose that $P$ is well ordered. Assume that $\mathbb{B}$ is not metrizable so cf $P > \aleph_0$. By Theorem 4.1, there exists an unbounded function $f : X \longrightarrow \mathbb{R}$ of bounded oscilation. Choose a countable subset $X' \subseteq X$ such that $f(X')$ is unbounded in $\mathbb{R}$. Since cf $P > \aleph_0$, there exists $X_0 \in X$, $\alpha \in P$ such that $X' \subseteq B(x_0, \alpha)$. We get a contradiction to the definition of function of bounded oscilation. □

The next results give us examples of nonmetrizable submetrizable ball structures.

**Theorem 4.3.** *If a group $G$ has a normal subgroup $H$ of countable index, then $\mathbb{B}_l(G)$ is submetrizable.*

*Proof.* Let $\Im$ be a family of all finite subsets of $G$ containing the identity $e$ of $G$. Given any $g \in G$, $F \in \Im$ put

$$B'_l(g, FN) = FNg.$$

Denote by $\Im'$ the family of all subsets $Y \subseteq G$ of the form $Y = FN$, $F \in \Im$. Then $\mathbb{B}' = (G, \Im', B')$ is metrizable ball structure and $\mathbb{B} \subseteq \mathbb{B}'$. □

**Theorem 4.4.** *Let $\varphi$ be a filter on an infinite set $X$, $Y = \bigcap \varphi$. Then $\mathbb{B}(X, \varphi)$ is submetrizable if and only if one of the following statements holds*

*(i) $Y$ is infinite;*

*(ii) $Y$ is finite and there exists a filter $\psi$ on $X$ with a countable base such that $\varphi \subseteq \psi$ and $Y \notin \psi$.*

*Proof.* If $(i)$ holds, choose a countable subset $\{y_n : n \in \omega\}$ of $Y$ and put $f(y_n) = n$, $n \in \omega$ and $f(x) = 0$ for every $x \in X \setminus \{y_n : n \in \omega\}$. Then $f : X \longrightarrow \mathbb{R}$ is an unbounded function of bounded oscilation. Apply Theorem 4.1.

If $(ii)$ holds, choose a base $\{F_n : n \in \omega\}$ of $\psi$ such that $F_{n+1} \subset F_n$ for every $n \in \omega$. For every $x \in X$, put

$$f(x) = \begin{cases} n, & \text{if } x \in F_n \setminus F_{n+1}; \\ 0, & \text{otherwise.} \end{cases}$$

Since $f$ is an unbounded function of bounded oscilation, we can apply Theorem 4.1.

Assume that $Y$ is finite and $\mathbb{B}(X, \varphi)$ is submetrizable. By Theorem 4.1, there exists an unbounded function $f : X \longrightarrow \mathbb{R}$ of bounded oscilation. For every $n \in \omega$, put $X_n = \{x \in X : |f(x)| > n\}$. Let $\psi$ be a filter on $X$ with the base $X_n : n \in \omega$. Take an arbitrary $F \in \varphi$. Since $f$ is of bounded oscilation, $f$ is bounded on $X \setminus F$. If follows that $X_m \subseteq F$ for some $m \in \omega$, so $\varphi \subseteq \psi$. □

**Theorem 4.5.** *Let a ball structure $\mathbb{B}$ is a disjoint union of the family $\{\mathbb{B}_\lambda = (X_\lambda, P, B_\lambda) : \lambda \in I\}$ of uniform ball structures. Then $\mathbb{B}$ is submetrizable if and only if either $I$ is infinite of there exists $\lambda \in I$ such that $\mathbb{B}_\lambda$ is submetrizable.*

*Proof.* If $I$ is infinite, choose a countable subset $\{\lambda_n : n \in \omega\}$ of $I$, put $f|X_{\lambda_n} \equiv n$, $n \in \omega$ and $f|X_\lambda \equiv 0$, $\lambda \notin \{\lambda_n : n \in \omega\}$. Apply Theorem 4.1.

If $\mathbb{B}_\lambda$ is submetrizable, we fix an unbounded function $f : X_\gamma \longrightarrow \mathbb{R}$ of bounded oscilation and put $f|X_\gamma \equiv 0$ for every $\gamma \neq \lambda$. Apply Theorem 4.1.

On the other hand, assume that $I$ is finite and $\mathbb{B}_\lambda$ is not submetrizable for every $\lambda \in I$. By Theorem 4.1, $\mathbb{B}$ is not submetrizable. □

## §5 Extremalities

Fix a set $X$ and denote by $\mathcal{B}(X)$ the class of all ball structures with the support $X$. Clearly, every bounded ball structure is a maximal by inclusion $\subseteq$ element of $\mathcal{B}(X)$, and any two bounded ball structures on $X$ coincide. On the other hand, the discrete ball structure $(X, \{p\}, B)$, $B(x, p) = \{x\}$, $x \in X$ is a minimal by inclusion element of $\mathcal{B}(X)$. We can easily avoid these trivialities considering some natural subclasses of $\mathcal{B}(X)$.

An unbounded uniform ball structure $\mathbb{B}$ is called *prebounded* if $\mathbb{B} \subseteq \mathbb{B}'$, $\mathbb{B}'$ is unbounded and uniform, implies $\mathbb{B} \subseteq \mathbb{B}'$.

**Theorem 5.1.** *For every unbounded uniform ball structure $\mathbb{B} = (X, P, B)$, there exists a prebounded ball structure $\mathbb{B}'$ such that $\mathbb{B} \subseteq \mathbb{B}'$. Every prebounded ball structure is not submetrizable.*

*Proof.* The first statement follows directly from Zorn lemma and the construction of inductive limit. To prove the second statement, we take an unbounded metric space $(X, d)$ and define a new metric $d'$ on $X$

such that $(X, d')$ is unbounded and $\mathbb{B}(X, d) \subset \mathbb{B}(X, d')$. Fix an arbitrary element $x_0 \in X$. For every $x \in X$, choose $n \in \omega$ such that

$$n^2 \leq d(x) < (n+1)^2$$

and put $f(x) = n$. For every $x \in X$, $m \in \omega$, denote

$$B'(x, m) = \{y \in X : |f(x) - f(y)| \leq m\}.$$

Consider the ball structure $B' = (X, \omega, B')$. By metrizability criterion, $\mathbb{B}'$ is metrizable. Clearly, $\mathbb{B} \subset \mathbb{B}'$. $\square$

**Theorem 5.2.** *Let $X$ be an infinite set and let $\varphi = \{F \subseteq X : X \backslash F \text{ is finite}\}$. Then $\mathbb{B}(X, \varphi) \subseteq \mathbb{B}$ for every connected uniform ball structure $\mathbb{B} = (X, P, B)$.*

*Proof.* Take an arbitrary $F \in \varphi$. Since $\mathbb{B}$ is connected and uniform, the finite subset $X \backslash F$ is bounded. Choose $\alpha \in P$ such that $X \backslash F \subseteq B(x, \alpha)$ for every $x \in X \backslash F$. Then *id* $B(x, F) \subseteq B(x, \alpha)$ for every $x \in X$. Hence, $\mathbb{B}(X, \varphi) \subseteq \mathbb{B}$. $\square$

**Acknowledgements.** The author express his sincere thanks to Michael Zarichniĭ for the question about asymptotic counterpart of uniform topological spaces.

## References

[1] Burbaki N. *Obschaya topologiya. Ispol'zovanie veschestvennih chisel v obschej topologii*, M., "Nauka", 1975, in russian.

[2] Protasov I.V. *Mertizable ball structures*, Algebra and Discrete Math., 1 (2002).

CONTACT INFORMATION

**I. V. Protasov** Department Cybernetics, Kyiv State University, Volodimirska 64, Kyiv 01033, Ukraine
*E-Mail:* kseniya@profit.net.ua

Received by the editors: 31.01.2003.

Algebra and Discrete Mathematics
Number 2. **(2003).** pp. 103–110 RESEARCH ARTICLE


# An additive divisor problem in $\mathbb{Z}[i]$

**O. V. Savasrtu and P. D. Varbanets**

Communicated by V. V. Kirichenko

Abstract. Let $\tau(\alpha)$ be the number of divisors of the Gaussian integer $\alpha$. An asymptotic formula for the summatory function $\sum\limits_{N(\alpha)\le x} \tau(\alpha)\tau(\alpha+\beta)$ is obtained under the condition $N(\beta) \le x^{3/8}$. This is a generalization of the well-known additive divisor problem for the natural numbers.

## 1. Introduction

In 1927 A.E. Ingham [1] obtained the asymptotic formula for the number of solutions $I(x)$ the diophantic equation

$$u_1u_2 - v_1v_2 = 1$$

under conditions: $u_1, u_2, v_1, v_2 \in \mathbb{N}$, $u_1u_2 \le x$.

Obviously

$$I(x) = \sum_{n\le x} \tau(n)\tau(n+1),$$

where $\tau(n) = \sum\limits_{n=ab} 1$ denote the number of ways $n$ may be written as a product of two natural numbers.

Ingham proved that

$$I(x) = \frac{6}{\pi^2} x \log^2 x + O(x \log x).$$

T. Estermann [2] improved this result in form

$$I(x) = xP_2(\log x) + E(x),$$

**2001 Mathematics Subject Classification:** *11N37; 11R42.*
**Key words and phrases:** *additive divisor problem; asymptotic formula.*

where $P_2(u)$ is a polynom $a_0u^2 + a_1u + a_2$ and $E(x)$ is an error term.

Estermann gave $E(x) \ll x^{\theta+\varepsilon}$, $\theta = \frac{11}{12}$. The exponent $\theta$ was subsequently improved to $\frac{5}{6}$ by D.R. Heath-Brown [3] and then to $\frac{2}{3}$ by J.-M. Deshouillers and H. Ivaniec [4]. In 1994 Y. Motohashi [5] employed powerful methods from the spectral theory of automorphic forms and obtained very precise result:

$$I(x) = \sum_{n\leq x} \tau(n)\tau(n+h) = x\sum_{i=0}^{2}(\log x)^i \sum_{j=0}^{2} c_{ij} \sum_{d|h} \frac{(\log d)^j}{d} + O\left(x^{\frac{2}{3}+\varepsilon}\right)$$

holds uniformly for $1 \leq h \leq x^{20/27}$.

The purpose of this paper is to build the asymptotic formula for sum

$$\sum_{\substack{\alpha\in\mathbb{Z}[i]\\ 0<N(\alpha)\leq x}} \tau(\alpha)\tau(\alpha+\beta)$$

where $\tau(\alpha) = \sum\limits_{\delta|\alpha} 1$ is a number of divisors of a Gaussian integer $\alpha$.

**Notations.** Denote by $\mathbb{Z}$ the ring of Gaussian integers. We write $N(\alpha) = a^2 + b^2$, $Sp(\alpha) = 2a$ for $\alpha = a + bi \in \mathbb{Z}[\alpha]$; $\varphi(\alpha) = N(\alpha)\prod\limits_{p|\alpha}(1 - N(p)^{-1})$, $p$ is prime divisor $\alpha$; $e(x) = \exp(2\pi i x)$ for the real number $x$; the Vinogradov symbol $f \ll g$ means $f = O(g)$; $\varepsilon$ is an arbitrary small positive number that is not necessarily the same at each occurrence; the constants implied by the $O$ (or $\ll$) – notation depend at most on $\varepsilon$.

## 2. Statement of Result

Let $\beta$ be Gaussian integer and $x$ be real positive number. By $I(x,\beta)$ we denote the number of solutions in Gaussian integers of the equation

$$\alpha_1\alpha_2 - \alpha_3\alpha_4 = \beta, \quad N(\alpha_1\alpha_1) \leq x.$$

**Theorem.** *For $N(\beta) \ll x^{3/8}$ and any $\varepsilon > 0$ the following formula*

$$I(x,\beta) = xP_2(\log x) + O(x^{\frac{7}{8}+\varepsilon})$$

*holds.*

Here $P_2(u) = A_0u^2 + A_1u + A_2$, $A_i = A_i(\beta)$, $i = 0,1,2$, moreover $A_i(\beta)$ are computable and $1 \ll A_i(\beta) \ll \tau(\beta)$, $A_0(\beta) > 0$.

## 3. Auxiliary Results

Let $\delta_0, \delta$ be the Gaussian rationals ($\delta_0, \delta \in \mathbb{Q}(i)$) not necessarily integers. Let for $Re(s) > 1$

$$\zeta(s,\delta,\delta_0) = \sum_{\substack{\omega\in\mathbb{Z}[i]\\ \omega\neq-\delta}} e\left(\frac{1}{2}Sp(\delta_0\omega)\right) N(\delta+\omega)^{-s}.$$

**Lemma 1 (see [5],lemmas 1 and 3).** *The function $\zeta(s,\delta,\delta_0)$ is entire function if $\delta_0 \notin \mathbb{Z}[i]$. For $\delta_0 \in \mathbb{Z}[i]$, $\zeta(s,\delta,\delta_0)$ is holomorphic except at $s=1$, where it has a simple pole and*

$$\zeta(s,\delta,0) = \frac{\pi}{s-1} + a_0(\delta) + a_1(\delta)(s-1) + \ldots$$

*where*

$$a_0(\delta) = \begin{cases} \pi E + 4L'(1,\chi_4) & \text{if } \delta \in \mathbb{Z}[i], \\ \pi E + 4L'(1,\chi_4) + \sum\limits_{\beta\in B} N(\delta+\beta) + b_0(\delta) & \text{if } 0 < N(\delta) < 1; \end{cases}$$

*$E$ is the Euler constant, $L'(s,\chi_4) = \frac{d}{ds}L(s,\chi_4)$, $L(s,\chi_4)$ is L-Dirichlet function with non-principal character mod 4; $b_0(\delta) = -4 + O\left(N^{1/2}(\delta)\right)$, $B$ denotes the set $\{0, \pm1, \pm i\}$. Moreover, the functional equation*

$$\pi^{-s}\Gamma(s)\zeta(s,\delta,\delta_0) = \pi^{-(1-s)}\Gamma(1-s)\zeta(1-s,-\delta_0,\delta)e\left(-\frac{1}{2}Sp(\delta_0\delta)\right) \quad (1)$$

*holds.*

Let $\alpha, \beta, \gamma \in \mathbb{Z}[i]$. We define the Kloosterman sum for the ring of Gaussian integer

$$K(\alpha,\beta;\gamma) = \sum_{\substack{\xi,\xi' \pmod{\gamma}\\ \xi\cdot\xi'\equiv1(\gamma)}} e\left(\frac{1}{2}Sp\left(\frac{\alpha\xi+\beta\xi'}{\gamma}\right)\right).$$

**Lemma 2.** *Let $\alpha, \beta, \gamma$ be Gaussian integers, $\gamma \neq 0$. Then the estimate*

$$|K(\alpha,\beta;\gamma)| \ll (N(\gamma)N((\alpha,\beta;\gamma)))^{1/2}\tau(\gamma) \quad (2)$$

*holds, (where $(\alpha,\beta;\gamma)$ is the greate common divisor of $\alpha,\beta,\gamma$). Moreover,*

$$K(\alpha,\beta;\gamma) = \sum_{\delta|(\alpha,\beta,\gamma)} N(\delta)K\left(1,\frac{\alpha\beta}{\delta^2};\frac{\gamma}{\delta}\right). \quad (3)$$

This lemma follow from a multiplicative property of $K(\alpha,\beta;\gamma)$ on $\gamma$ and the Bombieri estimate of an exponential sum on the algebraic curve over the finite field. The formula (3) is a generalized Kuznetsov's identity for Kloosterman sums.

**Lemma 3.** *Let* $\alpha_0,\gamma\in\mathbb{Z}[i]$, $(\alpha_0,\gamma)=\beta$, $N(\beta)<N(\gamma)$. *Then for* $N(\gamma)\ll x^{2/3+\varepsilon}$ *we have*

$$\sum_{\substack{\alpha\equiv\alpha_0(\gamma)\\ N(\alpha)\le x}}\tau(\alpha)=c_0(\alpha_0,\gamma)\frac{x}{N(\gamma)}\log\frac{x}{N(\beta)}+$$

$$+c_1(\alpha_0,\gamma)\frac{x}{N(\gamma)}+O\left(x^{1/2+\varepsilon}N(\gamma)^{-1/4}\right),$$

*where* $c_0(\alpha_0,\gamma)=\pi^2N(\beta)\varphi\left(\frac{\gamma}{\beta}\right)N^{-1}(\gamma)\tau(\beta)$,

$$c_1(\alpha_0,\gamma)=\pi^2\sum_{\delta|\beta}\left[2E-1+2\frac{L'(1,\chi_4)}{L(1,\chi_4)}+\sum_{p|\gamma/\delta}{}^{*}\log\frac{N(p)}{N(p)-1}\right]$$

$$\prod_{\gamma|\gamma/\delta}{}^{*}(1-N^{-1}(p)).$$

*Proof.* Without loss of generality we will consider only a case $(\alpha_0,\gamma)=1$.

We have for $c=1+\varepsilon$:

$$\sum_{\substack{\alpha\equiv\alpha_0(\gamma)\\ N(\alpha)\le x}}\tau(\alpha)-\sum_{\substack{\alpha=\alpha_0+\beta\gamma\\ \beta\in B}}\tau(\alpha)=$$

$$=\frac{1}{2\pi i}\int_{c-iT}^{c+iT}\left(F(s)-\sum_{\beta\in B}\frac{\tau(\alpha_0+\beta\gamma)}{N(\alpha_0+\beta\gamma)^s}\right)\frac{x^s}{s}ds+O\left(\frac{x^c}{TN(\gamma)}\right), \qquad (4)$$

where

$$F(s)=N(\gamma)^{-2s}\sum_{\substack{\alpha_1,\alpha_2 (\mathrm{mod}\ \gamma)\\ \alpha_1\alpha_2\equiv\alpha_0(\gamma)}}\zeta\left(s,\frac{\alpha_1}{\gamma},0\right)\zeta\left(s,\frac{\alpha_2}{\gamma},0\right)=\sum_{\substack{\alpha\equiv\alpha_0(\gamma)\\ \alpha\in\mathbb{Z}[i]}}\frac{\tau(\alpha)}{N(\alpha)^s}.$$

From lemma 1 we have the functional equation

$$F(s)=\frac{\pi^{2(2s-1)}}{N^{2s}(\gamma)}\frac{\Gamma^2(1-s)}{\Gamma^2(s)}\Psi(1-s),$$

where

$$\Psi(s)=\sum_{\omega}\frac{1}{N(\omega)^s}\sum_{\alpha\beta=\omega}\Phi(\alpha,\beta;\gamma),$$

$$\Phi(\alpha,\beta;\gamma)\sum_{\substack{\alpha_1,\alpha_2 (\mathrm{mod}\ \gamma)\\ \alpha_1\alpha_2\equiv\alpha_0(\gamma)}} e\left(\frac{1}{2}Sp\left(\frac{\alpha\alpha_1+\beta\alpha_2}{\gamma}\right)\right).$$

Moreover, $F(0)=0$ if $N(\gamma)>1$ and $\alpha\not\equiv 0(\mathrm{mod}\ \gamma)$.

By lemma 1 we obtain

$$G(s)=F(s)-\sum_{\beta\in B}\frac{\tau(\alpha_0+\beta\gamma)}{N(\alpha_0+\beta\gamma)^s}\ll\begin{cases} N(\gamma)^{-1+\varepsilon} & \text{if } Re(s)=1+\varepsilon,\\ N(\gamma)^{1/2+\varepsilon}T^3 & \text{if } Re(s)=-\frac{1}{4}.\end{cases} \tag{5}$$

Applying Phragmen-Lindelöf principle we infer

$$G(-\varepsilon+it)\ll N(\gamma)^{1/5+\varepsilon}T^{12/5+\varepsilon}\quad \text{for } |t|\le T.$$

To deal with integral in (4) we move the segment of integration to $Re(s)=-\varepsilon$.

By the theorem of residues we obtain

$$\sum_{\substack{\alpha\equiv\alpha_0(\gamma)\\ N(\alpha)\le x}}\tau(\alpha)=\mathrm{res}_{s=0}\left(G(s)\frac{x^s}{s}\right)+\mathrm{res}_{s=1}\left(G(s)\frac{x^s}{s}\right)+$$

$$+\frac{1}{2\pi i}\int\limits_{-\varepsilon-iT}^{-\varepsilon+iT}G(s)\frac{x^s}{s}ds++O(x^\varepsilon)+O\left(N(\gamma)^{1/5+\varepsilon}T^{12/5+\varepsilon}\right)+$$

$$+O\left(\frac{x^{1+\varepsilon}}{TN(\gamma)}\right). \tag{6}$$

Further,

$$\mathrm{res}_{s=0}\left(G(s)\frac{x^s}{s}\right)=\frac{\pi^2x\log x}{N(\gamma)}\prod_{\gamma|\alpha}(1-N(\gamma)^{-1})+$$

$$+\frac{\pi^2x}{N(\gamma)}\prod_{p|\alpha}(1-N(p)^{-1})\left[-1+2\left(E+\frac{L'(1,\chi_4)}{L(1,\chi_4)}+\sum_{p|\delta}\frac{\log N(p)}{N(p)-1}\right)\right], \tag{7}$$

$$\mathrm{res}_{s=0}\left(G(s)\frac{x^s}{s}\right)=\mathrm{res}_{s=0}\left(-\sum_{\beta\in B}\frac{\tau(\alpha_0+\beta\gamma)}{N(\alpha_0+\beta\gamma)^s}\frac{x^s}{s}\right)\ll N(\gamma)^\varepsilon.$$

Observe that by lemma 2

$$\sum_{\alpha\beta=\omega}|\Phi(\alpha,\beta;\gamma)|=\sum_{\alpha\beta=\omega}|K(\alpha,\beta\alpha_0;\gamma)|\ll N(\gamma)^{1/2}N((\omega,\gamma))^{1/2}\tau(\gamma)\tau(\omega).$$

Now by termwise integration and applying the Stirling formula for the gamma function and the method of stationary phase we get

$$\frac{1}{2\pi i}\int\limits_{-\varepsilon-iT}^{-\varepsilon+iT} G(s)\frac{x^s}{s}ds =$$

$$= \sum_{\substack{\omega \\ 0<N(\omega)\le Y}} \frac{\pi^2}{N(\omega)} \sum_{\alpha\beta=\omega} \Phi(\alpha,\beta;\gamma)\frac{y^{3/8}}{4\sqrt{2/\pi}} e\left(-\frac{1}{8}-\frac{1}{2\pi}y^{1/4}\right)\cdot$$

$$\cdot\left(1+O\left(y^{-1/8}\right)\right)+O\left(\frac{x^{1+\varepsilon}}{TN(\gamma)}\right)+O\left(x^{\varepsilon}\right)+$$

$$+O\left(\sum_{N(\omega)>Y} y^{-\varepsilon}T^{1+4\varepsilon}N(\gamma)^{1/2+\varepsilon}N((\omega,\gamma))^{1/2}\tau(\omega)N(\omega)^{-1}\right), \quad (8)$$

where $Y \le X = \left(\frac{4}{\pi}\right)^4 \frac{T^4N^2(\gamma)}{x}$, $y = \frac{\pi^4 xN(\omega)}{N^2(\gamma)}$.

The assertion of the lemma follow from (4),(6)–(8) if we put

$$T = x^{1/2}N(\gamma)^{-3/4},\ Y = x^{1/3}.$$

□

## 4. Proof of the theorem

We start the proof of our theorem by observing that

$$\tau(\alpha) = 2\#\{\gamma|\alpha; N(\gamma)\le x^{1/2}\} - \#\{\gamma|\alpha; N(\alpha)x^{-1/2}\le N(\gamma)\le x^{1/2}\}$$

whenever $\alpha$, $N(\alpha)\le x$.

Hence

$$\sum_{N(\alpha)\le x}\tau(\alpha)\tau(\alpha+\beta) =$$

$$= \sum_{N(\gamma)\le x^{1/2}}\left(2\left\{\sum_{\substack{\alpha\equiv\beta(\gamma)\\ N(\alpha-\beta)\le x}}\tau(\alpha)-1\right\}-\left\{\sum_{\substack{\alpha\equiv\beta(\gamma)\\ N(\alpha-\beta)\le N(\gamma)x^{1/2}}}\tau(\alpha)-1\right\}\right) =$$

$$= 2\sum_{N(\gamma)\le x^{1/2}}\sum_{\substack{\alpha\equiv\beta(\gamma)\\ N(\alpha)\le x}}\tau(\alpha) - \sum_{N(\gamma)\le x^{1/2}}\sum_{\substack{\alpha\equiv\beta(\gamma)\\ N(\alpha)\le N(\gamma)x^{1/2}}}\tau(\alpha)+O(x^{7/8+\varepsilon})$$

Indeed, we have

$$N(\alpha-\beta) = |\alpha-\beta|^2 \ge \left||\alpha|^2-|\beta|^2\right| = N(\alpha)-N(\beta) \quad \text{for } N(\alpha)\ge N(\beta),$$

and

$$N(\alpha-\beta)\le|\alpha|^2+|\beta|^2=N(\alpha)+N(\beta).$$

Therefore we carry an error in the asymptotic formula $\ll N(\beta)x^{1/2}\ll x^{7/8}$ if we replace the condition $N(\alpha-\beta)\le x$ on the condition $N(\alpha)\le x$ (we take into account that $N(\beta)\le x^{3/8}$).

Now, by lemma 3 we obtain

$$\sum_{N(\alpha)\le x^{1/2}}\sum_{\substack{\alpha\equiv\beta(\gamma)\\ N(\alpha)\le x}}=\sum_{\delta|\beta}\sum_{\substack{N(\gamma)\le x^{1/2}N(\delta)^{-1}\\ (\gamma,\beta/\delta)=1}}\sum_{\substack{\alpha\equiv\beta(\mathrm{mod}\ \gamma\delta)\\ N(\alpha)\le x}}\tau(\alpha)=$$

$$=\sum_{\delta|\beta}\sum_{\substack{N(\gamma)\le\frac{x^{1/2}}{N(\delta)}\\ (\gamma,\alpha_0/\delta)=1}}\left\{\frac{\pi^2x}{N^2(\gamma\delta)}N(\delta)\varphi(\gamma)\tau(\delta)\left(\log\frac{x}{N(\delta)}-1\right)+\right.$$

$$\left.+\frac{2\pi^2x}{N(\gamma)}\sum_{t|\delta}\left(E+\frac{L'(1,\chi_4)}{L(1,\chi_4)}+\sum_{p|\gamma\delta/t}\log\frac{N(p)}{N(p)-1}\right)\prod_{p|\gamma\delta/t}(1-N(p)^{-1}\right\}+$$

$$+O\left(\sum_{\delta|\beta}\sum_{N(\gamma)\le\frac{x^{1/2}}{N(\delta)}}x^{1/2}N(\gamma\delta)^{-1/4}\right).$$

Using the equality

$$\varphi(\alpha)=N(\alpha)\prod_{p|\alpha}(1-N(p)^{-1})=N(\alpha)\sum_{\delta|\alpha}\frac{\mu(\delta)}{N(\delta)}$$

we infer

$$\sum_{\substack{N(\alpha)\le x\\ (\alpha,\beta)=1}}\frac{\varphi(\delta)}{N(\delta)}=\sum_{\substack{N(\alpha)\le x\\ (\alpha,\beta)=1}}\sum_{\delta|\alpha}\frac{\mu(\delta)}{N(\delta)}=\sum_{\substack{N(\delta)\le x\\ (\alpha\delta,\beta)=1}}\frac{\mu(\delta)}{N(\delta)}\sum_{N(\alpha)\le\frac{x}{N(\delta)}}1=$$

$$=\prod_{p|\beta}(1-N(p)^{-1})\left(\pi x\sum_{\substack{N(\alpha)\le x\\ (\delta,\beta)=1}}\frac{\mu(\delta)}{N^2(\delta)}+O(x^{1/3})\right)$$

$$=c_0(\beta)x+O(x^{1/3}).\tag{11}$$

where $c_0(\beta)=c_0\frac{\varphi(\beta)}{N(\beta)}\prod_{p|\beta}(1-N(\rho)^{-2}),\quad c_0=const.$

Therefore

$$\sum_{\substack{N(\gamma)\le\frac{x^{1/2}}{N(\delta)}\\ (\gamma,\beta/\delta)=1}}c_0(\beta/\delta)\left(\log x+1+O\left(x^{1/3}\right)\right).\tag{12}$$

Hence, from (10),(12), we get

$$\sum_{N(\gamma)\le x^{1/2}} \sum_{\substack{\alpha\equiv\beta(\gamma)\\ N(\alpha)\le x}} \tau(\alpha) = x(a_0(\beta)\log^2 x + \\ + a_1(\beta)\log x + a_2(\beta)) + O(x^{7/8+\varepsilon}). \quad (13)$$

Similarly

$$\sum_{N(\gamma)\le x^{1/2}} \sum_{\substack{\alpha\equiv\beta(\gamma)\\ N(\alpha)\le N(\gamma)x^{1/2}}} \tau(\alpha) = x(b_1(\beta)\log x + b_2(\beta)) + O(x^{7/8+\varepsilon}). \quad (14)$$

From (9),(13),(14) we obtain the assertion of theorem. □

## References


[1] A.E. Ingham, Some asymptotic formulae in the theory of numbers, 2. J. London Math. Soc. 2(1927), 202-208.

[2] T. Estermann, Über die Darstellungen einer Zahl als Differenz von zwei Producten, J. Reine Angew. Math. 164(1931), 173-182.

[3] D.R.Heath-Brown, The divisor function $d_3(n)$ in arithmetic progression, Acta Arithm. XLVII.1 (1986), 29-56.

[4] J.-M. Deshoillers and H. Ivaniec, An additive divisor problem, Proc. London Math. Soc (2) 26(1982), 1-14.

[5] Y.Motohashi, The binary additive divisor problem, Ann. Sci. École Normale Supériure (4) 27 (1994), 529-572.

[6] P.D. Vatbanets and P.A. Zarzycki, Divisors of the Gaussian integers in an Arithmetic Progression, J. Number Theory (2) 33(1989), 152-269.



Contact information

**O. V. Savasrtu** ul.Dvoryanskaya 2, Dept. of computer algebra and discrete mathematics, Odessa national university, Odessa 65026 Ukraine
*E-Mail:* prik@imem.odessa.ua

**P. D. Varbanets** ul.Solnechnaya 7/9 apt.18, Odessa 65009 Ukraine
*E-Mail:* varb@te.net.ua


Received by the editors: 22.02.2003.

Algebra and Discrete Mathematics RESEARCH ARTICLE
Number 2. **(2003).** pp. 111–124


# Ramseyan variations on symmetric subsequences

**Oleg Verbitsky**

Communicated by V. I. Sushchansky

Abstract. A theorem of Dekking in the combinatorics of words implies that there exists an injective order-preserving transformation $f : \{0, 1, \ldots, n\} \to \{0, 1, \ldots, 2n\}$ with the restriction $f(i + 1) \le f(i) + 2$ such that for every 5-term arithmetic progression $P$ its image $f(P)$ is not an arithmetic progression. In this paper we consider symmetric sets in place of arithmetic progressions and prove lower and upper bounds for the maximum $M = M(n)$ such that every $f$ as above preserves the symmetry of at least one symmetric set $S \subseteq \{0, 1, \ldots, n\}$ with $|S| \ge M$.

## 1. Introduction

Let $[n] = \{0, 1, 2, \ldots, n\}$, with 0 included for our convenience. We consider injective order-preserving transformations $f : [n] \to [2n]$ with restriction $f(i + 1) - f(i) \le 2$ for all $i < n$. We wonder to which extent such transformations can violate the regular structure of $[n]$. Namely, suppose that $\mathcal{P}$ is a regularity property of a set of integers, say, one of being an arithmetic progression. We then wish to know the maximum $M = M(n)$ such that, for every $f$ as above, at least one set $S \subseteq [n]$ with $|S| \ge M$ has property $\mathcal{P}$ and its image $f(S)$ still has the same property.

In the case of arithmetic progressions, it is easy to observe an equivalent reformulation of the question. Let $V = \{v_0, \ldots, v_n\}$ be a sequence of points in the grid $\mathbf{Z}^2$ with each difference $v_{i+1} - v_i$ being either $a = (1, 1)$ or $b = (1, 2)$. Now the question is what is the maximum $M$ such that every $V$ contains an $M$-term arithmetic progression of vectors. To see the equivalence of the two problems, it suffices to view a set $V$ as the

*An extended version of this paper has been published in* INTEGERS: Electronic Journal of Combinatorial Number Theory.

graph of a map $f$. Clearly, $f$ preserves an arithmetic progression $S \subseteq [n]$ iff $\{ (x, f(x)) : x \in S\}$ is an arithmetic progression in $V$. Notice that the specification of differences $a$ and $b$ is actually irrelevant — those could be any other pair of non-collinear vectors as well, say, $a = (1, 0)$ and $b = (0, 1)$.

As the choice of the initial point $v_0$ does not affect anything, a set $V$ is characterized by the sequence of differences $v_1 - v_0, \ldots, v_n - v_{n-1}$, which can be regarded as a word $w(V)$ of length $n$ over alphabet $\{a, b\}$. In this way we arrive at yet another reformulation of the problem under consideration. We call an arbitrary sequence of variables a *pattern*. An *abelian occurrence* of a pattern in a word is a subword obtainable from the pattern by substituting nonempty words in place of variables so that words replacing the same variable may differ only in order of letters (see Section 2 for more details). It is not hard to observe a one-to-one correspondence between $(m + 1)$-term arithmetic progressions in $V$ and abelian occurrences of the pattern $x^m$ in $w(V)$. Thus, the value of $M(n)$ is the maximum number $M$ such that every word of length $n$ over the binary alphabet has an abelian occurrence of $x^{M-1}$.

Dekking [6] constructs an infinite word in the binary alphabet without abelian occurrences of $x^4$. It immediately follows [13, theorem 6.13] that $M(n) \leq 4$, i.e. 5-term arithmetic progressions can all be destroyed by some transformation $f$.

This motivates an extension of property $\mathcal{P}$. A set $S \subseteq \mathbf{Z}^k$ such that $S = g - S$ for a lattice point $g \in \mathbf{Z}^k$ is called *symmetric* (with respect to the center at rational point $\frac{1}{2}g$). From now on the property $\mathcal{P}$ extended to being symmetric will be our main concern. Given $V \subseteq \mathbf{Z}^k$, let $MS(V)$ denote the maximum cardinality of a symmetric subset of $V$.

A pattern is *symmetric* if it reads the same backward as forward, like $xyx$. With notation introduced above, we again have a one-to-one correspondence between sets $S \subseteq [n]$ whose symmetry is preserved by $f$, symmetric subsets of the graph $V$ of $f$, and abelian occurrences of symmetric patterns in the word $w(V)$. Correspondingly, we have the following equivalences whose proof is given in more detail in Section 2.

**Lemma 1.1.** *The statements below are equivalent.*

1. $M(n) = \min_{f:[n]\to[2n]} \max_{S\subseteq[n]} \{ |S| : \textit{both } S \textit{ and } f(S) \textit{ are symmetric}\}$, *where the minimum is taken over all injective $f$ with*

$$1 \leq f(i + 1) - f(i) \leq 2 \tag{1}$$

*for $i < n$.*

*2.* $M(n)$ *is the minimum of* $MS(V)$ *over all subsets* $V = \{v_0, v_1, \ldots, v_n\}$ *of* $\mathbf{Z}^2$ *with each* $v_{i+1} - v_i$ *equal to either* $a$ *or* $b$*, where* $a$ *and* $b$ *are arbitrarily fixed non-collinear vectors.*

*3.* $M(n)$ *is the maximum* $M$ *such that every word of length* $n$ *over the binary alphabet has abelian occurrence of a symmetric pattern of length at least* $M - 1$*.*

In contrast to the case of arithmetic progressions, $M(n)$ now grows with $n$, that is, no $f$ is able to destroy symmetric subsets so well as arithmetic progressions. To show this, consider an infinite sequence of symmetric patterns

$$\begin{aligned} P_1 &= x, \\ P_2 &= xyx, \\ P_3 &= xyxzxyx, \\ P_4 &= xyxzxyxuxyxzxyx, \\ &\vdots \end{aligned} \tag{2}$$

where $P_{i+1}$ is the result of inserting a new variable between two copies of $P_i$. In combinatorics of words, members of this sequence are called *sesquipowers* or *Zimin's patterns.* Coudrain and Schützenberger [5] proved that each $P_i$ must occur in all long enough words over a finite alphabet. Here we mean literal rather than abelian occurrence, i.e. the same variable is substituted everywhere by the same word. The unavoidability of sesquipowers immediately implies that $M(n)$ goes to the infinity with $n$ increasing. However, this argument gives a very small lower bound for $M(n)$, actually, a kind of the inverse tower function (see Lemma 2.3).

In Section 3 we prove a better lower bound $M(n) = \Omega(\ln n)$ based on estimation of how long symmetric pattern is represented by an abelian occurrence in every binary word of length $n$. Similarly to the $O$-notation, we write $\Omega(h(n))$ to refer to a function of $n$ that everywhere exceeds $c \cdot h(n)$ for a positive constant $c$.

In Section 4 we prove upper bound $M(n) = O(\sqrt{n})$. As the main technical tool we use $B_2$-sequences introduced by Sidon and investigated by many authors (see [13, section 4.1] for survey and references). A set $X$ of integers is called a $B_2$*-sequence* if for any integer $g$ the equation $x + y = g$ has at most one solution in $X$ with $x \leq y$. In other words, a $B_2$-sequence $X$ is a highly asymmetric set characterized by $MS(X) \leq 2$. There are several constructions of dense $B_2$-sequences in $[n]$. We employ a fairly simple and explicit construction of [12], making use of an additional uniformity property of it.

**Related work.** The van der Waerden theorem can be restated so that every infinite subsequence $v_0, v_1, \ldots$ of $\mathbf{N}$ with $v_{i+1} - v_i = O(1)$ contains arbitrarily long arithmetic progressions (see [4]). As Dekking's result shows, a similar statement in $\mathbf{Z}^2$ is false. However, Ramsey and Gerver [15] prove that every infinite sequence $v_0, v_1, \ldots$ in $\mathbf{Z}^2$ with bounded distances $||v_{i+1} - v_i||$ between any two successive points contains arbitrarily large subsets of collinear points. Pomerance [14] shows this holds true even under the weaker assumption that

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} ||v_{i+1} - v_i|| < \infty. \tag{3}$$

These results can be viewed as two-dimensional analogs of the van der Waerden theorem and its density version of Szemerédi, with collinear subsets instead of arithmetic progressions. In this respect our result on behavior of $M(n)$, in view of item 2 of Lemma 1.1, can serve as yet another two-dimensional analog of van der Waerden's theorem, with arithmetic progressions replaced by symmetric subsets. The multi-dimensional analog of Szemerédi's theorem is also true as shown by Banakh [3], who observed that condition (3) guarantees the existence of arbitrarily long symmetric subsequences in an infinite sequence $v_0, v_1, \ldots$ of points in $\mathbf{Z}^k$, $k \geq 1$. It should be noted that in the case of $k = 2$ the latter result strengthes the claim that $M(n) \to \infty$ but provides no satisfactory lower bound for $M(n)$.

Banakh and Protasov [2] prove that the minimal number of colors required for coloring the $n$-dimensional integer grid $\mathbf{Z}^n$ avoiding infinite symmetric monochromatic subsets is $n + 1$. Unavoidable symmetries in words are investigated by Fouché [10].

## 2. Preliminaries

In this section we prove Lemma 1.1 and then show that $M(n) \to \infty$ as $n \to \infty$. Recall that throughout the paper $MS(V)$ denotes the cardinality of the largest symmetric subset of $V$.

The proof of the equivalence of statements 1 and 2 of Lemma 1.1 in the case that

$$a = (1, 1), \quad b = (1, 2) \tag{4}$$

follows arguments outlined in the introduction for arithmetic progressions. With a function $f$ we associate its graph $V = \{v_0, \ldots, v_n\}$, where $v_i = (i, f(i))$. The bounds (1) imply that $v_{i+1} - v_i \in \{a, b\}$. Vice versa, any set $V = \{v_0, \ldots, v_n\}$ in $\mathbf{Z}^2$ with the latter condition can be viewed as the graph of a function $f$ of the prescribed kind. A set $S \subseteq [n]$ and its image $f(S)$ are both symmetric iff $S' = \{\,(i, f(i)) : i \in S\}$ is a symmetric subset of $V$. This completes the proof in the case (4).

The case of arbitrary non-collinear $a$ and $b$ reduces to the case (4). Really, consider two sets $V = \{v_0, \ldots, v_n\}$ and $V' = \{v'_0, \ldots, v'_n\}$ in $\mathbf{Z}^2$ with all $v_{i+1} - v_i \in \{(1,1), (1,2)\}$ and $v'_{i+1} - v'_i \in \{a, b\}$, where $a$ and $b$ are non-collinear. Let $\phi$ be the affine transformation of $\mathbf{Z}^2$ into itself that takes $v_0$ to $v'_0$, $(1,1)$ to $a$, and $(1,2)$ to $b$. Then $\phi$ establishes a one-to-one correspondence between $V$ and $V'$ that matches symmetric subsets in $V$ and symmetric subsets in $V'$. It follows that $MS(V) = MS(V')$, thereby proving the equivalence of statements 1 and 2.

Before proving the equivalence of statements 2 and 3, let us recall the relevant notions of the formal language theory. A *pattern* is a word over the alphabet of variables $\{x_1, x_2, \ldots\}$. Pattern $x_{i_1} x_{i_2} \ldots x_{i_l}$ is *symmetric* if $i_j = i_{l+1-j}$ for all $j \le l$. Let $A = \{a_1, \ldots, a_m\}$ be a finite alphabet. The number of occurrences of letter $a_i$ in a word $w$ is denoted by $|w|_{a_i}$. A *commutative index* of $w$ over $A$ is the tuple $\langle |w|_{a_1}, \ldots, |w|_{a_m} \rangle$. A subword $u$ of a word $w$ is an *occurrence* of a pattern $P = x_{i_1} \ldots x_{i_l}$ if $u$ can be obtained from $P$ by substituting nonempty words in place of each variable, where the same variable is everywhere replaced with the same word. If the same variable may be replaced by (possibly distinct) words with the same commutative index, $u$ is called an *abelian occurrence* of $P$.

*Example.* In word $a_1 a_1 a_2 a_1 a_2 a_1 a_3$, subwords $a_1 a_1$, $a_1 a_2 a_1 a_2$, and $a_2 a_1 a_2 a_1$ are occurrences of pattern $x_1 x_1$. In addition, $a_1 a_1 a_2 a_1 a_2 a_1$ is an abelian occurrence of the same pattern.

Given a sequence of vectors $V = \{v_0, v_1, \ldots, v_n\}$ in $\mathbf{Z}^k$ with all $v_i - v_{i-1}$ in a finite set $A \subset \mathbf{Z}^k$, we associate with $V$ the sequence $w(V)$ of differences $v_1 - v_0, v_2 - v_1, \ldots, v_k - v_{k-1}$ which will be viewed as a word of length $n$ over alphabet $A$.

**Lemma 2.1.**

1. *If $w(V)$ has an abelian occurrence of a symmetric pattern of length $l$, then $MS(V) \ge l + 1$.*

2. *Conversely, suppose that $A$ is a linearly independent set of vectors. Then $w(V)$ has an abelian occurrence of a symmetric pattern of length at least $MS(V) - 1$.*

*Proof.* 1. Recall that word $w(V)$ is a sequence of vectors $v_1 - v_0, \ldots, v_n - v_{n-1}$. Given a subword $u = v_{i+1} - v_i \ldots v_j - v_{j-1}$, $i < j$, we call $v_i$ the initial point and $v_j$ the terminal point of $u$. Let $u = u_1 \ldots u_l$ be an abelian occurrence of a symmetric pattern $P$ of length $l$, where $u_s$ is substituted in place of $s$-th variable of $P$. Let $v_{i_{s-1}}$ and $v_{i_s}$ be the initial and terminal points of $u_s$. Then the set $\{v_{i_0}, \dots, v_{i_l}\}$ is symmetric. This can be shown by easy induction. Really, assume that $v_{i_1}$ and $v_{i_{l-1}}$ are symmetric with respect to the center $\frac{1}{2}g$, that is, $v_{i_1} + v_{i_{l-1}} = g$. As $u_1$ and $u_l$ differ only in order of their letters, we have $v_{i_1} - v_{i_0} = v_{i_l} - v_{i_{l-1}}$. Consequently, $v_{i_0}$ and $v_{i_l}$ are symmetric with respect to $\frac{1}{2}g$ too.

2. Let $l = MS(V)$ and $v_{i_0}, \dots, v_{i_l}$ be a symmetric subsequence of $V$. Denote a subword of $w(V)$ whose initial and terminal points are $v_{i_{s-1}}$ and $v_{i_s}$ by $u_s$. Then $u = u_i \dots u_l$ is an abelian occurrence of a symmetric pattern of length $l$. It suffices to show that commutative indices of words $u_s$ and $u_{l+1-s}$ are the same. Those are uniquely determined by expansions of vectors $v_{i_s} - v_{i_{s-1}}$ and $v_{i_{l+1-s}} - v_{i_{l-s}}$ in basis $A$. It remains to notice that the last two vectors are equal by symmetricalness of $\{v_{i_0}, \dots, v_{i_l}\}$. □

The equivalence of statements 2 and 3 of Lemma 1.1 now follows directly from Lemma 2.1. The proof of Lemma 1.1 is complete.

**Proposition 2.2.** *$M(n) \to \infty$ as $n \to \infty$.*

At this point we prefer the statement 3 of Lemma 1.1.

Let $L^{non-abel}(n)$ be the maximal $l$ such that every word of length $n$ over the binary alphabet has an occurrence of a symmetric pattern of length at least $l$. As $M(n) > L^{non-abel}(n)$, it suffices to show that $L^{non-abel}(n) \to \infty$ for $n \to \infty$. The latter follows from a result of Coudrain and Schützenberger [5] which we state below in a form convenient for our purposes.

**Lemma 2.3 ([5]).** *If $L^{non-abel}(n) \geq l$, then $L^{non-abel}((n+1)(2^n+1)) \geq 2l+1$.*

*Proof.* Assume that every binary word of length $n$ has occurrence of a symmetric pattern $P$ of length $l$. Any binary word of length $(n+1)(2^n+1)$ contains two identical subwords of length $n$ separated by a nonempty word. Thus, there is an occurrence of the symmetric pattern $PxP$, where $x$ is a new variable absent in $P$. □

Notice that the above argument ensures that each pattern $P_i$ of the sequence (2) occurs in any long enough binary word.

## 3. Lower bound

The proof of Proposition 2.2 based on Lemma 2.3 gives us an extremely small lower bound for $M(n)$ that is even smaller than the inverse tower function. In this section we improve it to $M(n) \geq 2\ln n - O(1)$. We first prove an auxiliary fact. Notice that whenever below we refer to the number of subwords of a word, we distinguish all occurrences of a subword, that is, a subword is counted each time it occurs in the word.

**Lemma 3.1.** *Given a word $w$, let $\nu(w)$ denote the number of pairs $\{u_1, u_2\}$, where $u_1$ and $u_2$ are disjoint subwords of $w$ with the same commutative index. Let $N(n)$ be the minimum of $\nu(w)$ over all binary words $w$ of length $n$. Then*

$$N(n) \geq (\ln n - O(1))n^2/4.$$

*Proof.* Consider a binary word $w$ of length $n$ and estimate the value $\nu(w)$ from below. Expand $\nu(w)$ to the sum $\sum_t \nu_t(w)$, where the $t$-th term counts pairs of subwords with length $t$. Let $\sigma_t(i)$ denote the number of subwords of $w$ with length $t$ and commutative index $\langle i, t-i\rangle$. As the total number of subwords of length $t$ is equal to $n+1-t$, notice the equality $\sigma_t(0) + \sigma_t(1) + \ldots + \sigma_t(t) = n+1-t$. As a subword of length $t$ can overlap with at most $2t-1$ subwords of the same length, we have

$$\nu_t(w) \geq \frac{1}{2}\sum_{i=0}^{t} \sigma_t(i)(\sigma_t(i) - (2t-1)).$$

Taking into account that

$$\sum_{i=0}^{t} \sigma_t(i)^2 \geq (t+1)\left(\frac{\sum_{i=0}^{t}\sigma_t(i)}{t+1}\right)^2,$$

we conclude that

$$\begin{aligned} \nu_t(w) \quad &\geq \quad \frac{1}{2}\left((t+1)\left(\frac{n+1-t}{t+1}\right)^2 - (2t-1)(n+1-t)\right) \\ &= \quad \frac{(n+2)^2}{2(t+1)} - (t+\frac{1}{2})(n+1) + t^2 - \frac{1}{2}. \end{aligned}$$

Let us sum these inequalities over $t$ from 1 to $s$, dropping the last term $t^2 - \frac{1}{2}$ in the right hand side (anyway it would give us no gain). Summing the first term in the right hand side, we take into account that $\sum_{t=1}^{s} 1/t - \ln s$ approaches Euler's constant as $s$ increases. Therefore,

$$\sum_{t=1}^{s} \nu_t(w) \geq \frac{1}{2}(\ln s - O(1))(n+2)^2 - \frac{s(s+2)}{2}(n+1).$$

Setting $s = \lceil \sqrt{n}\,\rceil$, we obtain the proclaimed bound for $\nu(w)$ and hence for $N(n)$. ∎

**Theorem 3.2.** $M(n) \geq 2\ln n - O(1)$.

*Proof.* We adhere to the statement 2 of Lemma 1.1.

Let $V = \{v_0, v_1, \ldots, v_n\}$ be a set of points in $\mathbf{Z}^2$ with $v_{i+1} - v_i \in \{a, b\}$. Denote $G = \left\{ \frac{1}{2}(v_i + v_j) : 0 \leq i \leq j \leq n \right\}$, the set of all potential centers of symmetry. Let $m_g$ denote the "multiplicity" of an element $g$ in $G$, that is, the number of pairs $(i, j)$ such that $g = \frac{1}{2}(v_i + v_j)$ and $i \leq j$. Clearly,

$$\sum_{g \in G} m_g = (n+1)(n+2)/2.$$

Furthermore, let $N$ denote the total number of quadruples

$$(v_l, v_i, v_j, v_k) \text{ with } l < i \leq j < k \text{ and } v_i - v_l = v_k - v_j. \tag{5}$$

Clearly,

$$N \leq \sum_{g \in G} \binom{m_g}{2}$$

(actually, the linear independence of $a$ and $b$ implies the equality here). It follows that

$$N < \frac{1}{2} \sum_{g \in G} m_g^2 \leq \frac{1}{2} (\max_{g \in G} m_g) \sum_{g \in G} m_g = \frac{1}{4} n^2 (1 + O(\frac{1}{n})) \max_{g \in G} m_g. \tag{6}$$

Recall that with the set $V$ we associate a word $w(V)$ over alphabet $\{a, b\}$. It is easy to observe a one-to-one correspondence between quadruples (5) in $V$ and pairs of disjoint subwords $u_1$ and $u_2$ with the same commutative index in $w(V)$. By Lemma 3.1 we have

$$N \geq (\ln n - O(1)) n^2 / 4.$$

Together with (6), this gives

$$\max_{g \in G} m_g \geq \ln n - O(1).$$

It remains to observe that, for every center $g \in G$, the set $V$ contains a subset that is symmetric with respect to $g$ and has at least $2m_g - 1$ elements. □

## 4. Upper bound

In this section we prove an upper bound for $M(n)$.

**Theorem 4.1.** $M(n) \leq (7 + o(1))\sqrt{n}$.

We use a two-dimensional geometric interpretation of $M(n)$ given by statement 2 of Lemma 1.1. We will construct a set $V = \{v_0, v_1, \ldots, v_n\}$ of points in $\mathbf{Z}^2$ such that each difference $v_{i+1} - v_i$ is either $(1,0)$ or $(0,1)$ and $MS(V) \le (7 + o(1))\sqrt{n}$.

Our construction will be completely determined by two sets of integers $X = \{x_1, \ldots, x_q\}$ and $Y = \{y_1, \ldots, y_q\}$ listed in the ascending order. Given $X$ and $Y$, consider a sequence of points in $\mathbf{Z}^2$

$$(x_1, y_1),\ (x_2, y_1),\ (x_2, y_2),\ (x_3, y_2),\ (x_3, y_3),\ \ldots,\ (x_q, y_q) \tag{7}$$

We define $V$ by $V = V_1 \cup V_2$, where

$$V_1 = \bigcup_{i=1}^{q} \{\, (x_i, y) : y_{i-1} < y \le y_i \} \text{ and } V_2 = \bigcup_{i=1}^{q-1} \{\, (x, y_i) : x_i < x \le x_{i+1} \}$$

(we set $y_0 = y_1 - 1$ for convenience). Thus, (7) are "corner" points of $V$, at which difference $v_{i+1} - v_i$ changes its value from $(1,0)$ to $(0,1)$ or vice versa. Clearly, $V$ consists of $x_q + y_q + 1 - x_1 - y_1$ points.

Given a set $Z = \{z_1, \ldots, z_q\}$ of integers listed in the ascending order, define $D(Z) = \max_{1 \le i < q}(z_{i+1} - z_i)$.

**Lemma 4.2.** *Suppose that $V$ has been constructed based on $q$-element sets $X$ and $Y$ as described above. Then*

$$MS(V) < MS(X)D(Y) + MS(Y)D(X) + 2q. \tag{8}$$

*Proof.* Let $S$ be the maximum subset of $V$ symmetric with respect to center $\frac{1}{2}g$, i.e. $S = V \cap (g - V)$. Clearly,

$$S = (V_1 \cap g - V_1) \cup (V_2 \cap g - V_2) \cup (V_1 \cap g - V_2) \cup (V_2 \cap g - V_1).$$

Let us estimate the cardinality of each member of the union.

$V_1 \cap g - V_1$ is a symmetric subset of $V_1$. As the projection of $V_1 \cap g - V_1$ onto the first coordinate is symmetric too, the cardinality of this projection does not exceed $MS(X)$. As any cut of $V_1$ by vertical line (i.e. along the second coordinate) containes at most $D(Y)$ points, we have $|V_1 \cap g - V_1| \le MS(X)D(Y)$. Similarly, $|V_2 \cap g - V_2| \le MS(Y)D(X)$.

Observe now that all points of $V_1$ differ in the second coordinate and have only $q$ values for the first coordinate, while all points of $V_2$ differ in the first coordinate and have only $q$ values for the second coordinate. As a consequence, both $V_1 \cap g - V_2$ and $V_2 \cap g - V_1$ have less than $q$ points. The bound (8) follows. □

We now need to choose $X$ and $Y$ so as to make the right hand side of (8) as small as possible. The idea is to use a $B_2$-sequence $X = Y$, which gives us the best possible $MS(X) = MS(Y) = 2$. It easily follows from [8] that $D(X) \geq q(1 - o(1))$ for any $B_2$-sequence $X = \{x_1, \ldots, x_q\}$. We use a construction of [12] that provides us with $D(X) \leq (3 + o(1))q$.

**Lemma 4.3 (Krückeberg [12]).** *For any prime $q$ there is a sequence of integers $X = \{x_1, \ldots, x_q\}$ with $MS(X) = 2$ and $D(X) < 3q$. Moreover, $x_1 = 0$ and $x_q = 2q^2 - 2q - 1$.*

We include the proof of this lemma given in [12], because it contains a simple explicit construction of the needed $B_2$-sequences, thereby making our construction of $V$ explicit too.

*Proof.* Set $x_{i+1} = 2qi - (i^2 \bmod q)$ for $0 \leq i < q$, where expression $i^2$ mod $q$ stands for the least non-negative residue of $i^2$ modulo $q$. Obviously, $q < x_{i+1} - x_i < 3q$. To show that $X$ is a $B_2$-sequence, assume that $x_i + x_j = x_{i'} + x_{j'}$, $i \leq j$, $i' \leq j'$. It is easy to derive from this that

$$\begin{cases} i + j = i' + j' & (\bmod\, q) \\ i^2 + j^2 = (i')^2 + (j')^2 & (\bmod\, q) \end{cases}$$

Since in the field $\mathbf{F}_q$ a system of kind

$$\begin{cases} i + j = a \\ i^2 + j^2 = b \end{cases}$$

can have only a unique solution $i$, $j$ with $i \leq j$, we conclude that $i = i'$ and $j = j'$. □

Let us summarize our construction of the set $V = \{v_0, v_1, \ldots, v_n\}$. Let $q$ be the prime next to $(\sqrt{n+3} + 1)/2$ and $X$ be the $B_2$-set given by Lemma 4.3. Applying the construction described in the beginning of the section with $Y = X$, we obtain a set $V' = \{v_0, v_1, \ldots, v_n, \ldots\}$ of $4q^2 - 4q - 1 \geq n + 1$ points in $\mathbf{Z}^2$. Leaving aside some last elements of $V'$, we get the set $V$. By Lemma 4.2, $MS(V) \leq MS(V') < 14q$. Since the prime next to $m$ does not exceed $m + O(m^\alpha)$, where $0 < \alpha < 1$ [11], we have $MS(V) \leq (7 + o(1))\sqrt{n}$. The proof of Theorem 4.1 is complete.

**Remark 4.4.** The choice of Krückeberg's $B_2$-sequence is essentially best possible, because the right hand side of (8) cannot be smaller than $\sqrt{2n}$, whatever sets $X$ and $Y$ are. Let us prove this fact. First, observe relation

$$MS(X) \geq q/D(X). \tag{9}$$

for a set of integers $X = \{x_1, \dots, x_q\}$. This is a consequence of inclusion $X \subseteq \bigcup_{g=0}^{D(X)-1}(g + x_1 + x_q - X)$ which implies $|X \cap (g - X)| \geq q/D(X)$ for some $g$. By (9)

$$MS(X)D(Y) + MS(Y)D(X) \geq 2(MS(X)D(Y)MS(Y)D(X))^{1/2} \geq 2q$$

and therefore the right hand side of (8) is at least $4q$.

Further, observe that $MS(V) > \max\{D(X), D(Y)\}$. Using this, we have $n = |V| - 1 \leq q(D(X) + D(Y)) < 2q\,MS(V)$. Therefore, the right hand side of (8) exceeds $2n/MS(V)$. It remains to notice that one of the values $MS(V)$ and $2n/MS(V)$ is at least $\sqrt{2n}$.

**Remark 4.5.** Consider a random set $\mathbf{V} = \{v_0, v_1, \dots, v_n\}$ in $\mathbf{Z}^2$ with $v_{i+1} - v_i \in \{a, b\}$ for non-collinear $a$ and $b$. We mean that the underlying word $w(\mathbf{V})$ is uniformly distributed on $\{a, b\}^n$. The mean value of $MS(\mathbf{V})$ could serve as an upper bound for $M(n)$. Unfortunately, this probabilistic argument cannot give anything better that the constructive bound of Theorem 4.1 by the following reason.

Just for simplicity assume that $n = 2m$ is even. Let $\mathbf{s}$ denote the cardinality of the maximum subset of $\mathbf{V}$ symmetric with respect to the center at the medium point $v_m$. Consider now two independent sequences $\xi_1, \dots, \xi_m$ and $\zeta_1, \dots, \zeta_m$ of unbiased Bernoulli trials, that is, all $\xi_i$ and $\zeta_j$ are mutually independent random variables that take on equiprobable values 0 and 1. Denote the number of $k$ such that $\sum_{i=1}^k \xi_i = \sum_{i=1}^k \zeta_i$ by $\mathbf{t}$. In coding $a = 0$ and $b = 1$, it becomes clear that $\mathbf{s} = 2\mathbf{t} + 1$. Estimate the expectation of $\mathbf{t}$ from below.

Let $p_k = \mathbf{P}\left[\sum_{i=1}^k \xi_i = \sum_{i=1}^k \zeta_i\right]$. By linearity of mathematical expectation, $\mathbf{E}[\mathbf{t}] = \sum_{k=1}^m p_k$. Using Chernoff's bound, we have

$$p_k = \sum_{l=0}^{k} \mathbf{P}\left[\sum_{i=1}^{k} \xi_i = l\right]^2 > \sum_{k/2-\sqrt{k} \leq l \leq k/2+\sqrt{k}} \mathbf{P}\left[\sum_{i=1}^{k} \xi_i = l\right]^2 \geq$$

$$(2\sqrt{k} - 1)\left(\frac{\mathbf{P}\left[k/2 - \sqrt{k} \leq \sum_{i=1}^k \xi_i \leq k/2 + \sqrt{k}\right]}{2\sqrt{k} + 1}\right)^2 \geq$$

$$\geq \frac{(1 - 2\exp(-2))^2}{2\sqrt{k} + 7}.$$

Therefore, $\mathbf{E}[\mathbf{t}] = \Omega\left(\sum_{k=1}^m 1/\sqrt{k}\right) = \Omega(\sqrt{m})$. As $\mathbf{E}[\mathbf{s}] = 2\mathbf{E}[\mathbf{t}] + 1$, we conclude that the mean value of $MS(\mathbf{V})$ is $\Omega(\sqrt{n})$.

In conclusion we discuss one more aspect of the upper bound proven in this section. Given $n$, we have constructed a set $\{v_0, v_1, \dots, v_n\}$ with

$$MS(\{v_0, v_1, \dots, v_n\}) = O(\sqrt{n}). \tag{10}$$

**Question 4.6.** Is it possible to construct an infinite set $\{v_0, v_1, v_2 \dots\}$ such that (10) is true for all $n$?

We could achieve this goal with the same construction, if we had an infinite $B_2$-sequence $X = \{x_1, x_2, \dots\}$ with $D(\{x_1, \dots, x_q\}) = O(q)$ for all $q$. However, the latter condition implies $|X \cap [m]| = \Omega(\sqrt{m})$ for all $m$, whereas no $B_2$-sequence satisfies this condition by a result of Erdős. Erdős proves that there is a constant $c$ such that for any infinite $B_2$-sequence $X$ the inequality $|X \cap [m]| \le c\sqrt{m/\ln m}$ is true for infinitely many $m$ (see [9]). The best known construction of [1] gives $|X \cap [m]| = \Omega\left((m \ln m)^{1/3}\right)$. Nevertheless, we are able at least to approach (10) with an infinite $V$.

**Proposition 4.7.** *There is an infinite sequence $V = \{v_0, v_1, v_2, \dots\}$ with each difference $v_{i+1} - v_i$ either $(1,0)$ or $(0,1)$ and such that*

$$MS(\{v_0, v_1, \dots, v_n\}) = n^{1/2+O(1/\ln\ln n)} \tag{11}$$

*for all $n$.*

*Proof.* We apply the straightforward infinite version of the construction described in the beginning of this section with $X = Y = \{1, 4, 9, 16, \dots\}$, the set of integer squares. By Lemma 4.2, for any integer $q$ and $n = 2q^2 - 2$

$$MS(\{v_0, v_1, \dots, v_n\}) < 2MS(\{1, 4, \dots, q^2\})D(\{1, 4, \dots, q^2\}) + 2q.$$

We obviously have $D(\{1, 4, \dots, q^2\}) = 2q - 1$ and, by Lemma 4.8 below,

$$MS(\{1, 4, \dots, q^2\}) = q^{O(1/\ln\ln q)}.$$

This proves (11) for all $n = 2q^2 - 2$. Equality (11) is true for any other $n$ as well, because the next to $n$ number of kind $2q^2 - 2$ does not exceed $n + \sqrt{(n+2)/2} + 1 = n(1 + o(1))$. □

The following lemma in other terms estimates the number of representations of an integer as a sum of two squares. Though this estimate easily follows from the well-known number-theoretic facts, we give a proof for the sake of completeness.

**Lemma 4.8.** $MS(\{1, 4, \dots, q^2\}) = q^{O(1/\ln\ln q)}$.

*Proof.* It is easy to see that the maximum subset of $\{1, 4, \ldots, q^2\}$ symmetric with respect to $\frac{1}{2}g$ has as many elements as the number of solutions of equation $z_1 + z_2 = g$ in $\{1, 4, \ldots, q^2\}$. The Jacobi theorem (see e.g. [7, theorem 65]) says that if $g = 2^k m$ with odd $m$, then the total number of integer solutions of the equation $x^2 + y^2 = g$ is equal to $4E$, where $E$ is the excess of the number of divisors $t \equiv 1 \pmod 4$ of $m$ over the number of divisors $t \equiv 3 \pmod 4$ of $m$. We use the bound $E \le d(m)$, where $d(m)$ denotes the total number of positive divisors of $m$. It is known [16] that $d(m) = m^{O(1/\ln\ln m)}$. As $m \le g$ and it makes sense to consider only $g < 2q^2$, we have $d(m) = q^{O(1/\ln\ln q)}$. Summarizing, we obtain $MS(\{1, 4, \ldots, q^2\}) \le 4E \le 4d(m) = q^{O(1/\ln\ln q)}$. □

## Acknowledgments

I am grateful to Taras Banakh, whose proof of Proposition 2.2 in geometric terms inspired this work, and to Oleg Pikhurko, whose observation improved the original proof of Theorem 3.2 and led to a better lower bound for $M(n)$. I thank an anonymous referee of the electronic journal "INTEGERS" for several corrections.

## References

[1] M. Ajtai, J. Komlós and E. Szemerédi. A dense infinite Sidon sequence. *European J. Combin.*, 2:1–11, 1981.

[2] T. O. Banakh and I. V. Protasov. Asymmetric partitions of Abelian groups (in Russian). *Mat. Zametki*, 66(1):10–19, 1999.

[3] T. Banakh, I. Kmit, and O. Verbitsky. On asymmetric colorings of integer grids. *Ars Combinatoria*, 62:257–271, 2002.

[4] T. C. Brown. Variations on van der Waerden's and Ramsey's theorems. *Am. Math. Mon.*, 82:993–995, 1975.

[5] M. Coudrain and M. P. Schützenberger. Une condition de finitude des monoides finiment engendres. *C. R. Acad. Sci., Paris, Ser. A*, 262:1149–1151, 1966.

[6] F. M. Dekking. Strongly non-repetitive sequences and progression-free sets. *J. Combin. Theory*, 27:181–185, 1979.

[7] L. E. Dickson. *Introduction to the theory of numbers.* Dover Publ., New York, 1957.

[8] P. Erdős and P. Turan. On a problem of Sidon in additive number theory, and on some related problems. *J. London Math. Soc.*, 16:212–215, 1941.

[9] P. Erdős, message to A. Stöhr, published in: A. Stöhr, Gelöste und ungelöste Fragen über Basen der natürlichen Zahlenreihe. II. *J. Reine Angew. Math.*, 194:132–133, 1955.

[10] W. L. Fouché. Unavoidable regularities and factor permutations of words. *Proc. Royal Soc. Edinb.*, 125A(3):519–524, 1995.

[11] G. Hoheisel. Primzahlprobleme in der Analysis. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, 573:580–588, 1930.

[12] F. Krückeberg. $B_2$-Folgen und verwandte Zahlenfolgen. *J. Reine Angew. Math.*, 206:53–60, 1961.

[13] C. Pomerance and A. Sárközy. Combinatorial number theory. In *Handbook of Combinatorics*, chapter 20, pages 967–1018. Elsevier Publ., 1995.

[14] C. Pomerance. Collinear subsets of lattice point sequences — an analog of Szemerédi's theorem. *J. Combin. Theory A*, 28:140–149, 1980.

[15] L. T. Ramsey and J. L. Gerver. On certain sequences of lattice points. *Pacific J. Math*, 83:357–363, 1979.

[16] S. Wigert. Sur l'odre de grandeur du nombre des diviseurs d'un entier. *Arkiv för Matematik, Astronomi och Fisik*, 3(18):1–9, 1906–1907.

## Contact information

**O. Verbitsky**
Department of Algebra
Faculty of Mechanics & Mathematics
Kyiv National University
Volodymyrska 60
01033 Kyiv, Ukraine
*E-Mail:* oleg@ov.litech.net

Received by the editors: 13.12.2002.