UNIVERSAL ALGEBRA

Jaroslav Ježek

First edition, April 2008

Contents

PREFACE

This is a short text on universal algebra. It is a composition of my various notes that were collected with long breaks for many years, even decades; recently I put it all together to make it more coherent. Some parts were written and offered to my students during the past years.

The aim was to explain basics of universal algebra that can be useful for a starting advanced student, intending possibly to work in this area and having some background in mathematics and in algebra in particular. I will be concise. Many proofs could be considered as just hints for proving the results. The text could be easily doubled in size. Almost no mention of the history of universal algebra will be given; suffice it to say that foundations were laid by G. Birkhoff in the 1930's and 1940's. There will be not many remarks about motivation or connections to other topics. We start with two chapters collecting some useful knowledge of two different subjects – set theory and the theory of categories, just that knowledge that is useful for universal algebra. Also, the chapter on model theory is intended only as a server for our purposes.

The bibliography at the end of the book is not very extensive. I included only what I considered to be necessary and closely related to the material selected for exposition. Many results will be included without paying credit to their authors.

Selection of the topics was not given only by my esteem of their importance. The selection reflects also availability provided by my previous notes, and personal interest. Some most important modern topics will not be included, or will be just touched. This text contains no original results.

I do not keep to the usual convention of denoting algebras by boldface characters and then their underlying sets by the corresponding non-bold variants. Instead, I use boldface characters for constants (of any kind) and italics, as well as greek characters (both upper- and lower-case), for variables (running over objects of any kind). I do not reserve groups of characters for sorts of variables.

The rare cases when it is really necessary to distinguish between an algebra and its underlying set, can be treated by adding a few more words.

Other texts can be also recommended for alternative or further reading: Burris and Sankappanavar [81]; McKenzie, McNulty and Taylor [87]; Hobby and Mckenzie [88]; Freese and McKenzie [87]; Grätzer [79]; García and Taylor [84]; Gorbunov [99]. Some material in the present book has been also drawn from the first four of these books.

I would be grateful for comments of any kind, and in particular for pointing out errors or inconsistencies. I could use them for improvements that would be included in a possible second edition which may also contain some extensions. Please contact me at jezek@karlin.mff.cuni.cz.

My thanks are to Ralph McKenzie, Miklos Maróti and Petar Marković for many friendly discussions that were also of great value when I was working on this text.

CHAPTER 1

SET THEORY

The whole of mathematics is based on set theory. Because intuitive set theory can easily lead to the well-known paradoxes (the set A of all the sets that are not their own elements can satisfy neither $A \in A$ nor $A \notin A$), it is reasonable to work in a theory with a carefully selected system of axioms. Two such axiom systems, essentially equivalent, are the most common: the Gödel-Bernays and the Zermelo-Fraenkel systems. For universal algebra the first is the more convenient. The purpose of this chapter is to present foundations of set theory based on the Gödel-Bernays system of axioms. The books Gödel [40], Cohen [66] and Vopěnka, Hájek [72] can be recommended for further reading.

1. Formulas of set theory

Certain strings of symbols will be called formulas. Symbols that may occur in the strings are the following:

(1) Variables: both lower- and upper-case italic letters or letters of the Greek alphabet, possibly indexed by numerals (there should be no restriction on the number of variables)
(2) One unary connective: $\neg$
(3) Four binary connectives: $\& \vee \rightarrow \leftrightarrow$
(3) Two quantifiers: $\forall$ $\exists$
(4) Parentheses: ( )
(5) Equality symbol: $=$
(6) Membership symbol: $\in$

A *formula* is a string that can be obtained by several applications of the following rules. For every formula we also specify which variables are called *free* and which variables are called *bound* in it.

(1) For any two (not necessarily distinct) variables x and y, the strings x=y and x∈y are formulas; both x and y are free, no other variable is free, and no variable is bound in these formulas.
(2) If f is a formula then ¬(f) is a formula; a variable is free (or bound) in ¬(f) if and only if it is free (or bound, respectively) in f.
(3) If f and g are two formulas and if no variable is either simultaneously free in f and bound in g or simultaneously bound in f and free in g, then the four strings (f)&(g) and (f)∨(g) and (f)→(g) and (f)↔(g) are formulas; a variable is free (or bound) in the resulting formula if

and only if it is free (or bound, respectively) in at least one of the the formulas f and g.

(4) If f is a formula and if x is a variable that is not bound in f, then both $(\forall x)(f)$ and $(\exists x)(f)$ are formulas; the variable x is bound in the resulting formula; a variable other than x is free (or bound) in the resulting formula if and only if it is free (or bound, respectively) in f.

Observe that no variable is both free and bound in any formula. A variable occurs in a formula if and only if it is either free or bound in it. By a *sentence* we mean a formula without free variables.

Certain formulas are called *logical axioms.* If f, g and h are three formulas, then the following are logical axioms provided that they are formulas (some parentheses are omitted):

(1) $f\rightarrow(g\rightarrow f)$
(2) $(f\rightarrow(g\rightarrow h))\rightarrow((f\rightarrow g)\rightarrow(f\rightarrow h))$
(3) $((\neg f)\rightarrow(\neg g))\rightarrow(g\rightarrow f)$
(4) $(f\leftrightarrow g)\rightarrow(f\rightarrow g)$
(5) $(f\leftrightarrow g)\rightarrow(g\rightarrow f)$
(6) $(f\rightarrow g)\rightarrow((g\rightarrow f)\rightarrow(f\leftrightarrow g))$
(7) $(f\vee g)\leftrightarrow((\neg f)\rightarrow g)$
(8) $(f\&g)\leftrightarrow\neg((\neg f)\vee(\neg g))$
(9) $((\forall x)f)\rightarrow g$ where x and y are two variables not bound in f and g is obtained from f by replacing all the occurrences of x with y
(10) $((\forall x)(f\rightarrow g))\rightarrow(f\rightarrow(\forall x)g)$ where x is a variable not occurring in f
(11) $((\exists x)f)\leftrightarrow\neg((\forall x)\neg f)$ where x is a variable not bound in f
(12) $x=x$ where x is a variable
(13) $x=y\rightarrow y=x$ where x and y are two variables
(14) $(x=y\&y=z)\rightarrow x=z$ where x, y and z are three variables
(15) $(x=y\&z=u)\rightarrow(x\in z\leftrightarrow y\in u)$ where x, y, z and u are four variables

By a *theory* we mean a (finite) collection of formulas of the language; these formulas are called axioms of that theory.

By a *proof* in a given theory T we mean a finite sequence of formulas such that each member of the sequence is either a logical axiom or an axiom of T or can be obtained from one or two earlier members of the sequence by one of the following two rules:

(1) obtain g from f and $f\rightarrow g$;
(2) obtain f from $(\forall x)f$.

By a proof of a given formula in a given theory T we mean a proof in T which has the given formula as its last member. A formula is said to be provable in T if there exists a proof of the formula in T.

A theory S is said to be an extension of a theory T if every axiom of T is an axiom of S. A theory S is said to be stronger than a theory T if every axiom of T is provable in S. As it is easy to see, it follows that each formula provable in T is also provable in S. Two theories are said to be equivalent if each is stronger than the other one. Clearly, every theory is equivalent to a theory with the same number of axioms, all the axioms of which are sentences.

It is easy to see that for a sentence f and any formula g, the formula f$\rightarrow$g is provable in a theory T if and only if g is provable in the theory obtained from T by adding f as a new axiom.

A theory T is said to be inconsistent if there is a formula f such that f&¬f is provable in T. Clearly, in an inconsistent theory every formula is provable.

We will work in one particular theory, the set theory, and do mathematics informally. It is useful to have at mind, however, that our theorems should be expressible as sentences and that it should be possible to translate their informal proofs to obtain proofs in the above given rigorous sense. Definitions are to be understood as abbreviations for particular formulas. In order to introduce set theory, we need to start with a weaker theory.

2. Theory of classes

Any object under our investigation is a *class*. Thus to say, for an example, that there exists a class with some property is the same as to say that there exists an X with the property. By a *set* we mean a class a such that there exists an X with $a \in X$. If $a \in X$ then we say that a is an element of X. Thus to be a set is the same as to be an element of something. By a *proper class* we mean a class that is not a set.

Theory of classes has the following nine axioms:

(C1) If A and B are two classes such that $a \in A \leftrightarrow a \in B$ for all sets a, then $A = B$.

(C2) For any sets a and b there exists a set c such that for any set x, $x \in c$ if and only if either $x = a$ or $x = b$.

(C3) There exists a class A such that $a \in A$ for any set a.

Before we continue with the list of the axioms, we need to introduce two definitions. The set c, the existence of which is postulated in (C2) is, according to (C1), uniquely determined by a and b. It will be denoted by $\{a, b\}$; put $\{a\} = \{a, a\}$. For any sets a and b put $\langle a, b\rangle = \{\{a\}, \{a, b\}\}$. The set $\langle a, b\rangle$ is called the *ordered pair* (or just pair) of a and b. Put $\langle a\rangle = a$; for three sets a, b, c put $\langle a, b, c\rangle = \langle a, \langle b, c\rangle\rangle$; and similarly for four sets, etc.

(C4) For any class A there exists a class B such that for any set a, $a \in B$ if and only there exist two sets x, y with $a = \langle x, y\rangle$, $x \in y$ and $a \in A$.

(C5) For any two classes A and B there exists a class C such that for any set a, $a \in C$ if and only if $a \in A$ and $a \notin B$.

(C6) For any class A there exists a class B such that for any set a, $a \in B$ if and only if there exists an x with $\langle x, a\rangle \in A$.

(C7) For any two classes A and B there exists a class C such that for any set a, $a \in C$ if and only if there exist two sets x, y with $a = \langle x, y\rangle$, $y \in B$ and $a \in A$.

(C8) For any class A there exists a class B such that for any set a, $a \in B$ if and only if there exist two sets x, y with $a = \langle x, y\rangle$ and $\langle y, x\rangle \in A$.

(C9) For any class A there exists a class B such that for any set a, $a \in B$ if and only if there exist three sets x, y, z with $a = \langle x, \langle y, z\rangle\rangle$ and $\langle y, \langle z, x\rangle\rangle \in A$.

2.1. Theorem. *Let a, b, c, d be four sets. Then $\langle a, b\rangle = \langle c, d\rangle$ if and only if $a = c$ and $b = d$.*

Proof. This is easy to see. □

For two classes A and B we write $A \subseteq B$ (or also $B \supseteq A$) if for every set x, $x \in A$ implies $x \in B$. We say that A is a *subclass* of B, or that A is a *subset* of B in case that A is a set.

For two classes A and B we write $A \subset B$ (or also $B \supset A$) if $A \subseteq B$ and $A \neq B$. We say that A is a proper subclass of B.

It follows from (C3) and (C1) that there exists precisely one class such that every set is its element. We use the constant $\mathbf{V}$ to denote this class; the class $\mathbf{V}$ is called the universal class.

A class A is said to be a *relation* if its every element is an ordered pair.

It follows from (C4) that there exists precisely one relation A such that for any sets x and y, $\langle x, y\rangle \in A$ if and only if $x \in y$. This relation A will be denoted by $\mathbf{E}$.

For any two classes A and B, the uniquely determined class C from (C5) will be denoted by $A \setminus B$. It contains precisely the elements of A that do not belong to B. It will be called the *difference* of A and B.

For any two classes A and B put $A \cap B = A \setminus (A \setminus B)$. This class is called the *intersection* of A and B. It contains precisely the elements belonging to both A and B. We define $A \cap B \cap C = (A \cap B) \cap C$, etc.

For any two classes A and B put $A \cup B = \mathbf{V} \setminus ((\mathbf{V} \setminus A) \cap (\mathbf{V} \setminus B))$. This class is called the *union* of A and B. It contains precisely the elements belonging to at least one of the two classes, either A or B. We define $A \cup B \cup C = (A \cup B) \cup C$, etc.

Put $0 = \mathbf{V} \setminus \mathbf{V}$. This class is called the *empty* class. A class A is called nonempty if $A \neq 0$. Two classes A, B are said to be *disjoint* if $A \cap B = 0$.

For any class A, the uniquely determined class B from (C6) will be denoted by $\mathbf{Dom}(A)$. It will be called the *domain* of A. For a set a, we have $a \in \mathbf{Dom}(A)$ if and only if $\langle x, a\rangle \in A$ for at least one x.

For any two classes A and B, the uniquely determined class C from (C7) will be denoted by $A \restriction B$. This class is a relation and contains precisely the ordered pairs $\langle x, y\rangle \in A$ with $y \in B$. It will be called the *restriction* of A to B.

For any class A, the uniquely determined class B from (C8) will be denoted by $\mathbf{Inv}(A)$ and called the *inverse* of A. This relation contains precisely the ordered pairs $\langle x, y\rangle$ with $\langle y, x\rangle \in A$.

For any class A, the uniquely determined class B from (C9) will be denoted by $\mathbf{Inv}_3(A)$.

For any class A, the class $\mathbf{Dom}(\mathbf{Inv}(A))$ is called the *range* of A.

For any two classes A and B put $A \times B = (\mathbf{V} \restriction B) \cap \mathbf{Inv}(\mathbf{V} \restriction A)$. This relation will be called the *direct product* of A and B. It contains precisely the ordered pairs $\langle a, b \rangle$ such that $a \in A$ and $b \in B$. Put $A \times B \times C = A \times (B \times C)$, etc. Put $A^1 = A$, $A^2 = A \times A$, $A^3 = A \times A \times A$, etc.

We have $A \restriction B = A \cap (\mathbf{V} \times B)$.

A class A is said to be a *function* if it is a relation and for any three sets x, y, z, $\langle y, x \rangle \in A$ and $\langle z, x \rangle \in A$ imply $y = z$. If A is a function then instead of $\langle y, x \rangle \in A$ we write $A\text{'}x = y$ (or $A(x) = y$ if there is no confusion).

For two classes A and X, the range of $A \restriction X$ is denoted by $A\text{''}X$; if there is no confucion, if it is clear that we do not mean $A\text{'}X$, we also write $A(X)$ for $A\text{''}X$.

Let f be a formula. A sequence $\mathrm{x_1}$ through $\mathrm{x_n}$ of distinct variables is said to be free for f if none of these variables is bound in f. (The sequence is allowed to be empty, and need not consist of variables occurring in f.) We say that f is *CT-admissible* (or class theory admissible) with respect to such a free sequence of variables if the following is provable in the theory of classes: there exists a class A such that for any set a, $a \in A$ if and only if there exist $\mathrm{x_1}$ through $\mathrm{x_n}$ such that $a = \langle \mathrm{x_1}, \dots, \mathrm{x_n} \rangle$ and f is satisfied. A formula is said to be CT-admissible if it is CT-admissible with respect to any sequence of variables free for f.

Let X and Y be two different variables. We are going to show that the formula X $\in$ Y is CT-admissible. Let $\mathrm{x_1}$ through $\mathrm{x_n}$ be a sequence of variables free for this formula.

Consider first the case when neither X nor Y is among $\mathrm{x_1}$ through $\mathrm{x_n}$. We can take $A = 0$ if X $\in$ Y is satisfied, while $A = \mathbf{V}^{\mathrm{n}}$ in the opposite case.

Next consider the case when X is $\mathrm{x_i}$ for some i, while Y is not among $\mathrm{x_1}$ through $\mathrm{x_n}$. We can take $A = A_1 \times \dots \times A_{\mathrm{n}}$, where all the factors are $\mathbf{V}$, except the i-th factor, which is Y.

Next consider the case when Y is $\mathrm{x_i}$ for some i, while X is not among $\mathrm{x_1}$ through $\mathrm{x_n}$. If X is not a set, we can take $A = 0$. Otherwise, we can take $A = A_1 \times \dots \times A_{\mathrm{n}}$, where all the factors are $\mathbf{V}$, except the i-th factor, which is the range of $\mathbf{Inv}(\mathbf{E}) \restriction \{\mathrm{X}\}$.

It remains to consider the case when X is $\mathrm{x_i}$ and Y is $\mathrm{x_j}$. Consider first the subcase when i is less than j. If j is n, let $X = \mathbf{E}$; otherwise, let $X = \mathbf{Inv}_3(\mathbf{Inv}_3(\mathbf{V}^{\mathrm{n\text{-}j}} \times \mathbf{E}))$. We have $\langle \mathrm{x_i}, \mathrm{x_j}, \dots, \mathrm{x_n} \rangle \in X$ if and only if X $\in$ Y. If j is i+1, let $Y = X$; otherwise, let $Y = \mathbf{Inv}_3(\mathbf{V} \times \mathbf{Inv}(X))$. We have $\langle \mathrm{x_i}, \mathrm{x_{j\text{-}1}}, \mathrm{x_j}, \dots, \mathrm{x_n} \rangle \in Y$ if and only if X $\in$ Y. Repeating this process, we can find a class Z such that $\langle \mathrm{x_i}, \dots, \mathrm{x_n} \rangle \in Z$ if and only if X $\in$ Y. We can take $A = \mathbf{V} \times \dots \times \mathbf{V} \times Z$ where the number of factors is i. Now consider the subcase when j is less than i. We can proceed in the same way as above, taking $\mathbf{Inv}(\mathbf{E})$ instead of $\mathbf{E}$ and switching i and j.

We are going to show that if f is a CT-admissible formula, then $\neg$f is also CT-admissible. If a sequence $\mathrm{x_1}$ through $\mathrm{x_n}$ of variables is free for $\neg$f, then it is also free for f. The resulting class A for f can be replaced with $\mathbf{V}^{\mathrm{n}} \setminus A$ to obtain the resulting class for $\neg$f.

We are going to show that if f&g is a formula and f and g are both CT-admissible, then f&g is also CT-admissible. If a sequence x_1 through x_{n} is free for f&g, then it is free for both f and g. Let A be the resulting class for f and B be the resulting class for g. Then $A \cap B$ is the resulting class for f&g.

We are going to show that if f is a CT-admissible formula and X is a variable not bound in f, then $(\exists \mathrm{X})((\mathrm{X} \in \mathbf{V})\&\mathrm{f})$ is also CT-admissible. Let x_1 through x_{n} be a sequence of variables free for this formula. Then X followed by x_1 through x_{n} is a sequence of variables free for f. Let A be the resulting class for f. Then $\mathbf{Dom}(A)$ is the resulting class for $(\exists \mathrm{X})((\mathrm{X} \in \mathbf{V})\&\mathrm{f})$. (It would be more appropriate to write $(\exists \mathrm{Y})\mathrm{X} \in \mathrm{Y}$ instead of $\mathrm{X} \in \mathbf{V}$; we would have to say that Y is a variable different from X and not occurring in f.

With respect to an arbitrary extension of the theory of classes, certain formulas of that theory will be called *set-restricted formulas.* If X and Y are two variables, then $\mathrm{X} \in \mathrm{Y}$ is a set-restricted formula. If f is a set-restricted formula, then $\neg$f is a set-restricted formula. If f and g are two set-restricted formulas, then f&g, f$\vee$g, f$\rightarrow$g and f$\leftrightarrow$g are set-restricted formulas, under the assumption that they are formulas. If f is a set-restricted formula and X is a variable not bound in f, then $(\exists \mathrm{X})((\mathrm{X} \in \mathbf{V})\&\mathrm{f})$ and $(\forall \mathrm{X})((\mathrm{X} \in \mathbf{V}) \rightarrow \mathrm{f})$ are set-restricted formulas. Finally, if f$\leftrightarrow$g is provable in the theory under consideration and f is a set-restricted formula, then g is also a set-restricted formula.

It follows that every set-restricted formula is CT-admissible. Let f be a set-restricted formula and let x_1 through x_{n} be a sequence of variables free for f. Then we define $\{\langle \mathrm{x}_1, \dots, \mathrm{x}_{\mathrm{n}} \rangle : \mathrm{f}\}$ to be the uniquely determined class, resulting in the way described above. The following are examples.

For any class A put $\bigcup(A) = \{x : (\exists y)(x \in y \& y \in A\}$. This class is called the *union* of A. It is also sometimes denoted by $\bigcup A$.

For any class A put $\bigcap(A) = \{x : (\forall y)(y \in A \rightarrow x \in y)\}$. This class is called the *intersection* of A. It is also sometimes denoted by $\bigcap A$.

For any class A put $\mathbf{P}(A) = \{x : x \subseteq A\}$. This class is called the *power-class* of A.

3. Set theory

Set theory is the theory of classes extended by the following axioms:

(S1) For any set a there exists a set b such that whenever $x \in y \in a$ then $x \in b$.
(S2) For any set a there exists a set b such that for all sets x, whenever $x \subseteq a$ then $x \in b$.
(S3) For any set a and any function A, the class $A``a$ is a set.
(S4) Every nonempty class A contains an element a such that $A \cap a = 0$.
(S5) There exists a nonempty set a with $(\forall x)(x \in a \rightarrow (\exists y)(y \in a \& x \subset y))$.

(S6) There exists a function A with domain $\mathbf{V}$ such that $A\text{'}x \in x$ for any nonempty set x.

The axiom (S4) will be referred to as the axiom of regularity. The axiom (S5) will be referred to as the axiom of infinity. The axiom (S6) will be referred to as the axiom of choice.

3.1. Theorem. *Every subclass of a set is a set.*

Proof. Let A be a subclass of a set a. Put $B = \{\langle x, y\rangle : x = y \& x \in A\}$, so that B is a function and $A = B\text{''}a$. By (S3), A is a set. □

3.2. Theorem. *For every set a, the classes $\bigcup a$ and $\mathbf{P}(a)$ are sets.*

Proof. It follows from (S1), (S2) and 3.1. □

3.3. Theorem. *If a and b are two sets, then $a \cup b$ is a set.*

Proof. We have $a \cup b = \bigcup\{a, b\}$; this class is a set by (C2) and (S1). □

For three sets a, b, c put $\{a, b, c\} = \{a, b\} \cup \{c\}$. Similarly, for four sets a, b, c, d put $\{a, b, c, d\} = \{a, b, c\} \cup \{d\}$; etc.

3.4. Theorem. *If a and b are two sets, then $a \times b$ is a set.*

Proof. We have $a \times b \subseteq \mathbf{P}(\mathbf{P}(a \cup b))$. So, we can use 3.1, 3.2 and 3.3. □

3.5. Theorem. *If a is a set, then the domain of a, the range of a, the inverse of a and $\mathbf{Inv}_3(a)$ are sets.*

Proof. It is easy. □

3.6. Theorem. *Let A be a function. Then A is a set if and only if the domain of A is a set.*

Proof. We have $A \subseteq (A\text{''}\mathbf{Dom}(A)) \times \mathbf{Dom}(A)$. □

3.7. Theorem. *The empty class 0 is a set.*

Proof. By (S5), there exists a set a. We have $0 \subseteq a$. □

3.8. Theorem. *There is no set a with $a \in a$.*

Proof. Suppose $a \in a$. By (S4) applied to $\{a\}$, we have $\{a\} \cap a = 0$, a contradiction. □

Similarly, using (S4) it is possible to prove for any positive n the following: There do not exist sets $a_1, \dots, a_n$ with $a_1 \in a_2, \dots, a_{n-1} \in a_n, a_n \in a_1$.

3.9. Theorem. *The class $\mathbf{V}$ is a proper class.*

Proof. Suppose that $\mathbf{V}$ is a set. Then $\mathbf{V} \in \mathbf{V}$, a contradiction with 3.8. □

4. Relations and functions

By a *mapping* of a class A into a class B we mean a function F such that $\mathbf{Dom}(F) = A$ and the range of F is a subclass of B. If, moreover, the range of B equals B, we say that F is a mapping of A onto B.

A function F is said to be *injective* if $\mathbf{Inv}(F)$ is also a function. By a *bijection* of A onto B we mean an injective mapping of A onto B. By a permutation of A we mean a bijection of A onto itself.

If F is a function and $a \in \mathbf{Dom}(A)$, then the element $F\text{'}a$ of the range of F is called the *value* of F at a.

By a *family* we mean a mapping, the domain of which is a set. By a family of elements of B we mean a mapping of a set into the class B.

Let C be a family with domain I. The set of all mappings f with domain I such that $f\text{'}i \in C\text{'}i$ for all $i \in I$ is called the *direct product* of C and is denoted by ΠC. Clearly, if $C'i = 0$ for at least one $i \in I$, then $\Pi C = 0$; the direct product of the empty family is the one-element set $\{0\}$. It follows from the axiom of choice that if the sets $C\text{'}i$ are all nonempty then ΠC is nonempty.

For two classes A and B put

$$A \circ B = \{\langle a, b\rangle : (\exists c)(\langle a, c\rangle \in A \ \&\ \langle c, b\rangle \in B)\}.$$

This class is called the *composition* of A and B. Observe that $(A \circ B) \circ C = A \circ (B \circ C)$ for any classes A, B, C. We define $A \circ B \circ C = (A \circ B) \circ C$, $A \circ B \circ C \circ D = (A \circ B \circ C) \circ D$, etc.

Clearly, if F is a mapping of A into B and G is a mapping of B into C, then $G \circ F$ is a mapping of A into C, and for any $a \in A$ we have $(G \circ F)\text{'}a = G\text{'}(F\text{'}a)$.

For any class A put $\mathbf{id}_A = \{\langle a, a\rangle : a \in A\}$. This class is called the *identity* on A. Clearly, $\mathbf{id}_A$ is a bijection of A onto A.

For a set a and a class B we denote by B^a the class of all mappings of a into B. If B is a set then B^a is also a set (since it is a subclass of $\mathbf{P}(B \times a)$).

By a relation on a class A we mean a subclass of $A \times A$.

A relation R is said to be *reflexive* on a class A if $\langle a, a\rangle \in R$ for all $a \in A$.

A relation R is said to be *symmetric* if $\langle a, b\rangle \in R$ implies $\langle b, a\rangle \in R$.

A relation R is said to be *transitive* if $\langle a, b\rangle \in R$ and $\langle b, c\rangle \in R$ imply $\langle a, c\rangle \in R$.

A relation R is said to be *irreflexive* if $\langle a, a\rangle \notin R$ for all sets a.

A relation R is said to be *antisymmetric* if $\langle a, b\rangle \in R$ and $\langle b, a\rangle \in R$ imply $a = b$.

By an *equivalence* on a class A we mean a relation that is reflexive on A, symmetric and transitive. Clearly, for any class A, both $\mathbf{id}_A$ and A^2 are equivalences on A. For any equivalence R on A we have $\mathbf{id}_A \subseteq R \subseteq A^2$. If R is an equivalence on A and $a \in A$, then the class $R\text{''}\{a\} = \{x : \langle x, a\rangle \in R\}$ is called the *block* of a with respect to R, or the R-block of a.

If R is an equivalence on a set A, then the set $\{R\text{''}\{a\} : a \in A\}$ (the set of all R-blocks) will be denoted by A/R and called the *factor* of A through R; the mapping $\{\langle R\text{''}\{a\}, a\rangle : a \in A\}$ of A onto A/R is called the *canonical* *projection* of A onto A/R. Often (although there is some inconsistency in the notation) we also write a/R instead of $R``\{a\}$.

By a *partition* of a set A we mean a set P of nonempty subsets of A such that $\bigcup P = A$ and $X \cap Y = 0$ for any $X, Y \in P$ with $X \neq Y$.

4.1. Theorem. *For any given set A, the mapping assigning to any equivalence R on A the factor A/R, is a bijection of the set of all equivalences on A onto the set of all partitions of A. If P is a partition of A, then the corresponding equivalence R on A is defined by $\langle a, b\rangle \in R$ if and only if there is an $X \in P$ with $a, b \in X$.*

Proof. It is easy. □

For any function F we define

$$\mathbf{ker}(F) = \{\langle x, y\rangle : x \in \mathbf{Dom}(F) \ \&\ y \in \mathbf{Dom}(F) \ \&\ F(x) = F(y)\}.$$

This relation is called the *kernel* of F. Clearly, $\mathbf{ker}(F)$ is an equivalence on $\mathbf{Dom}(F)$.

4.2. Theorem. *Let F be a mapping of a set A into a set B. Then there exists precisely one mapping G of $A/\mathbf{ker}(F)$ into B such that $G \circ H = F$, where H is the canonical projection of A onto $A/\mathbf{ker}(F)$. This mapping G is injective; if F is a mapping of A onto B, then G is a bijection of $A/\mathbf{ker}(F)$ onto B.*

Proof. It is easy. □

4.3. Theorem. *Let F be a mapping of a class A into a class B. Then F is a bijection of A onto B if and only if there exists a mapping G of B into A such that $G \circ F = \mathbf{id}_A$ and $F \circ G = \mathbf{id}_B$.*

Proof. It is easy. □

Let R be a relation on a class A and S be a relation on a class B. By a *homomorphism* of A into B with respect to R, S we mean a mapping F of A into B such that $\langle x, y\rangle \in R$ implies $\langle F`x, F`y\rangle \in S$. By an *isomorphism* of A onto B with respect to R, S we mean a bijection F of A onto B such that F is a homomorphism of A into B with respect to R, S and $\mathbf{Inv}(F)$ is a homomorphism of B into A with respect to S, R. We say that A is *isomorphic* to B with respect to R, S if there exists an isomorphism of A onto B with respect to R, S.

By an *ordering* on a class A we mean a relation R on A which is reflexive on A, transitive and antisymmetric. Observe that the class A is uniquely determined by R: it is both the domain and the range of R. By an ordering we mean a class which is an ordering on its domain.

Let R be an ordering on A and let $B \subseteq A$. By a *minimal element* of B with respect to R we mean any element $a \in B$ such that $x = a$ for any $x \in B$ with $\langle x, a\rangle \in R$. By a *maximal element* of B with respect to R we mean any element $a \in B$ such that $x = a$ for any $x \in B$ with $\langle a, x\rangle \in R$. By the *least element* of

B with respect to R we mean any element $a \in B$ such that $\langle a, x\rangle \in R$ for all $x \in B$. By the *greatest element* of B with respect to R we mean any element $a \in B$ such that $\langle x, a\rangle \in R$ for all $x \in B$. Clearly, every subclass of A has at most one least and at most one greatest element with respect to R.

Let R be an ordering on A and let $B \subseteq A$. An element $a \in A$ is said to be a lower (or upper, respectively) bound of B with respect to R if $\langle a, x\rangle \in R$ (or $\langle x, a\rangle \in R$, respectively) for all $x \in B$. An element $a \in A$ is said to be the *infimum* of B with respect to R if it is the greatest element of the class of the lower bounds of B with respect to R. An element $a \in A$ is said to be the *supremum* of B with respect to R if it is the least element of the class of the upper bounds of B with respect to R. Clearly, every subclass of A has at most one infimum and at most one supremum with respect to R.

An ordering R is said to be *linear* if for any $a, b \in \mathbf{Dom}(R)$, either $\langle a, b\rangle \in R$ or $\langle b, a\rangle \in R$. A linear ordering R is said to be a *well ordering* if for any nonempty subclass B of the domain of R, there exists a minimal element of B with respect to R.

Let R be an ordering on A and let $a \in A$. The class $\{x \in A : \langle x, a\rangle \in R\}$ is called the *section* of a with respect to R (this can be a proper class).

5. Ordinal numbers

For a class A, we denote by $\in_A$ the relation $\mathbf{E} \cap (A \times A)$ and by $\in_A^=$ the relation $\mathbf{id}_A \cup \in_A$ on A.

By an *ordinal number* (or just an ordinal) we mean any set a such that $i \in j \in a$ implies $i \in a$ and the relation $\in_a^=$ is a well ordering. (The first condition can be also stated as follows: every element of a is a subset of a.) The class of ordinal numbers will be denoted by $\mathbf{On}$.

5.1. Lemma. *Let a be an ordinal number and s be a subset of a such that $i \in j \in s$ implies $i \in s$. Then either $s \in a$ or $s = a$.*

Proof. Let $s \neq a$, so that $a \smallsetminus s$ is nonempty and there exists an element $b \in a \smallsetminus s$ with $b \in c$ for any element $c \in a \smallsetminus s$ different from b. It is easy to see that $s \subseteq b$ and $b \subseteq s$, so that $s = b$ and $s \in a$. □

5.2. Theorem. *Every element of an ordinal number is an ordinal number.*

Proof. Let a be an ordinal and let $b \in a$. If $i \in j \in b$, then $i \in j \in a$, so that $i \in a$. Consequently, either $i \in b$ or $i = b$ or $b \in i$. The last two cases are impossible according to the remark following 3.8. Consequently, $i \in j \in b$ implies $i \in b$. The second condition is also satisfied, since b is a subset of a. □

Let a and b be two ordinal numbers. We write $a < b$ (or also $b > a$) if $a \in b$. We write $a \leq b$ (or also $b \geq a$) if either $a < b$ or $a = b$. According to 5.1, $a \leq b$ if and only if $a \subseteq b$. We write $a \not\leq b$ (or also $b \not\geq a$) if $a \leq b$ does not hold. For two ordinal numbers a and b, the larger of them is denoted by $\max(a, b)$ and the smaller by $\min(a, b)$.

5.3. Theorem. *The relation $\in_{\mathbf{On}}^{=}$, i.e., the class of the ordered pairs $\langle a, b\rangle$ such that a, b are ordinal numbers with $a \leq b$, is a well ordering on* **On**.

Proof. Clearly, $\in_{\mathbf{On}}^{=}$ is an ordering on **On**. Let a, b be two ordinals such that $a \not\leq b$. Then a is not a subset of b and there exists an element $c \in a \setminus b$ such that every element of c belongs to b. By 5.1 we get either $c \in b$ or $c = b$. But $c \notin b$, and hence $c = b$. It follows that $b < a$, and we have proved that $\in_{\mathbf{On}}^{=}$ is a linear ordering.

It remains to prove that if A is a nonempty subclass of **On**, then there exists a minimal element of A with respect to $\in_{\mathbf{On}}^{=}$. Take $a \in A$ arbitrarily. If a is minimal in A, we are through. If not, then the subset $a \cap A$ of a is nonempty, there exists a minimal element b of $a \cap A$ and it is easy to see that b is a minimal element of A. □

5.4. Theorem. **On** *is a proper class.*

Proof. Suppose **On** is a set. Then it follows from 5.3 that **On** is an ordinal number, so that $\mathbf{On} \in \mathbf{On}$, a contradiction. □

5.5. Theorem. 0 *is an ordinal number. If a is an ordinal number, then $a \cup \{a\}$ is an ordinal number and $a < a \cup \{a\}$. If s is a set of ordinal numbers, then $\bigcup s$ is an ordinal number; if s is nonempty, then $\bigcup s$ is the supremum of s in* **On** *with respect to the well ordering $\in_{\mathbf{On}}^{=}$.*

Proof. It is easy. □

For an ordinal number a, the ordinal number $b = a \cup \{a\}$ is called the *ordinal successor* of a. Clearly, $a < b$ and there is no ordinal c with $a < c < b$. The ordinal successor of 0 is denoted by 1. We have $1 = \{0\}$.

For a nonempty set s of ordinal numbers, the ordinal number $\bigcup s$ is called the supremum of s.

By a *limit ordinal* we mean any ordinal number which is not the ordinal successor of any ordinal number. Thus 0 is a limit ordinal. Every other limit ordinal a is the supremum of the set of the ordinal numbers b with $b < a$ (this set is equal to a).

For a non-limit ordinal number a there exists precisely one ordinal number b such that a is the ordinal successor of b. This b is called the *ordinal predecessor* of a.

In order to prove that all ordinal numbers have a given property, one can proceed in the following way: prove that 0 has the property; and, for any ordinal number a, prove that a has the property under the assumption that all the ordinal numbers less than a have the property. Equivalently, one can proceed in this way: prove that 0 has the property; for any ordinal number a prove that if a has the property, then the ordinal successor of a has the property; and, finally, prove for any limit ordinal number $a \neq 0$ that if all the ordinal numbers less than a have the property, then a has the property. In these cases we say that the proof is done by *transfinite induction.*

5.6. THEOREM. *For every function G with domain $\mathbf{V}$ and every ordinal number a there exists precisely one function f with domain a such that $f\text{'}i = G\text{'}(f \restriction i)$ for all $i \in a$. Also, for every G as above there exists precisely one function F with domain* **On** *such that $F\text{'}i = G\text{'}(F \restriction i)$ for all ordinal numbers i.*

PROOF. Let G be given. Suppose there exists an ordinal number a for which either such a function f does not exist, or there exist two different such functions. Then there exists a minimal ordinal a with this property; denote it by b. Clearly, $b > 0$. If b is the successor of an ordinal c, then take the unique function f corresponding to c; it is easy to see that $f \cup \{\langle G\text{'}f, b\rangle\}$ is the unique function corresponding to b. It remains to consider the case when b is a limit ordinal. Then it is easy to see that the union of the set of the unique functions corresponding to the elements of b is the unique function corresponding to b. The function F is the union of the class of all the functions f obtained in this way. □

We will usually apply this theorem informally. If we say that a function is defined by transfinite induction, we mean that the existence and the unicity of the defined function can be obtained from Theorem 5.6 in the obvious way.

5.7. THEOREM. *There exist limit ordinal numbers other than* 0.

PROOF. Suppose, on the contrary, that every ordinal number other than 0 has an ordinal predecessor. By (S5) there exists a nonempty set S such that for every $x \in S$ there exists a $y \in S$ with $x \subset y$. Let us define by transfinite induction a function F with domain **On** as follows: $F\text{'}0$ is an arbitrary element of S; if a is an ordinal number with ordinal predecessor b and if $F\text{'}b \in S$ is already defined, let $F\text{'}a$ be an arbitrary element of S such that $F\text{'}b \subset F\text{'}a$. Clearly, F is a bijection of the proper class **On** onto a subset of S, a contradiction with (S3). □

The least limit ordinal number different from 0 will be denoted by ω. The elements of ω are called *natural numbers*.

5.8. THEOREM. *Let r be a well ordering on a set s. Then there exists precisely one ordinal number a such that s is isomorphic to a with respect to $r, \in_a^=$.*

PROOF. Take an element e not belonging to s and define a mapping F with domain **On** by transfinite induction in this way: if i is an ordinal number and $F \restriction a$ is already defined, then $F\text{'}i$ is the least element of the subset $\{x \in s : x \neq F\text{'}j \text{ for all } j \in i\}$ of s with respect to r; if, however, this subset is empty, put $F\text{'}i = e$. Denote by a the least ordinal with $F\text{'}a = e$. Clearly, the restriction of F to a is a desired isomorphism.

In order to prove the converse, it is sufficient to show that if a, b are two ordinal numbers such that there exists an isomorphism f of a onto b with respect to $\in_a^=, \in_b^=$, then $a = b$. It is easy to see that $f\text{'}i = i$ for all $i \in a$. From this we get $a = b$. □

Given a set s and a well ordering r on s, the unique ordinal number a such that s is isomorphic to a with respect to $r, \in^{=}_{a}$ is called the ordinal number (or also the *ordinal type*) of s with respect to r. By the ordinal type of a set u of ordinal numbers we mean the ordinal type of u with respect to $\in^{=}_{\mathbf{On}} \cap (u \times u)$.

Let a and b be two ordinal numbers. We denote by $a+b$ the ordinal number of the set $s = (a \times \{0\}) \cup (b \times \{1\})$ with respect to the well ordering r on s, where r is defined as follows: $\langle\langle x, i\rangle, \langle y, j\rangle\rangle \in r$ if and only if either $i < j$ or $i = j$ and $x \leq y$. (It is easy to check that r is a well ordering.)

Let a and b be two ordinal numbers. We denote by $a \cdot b$ the ordinal number of the set $s = a \times b$ with respect to the well ordering r on s, where r is defined as follows: $\langle\langle x, y\rangle, \langle u, v\rangle\rangle \in r$ if and only if either $y < v$ or $y = v$ and $x \leq u$. (It is easy to check that r is a well ordering.)

It is easy to see that for any ordinal number a, $a+1$ is the ordinal successor of a. We define $2 = 1 + 1$.

5.9. Theorem. *For any ordinal numbers a, b, c we have*

(1) $(a + b) + c = a + (b + c)$
(2) $a + 0 = 0 + a = a$
(3) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$
(4) $a \cdot 0 = 0 \cdot a = 0$
(5) $a \cdot 1 = 1 \cdot a = a$
(6) $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$

Proof. It is easy. □

5.10. Example. We do not have $a + b = b + a$ for all ordinal numbers a, b. For example, $1 + \omega = \omega < \omega + 1$.

Similarly, we do not have $a \cdot b = b \cdot a$ for all a, b. For example, $2 \cdot \omega = \omega < \omega \cdot 2$.

Also, we do not have $(a + b) \cdot c = (a \cdot c) + (b \cdot c)$. For example, $(1 + 1) \cdot \omega = \omega < \omega + \omega$.

5.11. Theorem. *Let a, b, c be ordinal numbers such that $a < b$. Then*

(1) $a + c \leq b + c$
(2) $c + a < c + b$
(3) $a \cdot c \leq b \cdot c$
(4) *if c is not a limit ordinal then $a \cdot c < b \cdot c$*
(5) *if $c > 0$ then $c \cdot a < c \cdot b$*

Proof. It is easy. □

5.12. Theorem. *Let a, b, c be ordinal numbers. Then*

(1) *$c + a < c + b$ implies $a < b$*
(2) *$a + c < b + c$ implies $a < b$*
(3) *$c \cdot a < c \cdot b$ implies $a < b$*
(4) *$a \cdot c < b \cdot c$ implies $a < b$*
(5) *if $c > 0$ and $c \cdot a = c \cdot b$ then $a = b$*

Proof. It is easy. □

It is easy to see that for every pair a, b of ordinal numbers such that $a \leq b$ there exists a unique ordinal number c with $a + c = b$. This unique ordinal number c is denoted by $b - a$.

5.13. THEOREM. *Let a, b, c be ordinal numbers. Then*

(1) $(a + b) - a = b$
(2) *if* $a \leq b$ *then* $a + (b - a) = b$
(3) *if* $a \leq b$ *then* $c \cdot (b - a) = (c \cdot a) - (c \cdot b)$

PROOF. It is easy. □

5.14. THEOREM. *For every ordinal number a there exists a unique pair b, n such that b is an ordinal number, n is a natural number and $a = (\omega \cdot b) + n$.*

PROOF. It is easy. □

For a given ordinal number a, we define $\exp(a, b)$ for any ordinal number b by transfinite induction as follows: $\exp(a, 0) = 1$; $\exp(a, b + 1) = \exp(a, b) \cdot a$; if b is a limit ordinal number then $\exp(0, b) = 0$ while $\exp(a, b) = \bigcup\{\exp(a, c) : c \in b\}$ for $a > 0$.

5.15. THEOREM. *Let a, b, c be ordinal numbers. Then*

(1) $\exp(a, b + c) = \exp(a, b) \cdot \exp(a, c)$
(2) $\exp(\exp(a, b), c) = \exp(a, b \cdot c)$

PROOF. It is easy. □

In order to prove that all natural numbers have a given property, one can proceed in this way: prove that 0 has the property; and, for any natural number n, prove that if n has the property then also $n + 1$ has the property. We say in this case that the proof is done by induction (on n). Also, it follows from 5.6 that for every function g with domain ω and every set a there exists a unique function f with domain ω, such that $f\text{'}0 = a$ and $f\text{'}(n+1) = g\text{'}(f\text{'}n)$ for all $n \in \omega$; if we define a function f in this way, we say that it is defined by *induction.*

5.16. THEOREM. *Let a and b be natural numbers. Then*

(1) $a + b = b + a$
(2) $a \cdot b = b \cdot a$

PROOF. One can easily prove by induction that $x+1 = 1+x$ for all natural numbers x. Then one can proceed to prove by induction on b, for any fixed a, that $a + b = b + a$ and $a \cdot b = b \cdot a$. □

5.17. THEOREM. *For every ordinal number a there exists a limit ordinal ordinal number b such that $a < b$. Consequently, the class of limit ordinal numbers is a proper class.*

PROOF. For any ordinal number a, $a + \omega$ is a limit ordinal number and $a < a + \omega$. □

5.18. THEOREM. *Let A be a proper class and R be a well ordering of A such that every section with respect to R is a set. Then there exists a unique isomorphism of* **On** *onto A with respect to $\in_{\mathbf{On}}, R$.*

PROOF. For every ordinal number a define $F\text{'}a \in A$ by transfinite induction as follows: $F\text{'}a$ is the least element x of A with respect to R such that $\langle F\text{'}b, x\rangle \in R$ and $F\text{'}b \neq x$ for all $b \in a$. It is easy to see that F is the only isomorphism. □

The function $\mathbf{W}$ with domain **On** is defined by transfinite induction as follows: $\mathbf{W}\text{'}0 = 0$; if $a = b + 1$, then $\mathbf{W}\text{'}a = \mathbf{P}(\mathbf{W}\text{'}b)$; if a is a nonzero limit ordinal, then $\mathbf{W}\text{'}a = \bigcup(\mathbf{W}\text{''}a)$.

5.19. THEOREM. $\mathbf{V}$ *is the union of the range of* $\mathbf{W}$.

PROOF. Suppose, on the contrary, that there exists a set s not belonging to the union of the range of $\mathbf{W}$. Define a function F with domain **On** by transfinite induction in this way: $F\text{'}0 = s$; if a is an ordinal number with the ordinal predecessor b and if $F\text{'}b$ has been defined and if $F\text{'}b$ is nonempty, let $F\text{'}a$ be an arbitrary element of $F\text{'}b$; in all other cases define $F\text{'}a$ arbitrarily. It is easy to prove that the image of ω under F is a set contradicting (S4). □

We still did not use the axiom of choice anywhere. The following are consequences of the axiom of choice.

5.20. THEOREM. *For every set s there exist an ordinal number a and a bijection of a onto s.*

PROOF. Let A be a function as in (S6). Define a mapping F with domain **On** by transfinite induction as follows: if b is an ordinal and $F\text{'}i$ has been constructed for all $i \in b$, then if the set $t = s \setminus \{F\text{'}i : i \in b\}$ is nonempty, put $F\text{'}b = A\text{'}t$, and in the opposite case put $F\text{'}b = s$. We cannot have $F\text{'}b \in s$ for all ordinal numbers b. So, let a be the minimal ordinal number such that $F\text{'}b \notin s$; then the restriction of F to a is a bijection of a onto s. □

5.21. THEOREM. *Let R be an ordering on a set A satisfying the following condition: whenever B is a subset of A such that $R \cap (B \times B)$ is a well ordering then B has an upper bound in A with respect to R. Then for every element $a \in A$ there exists an element $b \in A$ such that $\langle a, b\rangle \in R$ and b is a maximal element of A with respect to R.*

PROOF. Let us take an element $z \notin A$. Let $a \in A$. By transfinite induction we can define a function F with domain **On** in this way: $F'0 = a$; if i is an ordinal number and there exists an element $x \in A$ such that $\langle F\text{'}j, x\rangle \in R$ and $F\text{'}j \neq x$ for all $j < i$, take one such element x and put $F\text{'}i = x$; in all other cases put $F\text{'}i = z$. (We have used the axiom of choice.) We cannot have $F\text{'}i \in A$ for all ordinal numbers i, since A is a set. Denote by k the least ordinal number such that $F\text{'}k = z$. It follows from the condition that k is not a limit ordinal number. So, $k = m + 1$ for an ordinal number n. Clearly, we can put $b = F\text{'}m$. □

By an *inclusion-chain* we mean any class A such that whenever $x, y \in A$ then either $x \subseteq y$ or $y \subseteq x$.

5.22. Corollary. *Let S be a set such that the union of every nonempty inclusion-chain that is a subset of S belongs to S. Then for every $a \in S$ there exists an element $b \in S$ such that $a \subseteq b$ and b is a maximal element of S (i.e., $b \subseteq c \in S$ implies $c = b$).*

Theorem 5.21 and also its corollary 5.22 will be referred to as Zorn's lemma.

6. Cardinal numbers

Two sets a, b are said to be *equipotent* if there exists a bijection of a onto b. By a *cardinal number* we mean any ordinal number a such that there is no ordinal number $b < a$ equipotent with a.

It follows from 5.20 that for every set a there exists a unique cardinal number equipotent with a. This cardinal number will be denoted by $\mathbf{card}(a)$; it is called the *cardinality* of a. Clearly, two sets a and b are equipotent if and only if $\mathbf{card}(a) = \mathbf{card}(b)$.

6.1. Theorem. *Let a and b be two sets. The following three conditions are equivalent:*

(1) $\mathbf{card}(a) \leq \mathbf{card}(b)$;
(2) *there exists an injective mapping of a into b;*
(3) *either $a = 0$ or there exists a mapping of b onto a.*

Proof. It is easy. □

6.2. Theorem. *Every natural number is a cardinal number. Also, ω is a cardinal number.*

Proof. The first statement can be proved by induction, and it can be proved by induction on a natural number n that n is not equipotent with ω. □

A set is said to be *finite* if its cardinality is a natural number. Clearly, natural numbers are precisely the finite ordinal numbers; also, natural numbers are precisely the finite cardinal numbers.

A set is said to be *infinite* if it is not finite; it is said to be countably infinite if its cardinality is ω. A set is said to be *countable* if it is either finite or countably infinite.

6.3. Theorem. *A set A is infinite if and only if there exists a bijection of A onto its proper subset.*

Proof. It is easy. □

6.4. Theorem. *Let a, b be two disjoint finite sets. Then $\mathbf{card}(a \cup b) = \mathbf{card}(a) + \mathbf{card}(b)$.*

Let a, b be two finite sets. Then $\mathbf{card}(a \times b) = \mathbf{card}(a) \cdot \mathbf{card}(b)$.

Proof. It is easy. □

6.5. Lemma. *Every infinite cardinal number is a limit ordinal number.*

Proof. It is easy. ☐

6.6. Lemma. *Let a be an infinite set. Then $a \times a$ is equipotent with a.*

Proof. Suppose, on the contrary, that there exists an infinite cardinal number c such that the cardinal number $d = \mathbf{card}(c \times c)$ is different from c, and take the least cardinal number c with this property. Clearly, $c < d$. Define a relation r on $c \times c$ in this way: $\langle\langle x, y\rangle, \langle u, v\rangle\rangle \in r$ if and only if either $\max(x, y) < \max(u, v)$ or $\max(x, y) = \max(u, v)$ and $x < u$ or $\max(x, y) = \max(u, v)$, $x = u$ and $y \le v$. One can easily check that r is a well ordering on $c \times c$. Consequently, by 5.8, there exists an ordinal number e and an isomorphism f of $c \times c$ onto e with respect to $r, \in_a^=$. We have $c < d \le e$. There exist two elements x, y of c with $f(\langle x, y\rangle) = c$. Put $z = \max(x, y) + 1$. Clearly, z is infinite and so $z < c$ by 6.5. Hence $\mathbf{card}(z) < c$ and by the minimality of c we get that $z \times z$ is equipotent with z. However, the range of the inverse of f is contained in $z \times z$, so that $c \le \mathbf{card}(z \times z)$ and we get a contradiction. ☐

6.7. Theorem. *Let a, b be two disjoint sets such that at least one of them is infinite. Then $\mathbf{card}(a \cup b) = \max(\mathbf{card}(a), \mathbf{card}(b))$.*

Let a, b be two nonempty sets such that at least one of them is infinite. Then $\mathbf{card}(a \times b) = \max(\mathbf{card}(a), \mathbf{card}(b))$.

Proof. It follows easily from 6.6. ☐

6.8. Theorem. *Let c be an infinite cardinal number. Let a be a set such that $\mathbf{card}(a) \le c$ and $\mathbf{card}(b) \le c$ for any $b \in a$. Then $\mathbf{card}(\bigcup a) \le c$.*

Proof. It is sufficient to prove this result under the assumption that a is nonempty and every element of a is nonempty. There exists a mapping f of c onto a. Also, there exists a mapping g with domain c such that for every $i \in c$, $g\text{'}i$ is a mapping of c onto $f\text{'}i$. For every $\langle i, j\rangle \in c \times c$ put $h\text{'}\langle i, j\rangle = (g\text{'}i)\text{'}j$. Then h is a mapping of $c \times c$ onto $\bigcup a$, so that $\mathbf{card}(\bigcup a) \le \mathbf{card}(c \times c) = c$ (we have used 6.1 and 6.7). ☐

6.9. Theorem. *The union of any set of cardinal numbers is a cardinal number.*

Proof. It is easy. ☐

6.10. Theorem. *For any set a, $\mathbf{card}(a) < \mathbf{card}(\mathbf{P}(a))$.*

Proof. Clearly, $\mathbf{card}(a) \le \mathbf{card}(\mathbf{P}(a))$. Suppose that there exists a bijection of a onto $\mathbf{P}(a)$. Denote by B the set of all elements $x \in a$ such that $x \notin f\text{'}x$. There is an element $b \in a$ with $f\text{'}b = B$. Then neither $b \in B$ nor $b \notin B$, a contradiction. ☐

6.11. Theorem. *The class of cardinal numbers is a proper class.*

Proof. It follows from 6.9 and 6.10. ☐

It follows that also the class of infinite cardinal numbers is a proper class, and hence there exists a unique isomorphism of **On** onto the class C of infinite cardinal numbers with respect to $\in_{\mathbf{On}}^{=}$ and $\in_{\mathbf{On}}^{=} \cap (C \times C)$. This bijection will be denoted by $\aleph$. For an ordinal number a we write $\aleph_a$ instead of $\aleph\text{'}a$. In particular, $\aleph_0 = \omega$.

By a *confinal subset* of an ordinal number a we mean any subset s of a such that $\bigcup s = a$. For an ordinal number a, the least cardinal number that equipotent with a confinal subset of a is called the *confinal* of a.

Clearly, if b is the confinal of a then $b \leq a$. The confinal of 0 is 0 and the confinal of any non-limit ordinal number is 1.

6.12. Theorem. *Let a be an ordinal number and b be its confinal. Then b is the ordinal type of a confinal subset of a.*

Proof. There is a bijection f of b onto a confinal subset of a. Let us define a function g with domain b by transfinite induction as follows: for $i \in b$, $g\text{'}i$ is the least ordinal number k such that $k > g\text{'}j$ and $k > f\text{'}j$ for all $j < i$. It is easy to see that the range of g is a confinal subset of a, and $i < j$ if and only if $g\text{'}i < g\text{'}j$. □

Consequently, the confinal of a could have been also defined as the least ordinal number that is the ordinal type of a confinal subset of a.

By a *regular cardinal number* we mean a cardinal number a such that the confinal of a equals a. Clearly, ω is the least regular cardinal number.

6.13. Theorem. *The confinal of any infinite limit ordinal number is a regular cardinal number.*

Proof. It is easy. □

6.14. Theorem. *If a is an infinite limit ordinal number then the confinal of $\aleph_a$ is equal to the confinal of a.*

Proof. It is easy. □

6.15. Theorem. *For every ordinal number a, $\aleph_{a+1}$ is a regular cardinal number.*

Proof. Suppose, on the contrary, that $\aleph_{a+1}$ has a confinal subset s of cardinality at most $\aleph_a$. Clearly, every element of s has cardinality at most $\aleph_a$. By 6.8, $\mathbf{card}(\bigcup s) \leq \aleph_a$. But $\bigcup s = \aleph_{a+1}$ and we get a contradiction. □

6.16. Theorem. *An infinite cardinal number c is regular if and only if $\mathbf{card}(\bigcup s) < c$ for any set s such that $\mathbf{card}(s) < c$ and $\mathbf{card}(x) < c$ for all $x \in s$.*

Proof. Let c be regular. If s is as above, then there exists an infinite cardinal number $d < c$ such that $\mathbf{card}(s) \leq d$ and $\mathbf{card}(x) \leq d$ for all $x \in s$. By 6.8, $\mathbf{card}(\bigcup s) \leq d$.

Let c be not regular. There exists a confinal subset s of c such that $\mathbf{card}(s) < c$. Then $\mathbf{card}(x) < c$ for all $x \in s$ and $\bigcup s = c$. □

Let C be a family of cardinal numbers, with domain I (so, I is a set). We put

$$\sum C = \mathbf{card}(\bigcup\{C\text{'}i \times \{i\} : i \in I\}) \qquad \text{and} \qquad \prod C = \mathbf{card}(\textstyle\prod_d C).$$

These cardinal numbers are called the sum and the product of the family C, respectively.

For two cardinal numbers a, b the cardinal number $\mathbf{card}(a^b)$ is also denoted by a^b. It should always follow from the context what does a^b mean. Actually, it can mean three different things: the cardinal number a^b, the set of all mappings of b into a, and if for example $b = 2$, the set of the ordered pairs of elements of a. We are not going to distinguish between these objects by some weird notation.

6.17. THEOREM. (1) *$c^0 = 0^c = 1$ for any cardinal number c; $0^d = 0$ for any cardinal number $d \neq 0$.*

(2) *If n, m are natural numbers then n^m is a natural number.*

(3) *$c < 2^c = \mathbf{card}(\mathbf{P}(c))$ for any cardinal number c.*

(4) *$d^n = d$ for any infinite cardinal number d and any natural number $n \neq 0$.*

(5) *If d is an infinite cardinal number and $2 \leq c \leq d$ then $c^d = 2^d$.*

PROOF. The first four statements are easy. Let $2 \leq c \leq d$ where d is infinite. Clearly, $2^d \leq c^d$. The set of all mappings of c into d is a subset of $\mathbf{P}(c \times d)$ where $\mathbf{card}(c \times d) = \max(c, d) = d$, so that $c^d \leq 2^d$. $\square$

Comments

It may seem strange that natural numbers were used (a little bit) already at the beginning of this chapter, while defined only later. Actually we did not mean to define old objects that were in use already before. We should distinguish between numbers 0, 1, 2, etc., as parts of our human language in which we can speak, among other things, about mathematics, and natural numbers as mathematical objects. Mixing the two things would lead to paradoxes. In fact, strictly speaking, in mathematics we do not define objects. We only introduce new formulas, accept their abbreviations and help ourselves to deal with them by imagining that they express some facts about some objects of mathematics. There may be a philosophical discussion about the existence of such mathematical objects, and we will not delve into such things. Anyway, mathematical objects should be considered as standing at a completely different level from that of metamathematics, which is a natural part of the human language.

The same situation will repeat later. In chapter 5 we will define formulas, variables, theories, proofs as mathematical objects. It is possible to imagine that they are extensions of the metamathematical concepts introduced in this present chapter, but it is better not to do so; they will have nothing in common with the metamathematical formulas, etc., except that we chose to use the same names in both cases.

From now on we will be a little bit less strict. If f is a function and a is an element in its domain, the value of f at a will not be denoted by $f'a$ but by $f(a)$, or sometimes also by f_a. If f is a function and a is a subset of its domain, the range of $f \restriction a$ will be also denoted by $f(a)$, unless this would lead to a misunderstanding. The class $A \smallsetminus B$ will be sometimes denoted by $A - B$. The empty set 0 will be sometimes denoted by $\emptyset$.

A few remarks should be made about various extensions of set theory.

As we have seen, axioms (C1) through (C9) provide the existence of the class of all sets with a given property only in the case when the property can be formulated in such a way that all quantifiers are restricted to sets. Imagine, for example, that we need to denote by K the class of all algebras A such that the variety generated by A has the amalgamation property (these notions will be defined in later chapters). The existence of K does not follow from the axioms of set theory immediately, since the property of A to be expressed needs to mention the existence of that variety, which is not a set. We have to avoid any mention of the variety generated by A. This can be done in this case, if we take a look at the definition of the amalgamation property. Without knowing what amalgamation property means, the introduction of the class K would be illegal. So, one should be careful. Such inconveniencies could be avoided if we added infinitely many axioms to the axioms of set theory: for each property, without the restriction on quantifiers, the existence of the corresponding class. We would obtain a larger system of axioms, perhaps a reasonable one, with infinitely many axioms instead of finitely many. This system would not be equivalent to its any finite subsystem. The resulting theory would be stronger. We will not do it.

It has been found that the sentence $2^{\aleph_0} = \aleph_1$, known as the continuum hypothesis, is both consistent and also independent with the axioms of set theory. This means that if we add either this sentence or its negation to the axioms, the resulting theory is in both cases consistent under the assumption that set theory is consistent. (Set theory is consistent most likely, but its consistency cannot be proved by finite means.)

Sometimes it is convenient to work in set theory extended by the generalized continuum hypothesis, which states that $2^{\aleph_\alpha} = \aleph_{\alpha+1}$ for all ordinal numbers α. This sentence is also both consistent and independent in the above sense.

There is an even stronger consistent axiom, that of constructibility, which has the generalized continuum hypothesis as its consequence.

A cardinal number κ is said to be strongly inaccessible if it is regular, larger than $\aleph_0$ and $2^\lambda < \kappa$ for every cardinal number $\lambda < \kappa$. The existence of a strongly inaccessible cardinal number implies the existence of a set U such that all axioms of set theory translate to provable sentences if all variables are let to run over subsets of U. However, the assumption of the existence would be a too strong assumption.

We are going to work in set theory with axioms (C1)–(C9) and (S1)–S(6) only.

CHAPTER 2

CATEGORIES

1. Basic definitions

We say that a *category* K is given if we are given

(1) a class K^o; elements of this class are called *objects* of K (or K-objects);
(2) a class K^m; elements of this class are called *morphisms* of K;
(3) two mappings κ_1, κ_2 of K^o into K^m; for a morphism a the objects $\kappa_1(a)$ and $\kappa_2(a)$ are called the *beginning* and the *end* of a, respectively; if $\kappa_1(a) = A$ and $\kappa_2(a) = B$, we write $a : A \to B$ and say that a is a morphism from A to B;
(4) a mapping, assigning to any pair a, b of morphisms with $\kappa_2(a) = \kappa_1(b)$ a morphism $ba : \kappa_1(a) \to \kappa_2(b)$ (instead of writing $\kappa_2(a) = \kappa_1(b)$ we can say that the product ba is defined);
(5) a mapping, assigning to any K-object A a K-morphism $1_A : A \to A$ (called the *identical morphism* of A);

the following two conditions must be satisfied:

(1) if $a : A \to B$, $b : B \to C$ and $c : C \to D$ then $(cb)a = c(ba)$ (so this morphism can be denoted by cba);
(2) if $a : A \to B$, then $a1_A = 1_B a = a$.

For a given category K we define a category K^∂, called the *dual* of K, in this way: K^∂ has the same objects and the same morphisms as K; the beginning of a morphism in K^∂ is its end in K, and the end of a morphism in K^∂ is its beginning in K; the product ba of two morphisms in K^∂ is the product ab in K; 1_A is the same in K^∂ as in K. Clearly, $(K^\partial)^\partial = K$ for every category K.

A morphism $a : A \to B$ is said to be an *isomorphism* if there exists a morphism $b : B \to A$ with $ba = 1_A$ and $ab = 1_B$. Clearly, the morphism b is uniquely determined by a; it is called the *inverse* of a. Two objects A, B are said to be isomorphic if there exists an isomorphism of A into B.

A morphism $a : A \to B$ is said to be a *monomorphism* if $ab = ac$ implies $b = c$ (for any object C and any morphisms $b : C \to A$ and $c : C \to A$).

A morphism $a : A \to B$ is said to be an *epimorphism* if it is a monomorphism in the dual category, i.e., if $ba = ca$ implies $b = c$.

It is easy to see that the product of two monomorphisms (if defined) is a monomorphism; if ba is a monomorphism, then a is a monomorphism. It follows by duality that the product of two epimorphisms is an epimorphism and if ba is an epimorphism, then b is an epimorphism. Every isomorphism is both a monomorphism and an epimorphism.

A category K is said to be *small* if its class of objects is a set. Of course, in that case also the class of morphisms is a set.

A category K is said to be *locally small* if for every object A there exists a set S of monomorphisms ending in A such that for every monomorphism a ending in A there are a monomorphism $b \in S$ and an isomorphism c with $a = bc$. A category is said to be *colocally small* if its dual is locally small.

Let K be a category and L be a subclass of K^o. Then L defines in a natural way a category, the objects of which are the elements of L and the morphisms of which are the morphisms a of K such that both the beginning and the end of a belongs to L. Such categories are called *full subcategories* of K; they can be identified with subclasses of K^o.

By a *functor* of a category K into a category L we mean a mapping F assigning to any K-object an L-object and to any K-morphism an L-morphism, such that the following three conditions are satisfied:

(1) if $a : A \to B$ in K, then $F(a) : F(A) \to F(B)$ in L;
(2) if $a : A \to B$ and $b : B \to C$ in K, then $F(ba) = F(b)F(a)$;
(3) for any K-object A, $1_{F(A)} = F(1_A)$.

Given two categories K and L and an L-object A, we define a functor $C_A = C_{A,K,L}$ of K into L in this way: $C_A(X) = A$ for every K-object X; $C_A(a) = 1_A$ for every K-morphism a. This functor is called the *constant functor* of K onto the L-object A.

Let F, G be two functors of a category K into a category L. By a *natural transformation* of F into G we mean a mapping μ, assigning to any K-object A an L-morphism $\mu_A : F(A) \to G(A)$, such that for every $a : A \to B$ in K we have $\mu_B F(a) = G(a)\mu_A$.

2. Limits and colimits

By a *diagram* in a category K we mean a pair, consisting of a small category D and a functor δ of D into K.

By a *limit* of a diagram D, δ in a category K we mean a K-object A together with a natural transformation μ of the constant functor C_A (of D into K) into δ such that for any K-object A' and any natural transformation μ' of $C_{A'}$ into δ there exists a unique morphism $a : A' \to A$ with $\mu'_i = \mu_i a$ for all D-objects i.

It is easy to see that the limit of a diagram D, δ in K (if it exists) is uniquely determined up to isomorphism in the following sense: if A, μ and A', μ' are two limits of D, δ in K, then there exists a unique isomorphism a of A into A' such that $\mu_i = \mu'_i a$ for all i.

It may be useful to remark that if A, μ is a limit of a diagram D, δ in K, and if a, b are two morphisms of a K-object B into A, then $a = b$ if and only if $\mu_i a = \mu_i b$ for all D-objects i.

Let H be a family (over a set I) of objects of a category K. We can define a diagram D, δ in K in this way: $D^o = I$; D has only identical morphisms; $\delta(i) = H_i$. A limit of this diagram is called a *product* of the family H in the kategory K. In other words, a product of H is a K-object A together with a family μ of morphisms $\mu_i : A \to H_i$ ($i \in I$) such that for any K-object A' and any family μ' of morphisms $\mu'_i : A' \to H_i$ there exists a unique morphism $a : A' \to A$ with $\mu'_i = \mu_i a$ for all i.

2.1. Theorem. *Products in a category K are associative in the following sense. Let A, μ be a product of a family H of K-objects over a set I; for every $i \in I$ let $A_i, \nu^{(i)}$ be a product of a family $G^{(i)}$ of K-objects over a set $J^{(i)}$. Denote by J the disjoint union of the sets $J^{(i)}$ and define a family G of K-objects by $G_j = G_j^{(i)}$, where i is the index such that $j \in J^{(i)}$. Then A, ν, where $\nu_j = \nu_j^{(i)} \mu_i$ with $j \in J^{(i)}$, is a product of G in K.*

Consequently, if every pair of K-objects has a product in K, then every nonempty finite family of K-objects has a product in K.

Proof. It is easy. □

For a K-object A and a set I, by an *I-power* of A in K we mean a product of the family H over I, defined by $H_i = A$ for all $i \in I$.

Let $a : A \to C$ and $b : B \to C$ be two K-morphisms with the same end. By a *pullback* of a, b we mean a K-object D together with two morphisms $c : D \to A$ and $d : D \to B$ such that $ac = bd$ and such that for any K-object D' and any morphisms $c' : D' \to A$ and $d' : D' \to A$ with $ac' = bd'$ there exists a unique morphism $e : D' \to D$ with $c' = ce$ and $d' = de$. Clearly, pullbacks of a, b are essentially the limits of the diagram D, δ defined in this way: D has three objects $1, 2, 3$ and two morphisms $p : 1 \to 3$, $q : 2 \to 3$ except the identical ones; $\delta(p) = a$ and $\delta(q) = b$.

2.2. Theorem. *Let D together with $c : D \to A$ and $d : D \to B$ be a pullback of two morphisms $a : A \to C$ and $b : B \to C$ in a category K. If a is a monomorphism then d is a monomorphism.*

Proof. It is easy. □

Let $a : A \to B$ and $b : A \to B$ be two K-morphisms with the same beginning and the same end. By an *equalizer* of a, b we mean a K-object C together with a morphism $c : C \to A$ such that $ac = bc$ and such that for any K-object C' and any morphism $c' : C' toA$ with $ac' = bc'$ there exists a unique morphism $d : C' \to C$ with $c' = cd$. Clearly, equalizers of a, b are essentially the limits of the diagram D, δ defined in this way: D has two objects $1, 2$ and two morphisms $p : 1 \to 2$ and $q : 1 \to 2$ except the identical ones; $\delta(p) = a$ and $\delta(q) = b$.

2.3. Theorem. *Let C together with $c : C \to A$ be an equalizer of a pair of morphisms in a category K. Then c is a monomorphism.*

Proof. It is easy. □

By a *final object* of a category K we mean a K-object A such that for every K-object B there exists precisely one morphism $a : B \to A$. Clearly, final objects of K are essentially the limits of the empty diagram in K.

By a *colimit* of a diagram D, δ in a category K we mean a limit of the diagram D^∂, δ in the category K^∂. By a *pushout* of a pair of morphisms with the same beginning in K we mean a pullback of these morphisms in K^∂. By a *coequalizer* of a pair of morphisms with the same beginning and the same end in K we mean an equalizer of these morphisms in K^∂. By an *initial object* of a category K we mean a final object of K^∂, i.e., an object A such that for any K-object B there exists precisely one morphism $a : A \to B$. If c, d is a pushout of a, b and a is an epimorphism, then d is an epimorphism. Every coequalizer is an epimorphism.

3. Complete and cocomplete categories

A category K is said to be *complete* if every diagram has a limit in K; it is said to be *cocomplete* if every diagram has a colimit in K.

3.1. THEOREM. *The following are equivalent for a category K:*

(1) *K is complete;*
(2) *every family of K-objects has a product in K and every pair of monomorphisms of K with the same end has a pullback in K;*
(3) *every family of K-objects has a product in K and every pair of K-morphisms with the same beginning and the same end has an equalizer in K.*

PROOF. Of course, (1) implies (3). Let us prove that (3) implies (2). Let $a : A \to C$ and $b : B \to C$ be two morphisms. Let P together with $p : P \to A$ and $q : P \to B$ be a product of the pair A, B. Let D together with $d : D \to P$ be an equalizer of the pair ap, bq. One can easily check that D together with pd, qd is a pullback of a, b in K.

It remains to prove that (2) implies (1). Let D, δ be a diagram in K. Denote by I the set of D-objects and by J the set of D-morphisms. For a morphism $j \in J$ denote by $\alpha(j)$ the begining and by $beta(j)$ the end of j. Define a family of K-objects E over I by $E_i = \delta(i)$ for $i \in I$; define a family F of K-objects over J by $F_j = \delta(\beta(j))$ for $j \in J$. Let P together with p be a product of E and let R together with r be a product of F in K. Since R is a product, there exists a unique morphism $h : P \to R$ with $p_{\beta(j)} = r_j h$ for all $j \in J$; also, there exists a unique morphism $k : P \to R$ with $\delta(j)p_{\alpha(j)} = r_j k$ for all $j \in J$.

Let us prove that h and k are monomorphisms. Let $ha = hb$ and $kc = kd$ for some morphisms a, b, c, d. In order to prove that $a = b$ and $c = d$, it is sufficient to prove that $p_i a = p_i b$ and $p_i c = p_i d$ for all $i \in I$. Where $j = 1_i$, we have

$$p_i a = p_{\beta(j)} a = r_j ha = r_j hb = p_{\beta(j)} b = p_i b,$$
$$p_i c = \delta(j)p_{\alpha(j)} c = r_j kc = r_j kd = \delta(j)p_{\alpha(j)} d = p_i d.$$

The pair h, k is a pair of monomorphisms with the same end R (and also the same beginning P). Let A together with h', k' be a pullback of this pair; its existence follows from (2). For all $i \in I$ we have

$$p_i h' = p_{\beta(j)} h' = r_j h h' = r_j k k' = \delta(j) p_{\alpha(j)} k' = p_i k'$$

(where again $j = 1_i$), so that $h' = k'$.

Let us define a natural transformation μ of the constant functor C_A into δ by $\mu_i = p_i h' = p_i k'$ for all $i \in I$. This is a natural transformation, since for a D-morphism $j : i_1 \to i_2$ we have

$$\begin{aligned} \delta(j)\mu_{i_1} &= \delta(j) p_{i_1} k' = \delta(j) p_{\alpha(j)} k' = r_j k k' \\ &= r_j h h' = p_{\beta(j)} h' = p_{i_2} h' = \mu_{i_2}. \end{aligned}$$

We are going to show that A together with μ is a limit of D, δ in K. Let B be a K-object and μ' be a natural transformation of C_B into δ. Since P is a product, there exists a unique morphism $v : B \to P$ with $\mu'_i = p_i v$ for all $i \in I$. For $j \in J$ we have

$$r_j h v = p_{\beta(j)} v = \mu'_{\beta(j)} = \delta(j)\mu'_{\alpha(j)} = \delta(j) p_{\alpha(j)} v = r_j k v,$$

so that $hv = kv$. By the definition of pullback there exists a unique morphism $a : B \to A$ with $v = h'a$, i.e., a unique morphism $a : B \to A$ such that $p_i v = p_i h' a$ for all $i \in I$, i.e., $\mu'_i = \mu_i a$ for all $i \in I$. □

3.2. Theorem. *Let K be a category such that K is either complete and locally small or cocomplete and colocally small. Then for every morphism $a : A \to B$ of K there exist a K-object C, an epimorphism $b : A \to C$ and a monomorphism $c : C \to B$ such that $a = cb$.*

Proof. Since the assertion is self-dual, it is sufficient to prove the theorem under the assumption that K is complete and locally small. There exists a set Y_0 of monomorphisms ending in B such that for every monomorphism m ending in B there are a monomorphism $m' \in Y_0$ and an isomorphism h with $m = m'h$. Denote by Y the set of the monomorphisms $f \in Y_0$ for which there exists a morphism f' with $a = ff'$. Define a diagram D, δ in K in this way: $D^o = Y \cup \{i_0\}$ where i_0 is an element not belonging to Y; except for the identical morphisms, the category D contains a unique morphism $p_f : f \to i_0$ for every $f \in Y$; $\delta(p_f) = f$. Let C together with μ be a limit of the diagram D, δ in K. It is easy to see that the morphisms μ_f are monomorphisms. Put $c = \mu_{i_0}$, so that $f\mu_f = c$ for all $f \in Y$. Then c is a monomorphism, since it is a product of two monomorphisms. For every $f \in Y$ take a morphism f' with $a = ff'$. By the definition of a limit, there exists a morphism $b : A \to C$ such that $f' = \mu_f b$ for all $f \in Y$. For any $f \in Y$ we have $a = ff' = f\mu_f b = cb$. So, it remains to prove that b is an epimorphism. Let $ub = vb$. Let E together with e be an equalizer of the pair u, v, so that e is a monomorphism and ce is a monomorphism ending in B. There exist a monomorphism $f_0 \in Y_0$ and an isomorphism i such that $f_0 = cei$. By the definition of equalizer there exists a morphism $\bar{e} : A \to E$ with $b = e\bar{e}$, so that $a = ce\bar{e}$; we get $f_0 \in Y$. We have $cei\mu_{f_0} = f_0\mu_{f_0} = c = c1_C$; since c is a monomorphism, it follows that $ei\mu_{f_0} = 1_C$. Hence $u = u1_C = uei\mu_{f_0} = vei\mu_{f_0} = v1_C = v$. □

4. Reflections

Let L be a full subcategory of a category K and let A be a K-object. By a *reflection* of A in Lwe mean an L-object B together with a morphism $a : A \to B$ such that for any L-object C and any morphism $b : A \to C$ there exists a unique morphism $c : B \to C$ with $b = ca$. Clearly, a reflection of a given object is unique up to isomorphism (if it exists) in the obvious sense. A full subcategory L of K (or a subclass of K^o) is said to be *reflective* if every K-object has a reflection in L.

4.1. THEOREM. *Let L be a reflective full subcategory of a category K such that L is closed under isomorphisms* (*i.e., for any isomorphism a of K, the beginning of a belongs to L if and only if the end of a belongs to L*) *and let D, δ be a diagram in L.*

(1) *Let A, μ be a limit of D, δ in K. Then $A \in L$ and A, μ is a limit of D, δ in L.*
(2) *Let A, μ be a colimit of D, δ in K and let B together with $a : A \to B$ be a reflection of A in L. Then B together with ν, where $\nu_i = a\mu_i$ for all $i \in D^o$, is a colimit of d, δ in L.*

Consequently, if K is complete then L is complete; if K is cocomplete then L is cocomplete.

PROOF. (1) Let $a : A \to B$ be a reflection of A in L. For every $i \in D^o$ there exists a unique morphism $\nu_i : B \to \delta(i)$ with $\mu_i = \nu_i a$. For a D-morphism $e : i \to j$ we have $\delta(e)\nu_i a = \delta(e)\mu_i = \mu_j = \nu_j a$; since a is a reflection, we get $\delta(e)\nu_i = \nu_j$, which shows that ν is a natural transformation. By the definition of a limit there exists a unique morphism $b : B \to A$ such that $\nu_i = \mu_i b$ for all $i \in D^o$. For $i \in D^o$ we have $\mu_i ba = \nu_i a = \mu_i = \mu_i 1_A$, from which we get $ba = 1_A$. We have $aba = a1_A = a$; from this we get $ab = 1_B$, since according to the definition of a reflection there is only one morphism $c : B \to B$ such that $a = ca$, and both ab and 1_B have this property. So, a is an isomorphism and $B \in l$. The rest is clear.

(2) Clearly, ν is a natural transformation of δ into the constant functor C_B. let E be an L-object and κ be a natural transformation of δ into C_E. Since A, μ is a colimit, there exists a unique morphism $b : A \to E$ such that $\kappa_i = b\mu_i$ for all $i \in D^o$. By the definition of a reflection there exists a unique morphism $c : B \to D$ such that $b = ca$, i.e., a unique morphism such that $b\mu_i = ca\mu_i$ for all $i \in D^o$, i.e., $\kappa_i = c\nu_i$. □

CHAPTER 3

STRUCTURES AND ALGEBRAS

1. Languages, structures, algebras, examples

By a *language* we mean a mapping σ, the domain of which is any set and the range of which is a set of integers. By a symbol of σ (or σ-symbol) we mean an element of the domain of σ. A σ-symbol s is said to be an *operation symbol* if $\sigma(s) \geq 0$; it is said to be a *relation symbol* if $\sigma(s) < 0$. For an operation symbol s of σ, the number $\sigma(s)$ is called its arity. For a relation symbol s, the arity of s is the number $-\sigma(s)$. Thus the arity of an operation symbol is a nonnegative integer, while the arity of a relation symbol is a positive integer. Operation symbols of arity 0 are called constants. Symbols of arity 1 are called unary, and symbols of arity 2 are called binary. By a purely relational language we mean a language without operation symbols. By an algebraic language, or *signature*, we mean a language without relation symbols.

Let n be a positive integer. By a relation of arity n (or n-ary relation) on a set A we mean a subset of A^n. Thus a unary (i.e., 1-ary) relation on A is a subset of A. A binary (i.e., 2-ary) relation is a relation in the previous sense. We often write $a\ r\ b$ instead of $\langle a, b\rangle \in r$ for a binary relation r.

Let n be a nonnegative integer. By a *partial operation* of arity n (or n-ary partial operation) on a set A we mean a mapping of a subset of A^n into A; a partial operation is said to be an *operation* if the domain is the set A^n. Thus nullary operations on A are in a natural one-to-one correspondence with elements of A, and will be usually identified with them. Unary partial operations on A are just mappings of a subset of A into A, and unary operations are mappings of A into A. If f is a binary operation on A, we often write $a\ f\ b$ instead of $f(a, b) = f`\langle a, b\rangle$.

By a *partial structure* of a language σ (or just partial σ-structure) we mean a pair $\langle A, p\rangle$ such that A is a nonempty set and p is a mapping, assigning to any relation symbol R of σ a relation of the same arity on A and to any operation symbol F of σ a partial operation of the same arity on A; if all the partial operations are operations then $\langle A, p\rangle$ is said to be a *structure*. By a *partial algebra* we mean a partial structure of an algebraic language. By an *algebra* we mean a structure of an algebraic language. The set A is called the *underlying set* of $\langle A, p\rangle$, and will be often identified with the structure. The relation, or partial operation $p(S)$ will be denoted by S_A (or just by S, if there is no confusion).

By the cardinality of a partial structure we mean the cardinality of its underlying set. Partial structures of cardinality 1 are called *trivial*; nontrivial partial structures are those of cardinality at least 2. A class of partial structures is said to be nontrivial if it contains at least one nontrivial partial structure.

Let σ be a given language. Unless otherwise stated, all symbols and partial structures will be symbols and partial structures of this one fixed language.

Observe that a subset of a language is a language. If $\tau \subseteq \sigma$ and A is a partial σ-structure then the partial τ-structure B with the same underlying set and $S_B = S_A$ for all $S \in \mathbf{Dom}(\tau)$ is called the *reduct* of A to τ, or the underlying partial τ-structure of A. The reduct of A to the set of operation symbols of σ is called the underlying partial algebra of A.

Algebras of the signature, containing just one binary operation symbol $\cdot$, are called *groupoids*. For two elements a, b of a groupoid A, we usually write ab instead of $\cdot_A(a, b)$. (If two groupoids are under consideration at a time and there may be elements belonging to both of them, we should say something like '$ab = c$ in A' instead of just '$ab = c$'.) In more complicated expressions, it is necessary to use parentheses; in order to avoid writing too many of them, let us make the following convention: $a_1a_2 \dots a_n$ stands for $((a_1a_2)\dots)a_n$, $ab \cdot cd$ stands for $(ab)(cd)$, $ab(c \cdot de)f$ stands for $((ab)(c(de)))f$, etc. This convention will be used also for arbitrary languages extending the signature of groupoids.

For every groupoid A we can define a groupoid B with the same underlying set by $ab = c$ in B if and only if $ba = c$ in A. We call B the *groupoid dual to* A.

A groupoid A is said to be *idempotent* if it satisfies $aa = a$ for all $a \in A$. It is said to be *commutative* if it satisfies $ab = ba$ for all $a, b \in A$. It is said to be *associative*, or to be a *semigroup*, if it satisfies $(ab)c = a(bc)$ for all $a, b, c \in A$. By a *semilattice* we mean an idempotent commutative semigroup.

By an *annihilating element* (or *zero element*) of a groupoid A we mean an element a such that $ax = xa = a$ for all $x \in A$. By a *unit element* of a groupoid A we mean an element a such that $ax = xa = x$ for all $x \in A$. It is easy to see that a groupoid contains at most one annihilating element and also at most one unit element.

By a *monoid* we mean an algebra of the signature $\{\cdot, 1\}$ where $\cdot$ is a binary operation symbol and 1 is a constant, such that its reduct to $\{\cdot\}$ is a semigroup with unit element 1.

For every element a of a semigroup A and every positive integer k we define the element a^k of A as follows: $a^1 = a$; $a^{k+1} = a^k a$. If A is a semigroup with unit 1, this definition can be extended to all nonnegative integers k by $a^0 = 1$.

Let A be a semigroup with unit 1 and let $a \in A$. An element $b \in A$ is called the *inverse* of a if $ab = ba = 1$. Clearly, every element of A has at most one inverse.

By a *cancellation groupoid* we mean a groupoid A such that $ab = ac$ implies $b = c$ and $ba = ca$ implies $b = c$ (for all $a, b, c \in A$). By a *division groupoid* we mean a groupoid A such that for every pair a, b of elements of A there exist elements $c, d \in A$ such that $ac = b$ and $da = b$.

By a *quasigroup* we mean an algebra of the signature $\{\cdot, /, \backslash\}$ such that

$$(a/b)b = a,$$
$$b(b\backslash a) = a,$$
$$(ab)/b = a,$$
$$b\backslash(ba) = a$$

for all $a, b \in A$.

It is easy to see that a quasigroup is uniquely determined by its groupoid reduct. The groupoid reducts of quasigroups are precisely the cancellation division groupoids.

By a *loop* we mean an algebra A of the signature $\{\cdot, /, \backslash, 1\}$ such that the reduct of A to $\{\cdot, /, \backslash\}$ is a quasigroup and 1 is a unit of A.

By a *group* we mean an algebra A of the signature $\{\cdot, *, 1\}$, where $*$ is a unary operation symbol, such that the reduct of A to $\{\cdot, 1\}$ is a monoid and for every $a \in A$, $*a$ is the inverse of a. For every element a of a group A and every integer k we define an element $a^k \in A$ as follows: if $k \geq 0$, define it as above; if $k < 0$, put $a^k = (*a)^{-k}$. (Thus $*a = a^{-1}$, which is a more usual notation.)

It is easy to see that a group is uniquely determined by its groupoid reduct. The groupoid reducts of groups are precisely the division semigroups.

By an *Abelian group* we mean an algebra of the signature $\{+, -, 0\}$, where $+$ is a binary, $-$ is a unary operation symbol and 0 is a constant, such that

$$(a+b)+c = a+(b+c),$$
$$a+b = b+a,$$
$$a+0 = a,$$
$$a+(-a) = 0$$

for all $a, b, c \in A$. Clearly, there is essentially no difference between Abelian groups and commutative groups. For two elements a, b of an Abelian group we write $a - b$ instead of $a + (-b)$; for an integer k, the element a^k of the corresponding commutative group is denoted by ka.

By a *ring* we mean an algebra R of the signature $\{+, \cdot, -, 0, 1\}$ such that the reduct of R to $\{+, -, 0\}$ is an Abelian group, the reduct of R to $\{\cdot, 1\}$ is a monoid and

$$a(b+c) = ab + ac,$$
$$(b+c)a = ba + ca$$

for all $a, b, c \in R$. A ring R is said to be commutative if $ab = ba$ for all $a, b \in R$.

By a *division ring* we mean a ring R such that every element $a \in R \smallsetminus \{0\}$ has an inverse element (an element b with $ab = ba = 1$). A *field* is a commutative division ring.

Let R be a ring. By an *R-module* we mean an algebra A of the signature $\{+, -, 0\} \cup R$, where the elements of R are taken as unary operation symbols, such that the reduct of A to $\{+, -, 0\}$ is an Abelian group and

$$r(a+b) = ra + rb,$$
$$(r+s)a = ra + sa,$$
$$(rs)a = r(sa),$$

$$1a = a$$

for all $r, s \in R$ and $a, b \in A$. If R is a field, then R-modules are called *vector spaces* over R.

Structures of the language, containing just one relation symbol $\to$, are called *graphs*. For every graph A we can define a graph B with the same underlying set by $a \to b$ in B if and only if $b \to a$ in A. We call B the *graph dual to* A. A graph A is said to be *reflexive* if $a \to a$ for all $a \in A$. It is said to be *antireflexive* if $a \not\to a$ for all $a \in A$. It is said to be *symmetric* if $a \to b$ implies $b \to a$. It is said to be *antisymmetric* if $a \to b$ and $b \to a$ imply $a = b$. It is said to be *transitive* if $a \to b$ and $b \to c$ imply $a \to c$.

A *quasiordered set* is a reflexive, transitive graph. An *ordered set* is an antisymmetric quasiordered set. If A is a quasiordered set then we write $a \leq b$ instead of $a \to b$; we write $a < b$ if $a \leq b$ and $b \nleq a$.

Let A be an ordered set. For two elements $a, b \in A$ such that $a \leq b$, the set $\{x \in A : a \leq x \leq b\}$ is denoted by $[a, b]$; such subsets of A are called *intervals*. An element a is said to be covered by an element b (and b is said to be a *cover* of a) if $a < b$ and there is no $c \in A$ with $a < c < b$. Clearly, a finite ordered set is uniquely determined by the set of its cover relations (the set of the pairs $\langle a, b \rangle$ such that a is covered by b).

By an *atom* of an ordered set with the least element o we mean any element that covers o. Coatoms are defined dually.

Let A be an ordered set. An element $c \in A$ is called the *meet* of two elements a, b in A if $c \leq a$, $c \leq b$ and $d \leq c$ for any element $d \in A$ such that $d \leq a$ and $d \leq b$. The notion of the *join* of two elements in an ordered set can be defined dually (i.e., the join of a, b in A is the meet of a, b in the dual of A). Clearly, every pair of elements of A has at most one meet and at most one join. By a *meet-semilattice* we mean an ordered set in which every two elements have a meet. By a *join-semilattice* we mean a dual of a meet-semilattice. By a *lattice ordered set* we mean an ordered set that is both a meet- and a join-semilattice.

There is a natural one-to-one correspondence between semilattices and meet-semilattices. For a given semilattice, the corresponding meet-semilattice is defined by $a \leq b$ iff $ab = a$. Given a meet-semilattice, the corresponding semilattice is defined by taking ab to be the meet of a, b. This makes it possible to identify semilattices with semilattice-ordered sets. (Of course, dually, there is also a one-to-one correspondence between semilattices and join-semilattices.)

By a *lattice* we mean an algebra of the signature, containing two binary symbols $\wedge$ and $\vee$ (meet and join), and satisfying

$$\begin{aligned} &(a \wedge b) \wedge c = a \wedge (b \wedge c), && (a \vee b) \vee c = a \vee (b \vee c), \\ &a \wedge b = b \wedge a, && a \vee b = b \vee a, \\ &(a \vee b) \wedge a = a, && (a \wedge b) \vee a = a \end{aligned}$$

for all a, b, c.

Similarly as for semilattices, there is a natural one-to-one correspondence between lattices and lattice-ordered sets; the two will be usually identified.

By a *complete lattice* we mean a lattice in which every subset has the meet. (The meet $\bigwedge S$ of a subset S is an element a such that $a \leq x$ for all $x \in S$, and $b \leq a$ for any element b such that $b \leq x$ for all $x \in S$; the join $\bigvee S$ of S is defined dually.) It is easy to see that in a complete lattice, every subset has also the join.

By an *ideal* of a lattice A we mean a nonempty subset X of A such that $a \leq b \in X$ implies $a \in X$ and $a, b \in X$ implies $a \vee b \in X$. For every element $a \in L$, the set $\{x \in A : x \leq a\}$ is an ideal, called the *principal ideal* of A generated by a. *Filters* and *principal filters* are defined dually.

The intersection of any nonempty set of ideals of a lattice A is an ideal if it is nonempty. Consequently, the set of all ideals of A, together with the empty set, is a complete lattice with respect to inclusion; we call it the lattice of ideals of A. Its subset consisting of the principal ideals of A is a sublattice isomorphic to A. Similarly, the set of all filters of A is a complete lattice, called the lattice of filters of A.

An element a of a complete lattice L is said to be *compact* if for any subset S of L, $a \leq \bigvee S$ implies $a \leq \bigvee S'$ for some finite subset S' of S. By an *algebraic lattice* we mean a complete lattice L such that every element of L is the join of a set of compact elements of L.

2. Homomorphisms

Let A and B be two partial σ-structures. By a *homomorphism* of A into B we mean a mapping h of A into B satisfying the following two conditions:

(1) whenever R is an n-ary relation symbol of σ then $\langle a_1, \dots, a_n \rangle \in R_A$ implies $\langle h(a_1), \dots, h(a_n) \rangle \in R_B$
(2) whenever F is an n-ary operation symbol of σ $F_A(a_1, \dots, a_n) = a$ implies $F_B(h(a_1), \dots, h(a_n)) = h(a)$.

The second condition can be also stated as

$$h(F_A(a_1, \dots, a_n)) = F_B(h(a_1), \dots, h(a_n))$$

whenever the left side is defined.

If f is a homomorphism of A into B and g is a homomorphism of B into C, then the composition gh is a homomorphism of A into C.

By an *isomorphism* of A onto B we mean a bijection h of A onto B such that h is a homomorphism of A into B and the inverse h^{-1} is a homomorphism of B into A. Clearly, if f is an isomorphism of A onto B, then f^{-1} is an isomorphism of B onto A. We write $A \simeq B$ if A, B are *isomorphic*, i.e., if there exists an isomorphism of A onto B.

Observe that for a given signature σ, any two trivial σ-algebras are isomorphic. We also express this fact by saying that there is just one trivial σ-algebra up to isomorphism.

2.1. THEOREM. *Let A and B be two σ-algebras. A bijection of A onto B is an isomorphism of A onto B if and only if it is a homomorphism of A into B.*

PROOF. It is easy. $\square$

By an *endomorphism* of a partial structure A we mean a homomorphism of A into A. By an *automorphism* of A we mean an isomorphism of A onto A. For any partial structure A, $\mathbf{id}_A$ is an automorphism of A.

The following observation is often used to prove that a given homomorphism is an isomorphism.

2.2. THEOREM. *A homomorphism f of A into B is an isomorphism of A onto B if and only if there exists a homomorphism g of B into A such that $gf = \mathbf{id}_A$ and $fg = \mathbf{id}_B$.*

PROOF. It is easy. $\square$

The set of endomorphisms of an algebra A is a monoid with respect to composition. The set of automorphisms of A is a group with respect to composition. These are called the *endomorphism monoid* and *automorphism group* of A.

Let K be a class of partial structures of a language σ and L be a class of partial structures of a language τ. By an *equivalence between* K and L we mean a bijection ε of K onto L such that for any $A \in K$, the partial structures A and $\varepsilon(A)$ have the same underlying sets and for any $A, B \in K$ and any mapping f of A into B, f is a homomorphism of A into B if and only if f is a homomorphism of $\varepsilon(A)$ into $\varepsilon(B)$. We say that the two classes are *equivalent* if there exists an equivalence between them.

It is easy to see that the class of groups is equivalent with the class of division semigroups. (On the other hand, the class of quasigroups is not equivalent with the class of cancellation division groupoids.)

3. Substructures

Let A be a partial σ-structure, and let S be a nonempty subset of A. We can define a partial σ-structure B with the underlying set S as follows: if R is an n-ary relation symbol of σ then $R_B = R_A \cap B^n$; if F is an n-ary operation symbol of σ then F_B is the restriction of F_A to the set of the n-tuples $\langle a_1, \dots, a_n \rangle \in S^n$ such that the element $F_A(a_1, \dots, a_n)$ is defined and belongs to S. This partial structure is called the *partial substructure* of A determined by S; it is denoted by $A \restriction S$. The reduct of $A \restriction S$ to a sublanguage τ of σ is denoted by $A \restriction S, \tau$. If σ is an algebraic language then partial substructures are called partial subalgebras.

So, for a given partial structure A, every nonempty subset of A is the underlying set of precisely one partial substructure of A. Observe that a partial subalgebra of an algebra is not necessarily an algebra.

By a *subuniverse* of a partial σ-structure A we mean a subset S of A such that $F_A(a_1, \dots, a_n) \in S$ for any n-ary operation symbol F of σ and any n-tuple

$\langle a_1, \dots, a_n \rangle \in S^n$ such that $F_A(a_1, \dots, a_n)$ is defined. By a *substructure* of a partial structure A we mean any partial substructure, the underlying set of which is a subuniverse. A subalgebra of a partial algebra is not necessarily an algebra. A subalgebra of an algebra is an algebra.

3.1. THEOREM. *Let A, B be two partial structures. Then B is a partial substructure of A if and only if $B \subseteq A$, $\mathbf{id}_B$ is a homomorphism of B into A and $\mathbf{id}_B$ is a homomorphism of C into B for any partial structure C with the underlying set B such that $\mathbf{id}_B$ is a homomorphism of C into A.*

An algebra B is a subalgebra *of an algebra A if and only if $B \subseteq A$ and $\mathbf{id}_B$ is a homomorphism of B into A.*

PROOF. It is easy. □

Clearly, a subset S of a structure A is a subuniverse of A if and only if it is either the underlying set of a substructure of A, or else σ contains no constants and S is empty.

By an *embedding* of a partial structure A into a partial structure B we mean an isomorphism of A onto a partial substructure of B. We say that A can be embedded into B if there exists such an ambedding. Clearly, an algebra A can be embedded into an algebra B if and only if there exists an injective homomorphism of A into B.

Clearly, the intersection of any nonempty collection of subuniverses of a partial structure A is a subuniverse of A. It follows that for every subset S of A there exists the least subuniverse of A containing S; it is called the subuniverse of A *generated by* S, and is denoted by $\mathbf{Sg}(S)$. If $\mathbf{Sg}(S)$ is nonempty, then the unique substructure of A with the underlying set $\mathbf{Sg}(S)$ is called the substructure of A generated by S. If $A = \mathbf{Sg}(S)$, then S is said to be a *generating subset* of A (or a set of generators of A). A partial structure is said to be *finitely generated* it is has a finite generating subset.

3.2. THEOREM. *Let f be a homomorphism of a partial structure A into a partial structure B. Then for every subuniverse S of B, f^{-1}''S is a subuniverse of A; if A is a structure then for every subuniverse S of A, f''S is a subuniverse of B.*

PROOF. It is easy. □

3.3. THEOREM. *Let f, g be two homomorphisms of a partial structure A into a partial structure B. Then $\{a \in A : f(a) = g(a)\}$ is a subuniverse of A. Consequently, if two homomorphisms f, g of A into B coincide on a generating subset of A, then $f = g$.*

PROOF. It is easy. □

3.4. THEOREM. *Let S be a generating subset of a partial structure A. Then $\mathbf{card}(A) \leq \max(\omega, \mathbf{card}(S), \mathbf{card}(\sigma))$.*

PROOF. Denote by k the maximum of the three cardinal numbers. Define subsets $S_0 \subseteq S_1 \subseteq \dots$ of A as follows: $S_0 = S$; S_{i+1} is the set of the elements that either belong to S_i or can be expressed as $F_A(a_1, \dots, a_n)$ for an

n-ary operation symbol F and an n-tuple $a_1, \dots, a_n$ of elements of S_i. Clearly, $\mathbf{card}(S_i) \le k$ for all i and A is the union of this chain of subsets. □

The set of subuniverses of a given partial structure A is a complete lattice with respect to inclusion. One can easily prove that a subuniverse of a partial structure A is a compact element of the lattice of subuniverses of A if and only if it is a finitely generated subuniverse of A. Consequently, the lattice of subuniverses of A is an algebraic lattice.

4. Congruences

By a *congruence* of a partial structure A we mean an equivalence relation r on A such that for any n-ary operation symbol F, $\langle a_1, b_1\rangle \in r, \dots, \langle a_n, b_n\rangle \in r$ imply $\langle F_A(a_1, \dots, a_n), F_A(b_1, \dots, b_n)\rangle \in r$ whenever $F_A(a_1, \dots, a_n)$ and $F_A(b_1, \dots, b_n)$ are both defined.

It is easy to see that the intersection of any nonempty set of congruences of A is a congruence. Consequently, the set of congruences of A is a complete lattice with respect to inclusion. The congruence lattice of A will be denoted by $\mathbf{Con}(A)$.

For a partial structure A, $\mathbf{id}_A$ is the least and A^2 is the greatest congruence of A. A partial structure is said to be *simple* if it has precisely two congruences (so, it must be nontrivial and the two congruences are $\mathbf{id}_A$ and A^2).

For a binary relation r on A, the congruence of A generated by r (the intersection of all congruences containing r) is denoted by $\mathbf{Cg}_A(r)$. For a pair $\langle a, b\rangle$ of elements of A we put $\mathbf{Cg}_A(a, b) = \mathbf{Cg}_A(\{\langle a, b\rangle\})$; these congruences are called *principal*. By a finitely generated congruence of A we mean any congruence of the form $\mathbf{Cg}(r)$, where r is a finite relation on A.

Let r be a congruence of a partial structure A. For $a \in A$, a/r is the block of r containing a. We define a partial structure A/r (of the same language) with the underlying set $\{a/r : a \in A\}$ as follows: for an n-ary relation symbol R, $\langle b_1, \dots, b_n\rangle \in R_{A/r}$ if and only if there exist elements $a_1 \in b_1, \dots, a_n \in b_n$ with $\langle a_1, \dots, a_n\rangle \in R_A$; for an n-ary operation symbol F and elements $b_1, \dots, b_n \in A/r$, $F_{A/r}(b_1, \dots, b_n)$ is defined if and only if there exist elements $a_1 \in b_1, \dots, a_n \in b_n$ such that $F_A(a_1, \dots, a_n)$ is defined; in the positive case we put $F_{A/r}(b_1, \dots, b_n) = F_A(a_1, \dots, a_n)/r$. (It follows from the definition of congruence that this definition is correct.) The partial structure A/r is called the *factor* of A through r. Of course, any factor of a structure is a structure; any factor of an algebra is an algebra.

Clearly, the mapping $a \to a/r$ is a homomorphism of A onto A/r. This mapping is called the *canonical homomorphism* of A onto A/r; it is denoted by π_r.

4.1. THEOREM. *The kernel of any homomorphism of a partial structure A into any partial structure is a congruence of A. Any congruence r of a partial structure A is the kernel of the canonical homomorphism of A onto A/r.*

PROOF. It is evident. □

4.2. THEOREM. *Let f be a homomorphism of an algebra A onto an algebra B and let r be a congruence of A such that $\mathbf{ker}(f) \subseteq r$. Then there exists a unique mapping g of B into A/r, such that gf is the canonical homomorphism of A onto A/r. This mapping g is a homomorphism of B onto A/r. If $r = \mathbf{ker}(f)$, then g is an isomorphism of B onto A/r. Consequently, every homomorphic image of an algebra A is isomorphic to a factor of A.*

PROOF. It is easy. □

4.3. THEOREM. *Let r be a congruence of an algebra A. For any congruence s of A/r define a congruence $f(s)$ of A by $\langle a, b\rangle \in f(s)$ if and only if $\langle a/r, b/r\rangle \in s$. Then f is an isomorphism of the congruence lattice of A/r onto the principal filter of the congruence lattice of A generated by r. For any congruence s of A/r, the algebras $(A/r)/s$ and $A/f(s)$ are isomorphic.*

PROOF. It follows from 4.2. □

If r and t are two congruences of an algebra A such that $r \subseteq t$, then the congruence of A/r corresponding to t will be denoted by t/r. Thus $\langle a/r, b/r\rangle \in t/r$ if and only if $\langle a, b\rangle \in t$.

4.4. THEOREM. *For any algebra A, $\mathbf{Con}(A)$ is a complete sublattice of the lattice of equivalences on A.*

PROOF. For a nonempty set R of equivalences on A, $\bigwedge R$ is the intersection of R and $\bigvee R$ is the set of ordered pairs $\langle a, b\rangle$ for which there exists a finite sequence $a_0, \dots, a_k$ of elements of A with $a_0 = a$, $a_k = b$, such that for any $i \in \{1, \dots, n\}$ there is an $s_i \in R$ with $\langle a_{i-1}, a_i\rangle \in s_i$. It is not difficult to prove that if R is a set of congruences, then both $\bigwedge R$ and $\bigvee R$ are again congruences. □

Clearly, finitely generated congruences of A are precisely the compact elements of the lattice $\mathbf{Con}(A)$. It follows that $\mathbf{Con}(A)$ is an algebraic lattice.

A congruence r of an algebra A is said to be *fully invariant* if $\langle a, b\rangle \in r$ implies $\langle f(a), f(b)\rangle \in r$ for every endomorphism f of A. It is said to be *invariant* if $\langle a, b\rangle \in r$ implies $\langle f(a), f(b)\rangle \in r$ for every automorphism f of A.

5. Direct and subdirect products

Let H be a family of sets over I, i.e., a mapping with domain I; for $i \in I$ write $H_i = H(i)$. Recall that the direct product ΠH of H is the set of all mappings f with domain I such that $f(i) \in H_i$ for all $i \in I$. For $i \in I$, the mapping $f \to f(i)$ of ΠH into H_i is called the i-th *projection* of ΠH to H_i.

Now let H be a family of partial σ-structures over a set I. We can define a partial structure B with the underlying set ΠH in this way: if R is an n-ary relation symbol then $\langle f_1, \dots, f_n\rangle \in R_B$ if and only if $\langle f_1(i), \dots, f_n(i)\rangle \in R_{H_i}$ for all $i \in I$; if F is an n-ary operation symbol then $F_B(f_1, \dots, f_n) = f$ if and only if $F_{H_i}(f_1(i), \dots, f_n(i)) = f(i)$ for all $i \in I$. This partial structure is called the *direct product* of the family H, and is also denoted by ΠH. So, the operations of the direct product are defined componentwise. Clearly, the i-th projection is a homomorphism of ΠH onto H_i for any $i \in I$.

Observe that according to this definition, the direct product of the empty family of partial structures is a one-element partial structure.

5.1. THEOREM. *Let H be a family of algebras over a set I. The direct product of H is the unique algebra with the underlying set ΠH such that for any $i \in I$, the i-th projection is a homomorphism of the algebra onto H_i.*

PROOF. It is easy. □

For a partial structure A and an arbitrary set I we denote by A^I the direct product of the family of partial structures, indexed by I, all the members of which are equal to A. This partial structure is called the I-th *direct power* of A.

It should be clear what we mean by the direct product $A_1 \times \cdots \times A_n$ of a finite collection of partial structures $A_1, \dots, A_n$.

An algebra is said to be *directly indecomposable* if it is not isomorphic to the direct product of any two nontrivial algebras.

5.2. THEOREM. *An equivalence on an algebra A is a congruence of A if and only if it is a subuniverse of the direct product $A \times A$.*

PROOF. It is easy. □

Let H be a family of algebras of signature σ over a set I. By a *subdirect product* of H we mean any subalgebra A of the direct product ΠH such that for any $i \in I$, the restriction of the i-th projection to A maps A onto H_i.

5.3. THEOREM. *Let A be a subdirect product of a family H of algebras over a set I. For every $i \in I$ denote by r_i the kernel of the restriction of the i-th projection to A, so that r_i is a congruence of A. Then $\bigcap\{r_i : i \in I\} = \mathbf{id}_A$.*

Conversely, let A be an arbitrary algebra and r be a family (over a set I) of congruences of A such that $\bigcap\{r_i : i \in I\} = \mathbf{id}_A$. Then A is isomorphic to a subdirect product of the family of algebras A/r_i ($i \in I$).

PROOF. The first statement is clear. Let A and r be as in the second statement. Define a family H over I by $H_i = A/r_i$. For $a \in A$, let $f(a)$ be the element of ΠH such that $f(a)(i) = a/r_i$ for all $i \in I$. One can easily verify that f is a homomorphism of A into the direct product ΠH, the range B of f is a subdirect product of H and the kernel of f is equal to $\bigcap\{r_i : i \in I\} = \mathbf{id}_A$, so that f is an isomorphism of A onto B. □

With respect to this correspondence between subdirect decompositions of a given algebra and families of congruences of the algebra with identical intersection, we introduce the following definition: An algebra A is said to be *subdirectly irreducible* if it is nontrivial and whenever $\mathbf{id}_A$ is the intersection of a nonempty family of congruences of A, then at least one of the congruences equals $\mathbf{id}_A$.

Clearly, every simple algebra is subdirectly irreducible.

The congruence lattice of a subdirectly irreducible algebra A contains precisely one atom; this atom is the intersection of all the congruences of A different from $\mathbf{id}_A$. This unique atom is called the *monolith* of A. The monolith of a subdirectly irreducible algebra A is contained in any congruence of A other than $\mathbf{id}_A$.

5.4. THEOREM. (Birkhoff [44]) *Every algebra is isomorphic to a subdirect product of a family of subdirectly irreducible algebras.*

PROOF. Let A be an algebra. Put $I = \{\langle a, b\rangle \in A^2 : a \neq b\}$. It follows easily from Zorn's lemma that for any $\langle a, b\rangle \in I$ there exists a maximal congruence among the congruences r of A not containing $\langle a, b\rangle$. For each $\langle a, b\rangle \in I$ choose one such maximal congruence and denote it by r_i. It follows from the maximal property of r_i that the algebra A/r_i is subdirectly irreducible. Clearly, the intersection of this family of congruences equals $\mathbf{id}_A$, and so, according to 5.3, A is isomorphic to a subdirect product of the family of algebras A/r_i. □

6. ISP-closed classes

Let K be a class of partial structures of the given language. We denote by $\mathbf{H}(K)$ the class of homomorphic images of elements of K, by $\mathbf{S}(K)$ the class of substructures of elements of K, by $\mathbf{P}(K)$ the class of direct products of arbitrary families of partial structures from K, and by $\mathbf{I}(K)$ the class of partial structures isomorphic to a partial structure from K. Observe that if K is a class of structures then all these classes are also classes of structures. A class K is said to be closed under homomorphic images (or substructures, or direct products, or isomorphisms) if $\mathbf{H}(K) \subseteq K$ (or $\mathbf{S}(K) \subseteq K$, or $\mathbf{P}(K) \subseteq K$, or $\mathbf{I}(K) \subseteq K$). K is said to be ISP-closed if it is closed under isomorphisms, substructures and direct products; it is said to be HSP-closed if it is closed under homomorphic images, substructures and direct products.

Observe that every ISP-closed class is nonempty. The class of one-element structures with all relations nonempty is the least ISP-closed class. Of course, the largest ISP-closed class is the class of all partial structures of the given language.

6.1. THEOREM. *Let K be a class of partial structures of the given language. Then*

$$\mathbf{SH}(K) \subseteq \mathbf{HS}(K),\ \mathbf{PH}(K) \subseteq \mathbf{HP}(K),\ \mathbf{PS}(K) \subseteq \mathbf{SP}(K),\ \mathbf{PP}(K) \subseteq \mathbf{IP}(K).$$

PROOF. It is easy. □

6.2. THEOREM. *Let K be a class of partial structures of the given signature. Then $\mathbf{ISP}(K)$ is the least ISP-closed class containing K and $\mathbf{HSP}(K)$ is the least HSP-closed class containing K.*

PROOF. It follows from 6.1. □

We call $\mathbf{ISP}(K)$ the ISP-closed class generated by K, and $\mathbf{HSP}(K)$ the HSP-closed class generated by K.

By a *reflection* of a partial structure A in a class K of partial structures we mean its reflection in the sense of category theory, i.e., a partial structure $B \in K$ together with a homomorphism f of A into B, such that for any $C \in K$ and any homomorphism g of A into C there exists precisely one homomorphism h of B into C with the property $g = hf$. Clearly, B is uniquely determined up to isomorphism by A and K. We often neglect the homomorphism f and by a reflection of A in K we mean just the partial structure B.

6.3. THEOREM. *Let K be an ISP-closed class of partial structures. Then every partial structure A of the given language has a reflection in K. If $f : A \to B$ is a reflection of A in K, then $f(A)$ is a generating subset of B.*

PROOF. Denote by Q the class of the ordered pairs $\langle g, C \rangle$ such that $C \in K$, g is a homomorphism of A into C and $g(A)$ is a generating subset of C. It follows from 3.4 that there exists a subset I of Q such that for every $\langle g, C \rangle \in Q$ there are a pair $\langle g', C' \rangle \in I$ and an isomorphism h of C onto C' with $g' = hg$. For $i = \langle g, C \rangle \in I$ put $H_i = C$, and denote by D the product of this family of partial structures; denote by p_i the i-th projection of D onto D_i. There exists a unique homomorphism $f : A \to D$ such that $g = p_i f$ for all $i = \langle g, C \rangle \in I$. Denote by B the substructure of D generated by the range of f. Since K is ISP-closed, B belongs to K. It is easy to check that $f : A \to B$ is a reflection of A in K. □

6.4. THEOREM. *Let A be an algebra and K be an ISP-closed class of algebras. Then there exists the least congruence r of A with the property $A/r \in K$. The algebra A/r, together with the canonical homomorphism of A onto A/r, is a reflection of A in K.*

PROOF. Define r as the intersection of all the congruences s such that $A/s \in K$. □

6.5. THEOREM. *Let S be a generating subset of a partial algebra A. The class of the partial algebras B such that every mapping of S into B can be extended to a homomorphism of A into B is a HSP-closed class.*

PROOF. It is easy to prove that the class is closed under subalgebras, homomorphic images and direct products. □

7. Free partial structures

A partial structure A is called *free* over a set X in a class K of partial structures (or also K-free over X), if $A \in K$, X is a generating subset of A and for any $B \in K$, any mapping of X into B can be extended to a homomorphism of A into B. (This extension is then unique due to 3.3).

A free partial structure over a set X in a given class K is uniquely determined up to isomorphism by the cardinality of X. If A is free over X in K and B is free over Y in K and if there is a bijection f of X onto Y, then f can be uniquely extended to an isomorphism of A onto B. (This follows from 2.2.)

So, for every class K and every cardinal number κ there exists at most one (up to isomorphism) free partial structure in K over a set of cardinality κ.

By a *discrete partial structure* we mean a partial structure A such that S_A is empty for any relation or operation symbol S. Clearly, a discrete partial structure is uniquely determined by its underlying set.

7.1. THEOREM. *A partial structure is free in the class of all partial structures if and only if it is discrete. The discrete partial structure with the underlying set X is free over X in the class of all partial structures of the given language.*

PROOF. It is easy. $\square$

7.2. THEOREM. *Let K and L be two nontrivial ISP-closed classes such that $L \subseteq K$; let A be a free partial structure over X in the class K, and let $f : A \to B$ be a reflection of A in L. Then the restriction of f to X is injective and B is free over $f(X)$ in L.*

PROOF. It is easy. $\square$

7.3. THEOREM. *Let K be a nontrivial ISP-closed class. Then for every nonempty set X there exists a free partial structure over X in K.*

PROOF. It follows from 6.3, 7.1 and 7.2. $\square$

8. The category of all partial structures of a given language

Every class K of partial structures (of a given language) can be considered as (and identified with) a category in the following way: the objects of the category are the elements of K; morphisms are triples $\langle f, A, B\rangle$ such that $A, B \in K$ and f is a homomorphism of A into B; A is the beginning and B is the end of $\langle f, A, B\rangle$; $\langle g, B, C\rangle\langle f, A, B\rangle = \langle gf, A, C\rangle$; $1_A = \langle \mathbf{id}_A, A, A\rangle$.

The category of all partial structures of a given language σ will be denoted by $\mathbf{Q}_\sigma$.

Let H be a family of partial structures over a set I. The direct product of H, together with the projections, is a product of H in the category $\mathbf{Q}_\sigma$. Under the assumption that I is nonempty, we are going to construct a coproduct of H in $\mathbf{Q}_\sigma$. Define a partial structure B as follows: its underlying set is the set of the ordered pairs $\langle i, a\rangle$ where $i \in I$ and $a \in H_i$; for an n-ary operation symbol F, $F_B(\langle i_1, a_1\rangle, \dots, \langle i_n, a_n\rangle)$ is defined if and only if $i_1 = \dots = i_n = i$ and $F_{H_i}(a_1, \dots, a_n)$ is defined for some $i \in I$, in which case the defined element is $\langle i, F_{H_i}(a_1, \dots, a_n)\rangle$; in particular, F_B is never defined for a constant F; for an n-ary relation symbol R, $\langle\langle i_1, a_1\rangle, \dots, \langle i_n, a_n\rangle\rangle \in R_B$ if and only if $i_1 = \dots = i_n = i$ and $\langle a_1, \dots, a_n\rangle \in R_{H_i}$ for some $i \in I$. Denote by r the least congruence of B containing all the pairs $\langle\langle i, F_{H_i}\rangle, \langle j, F_{H_j}\rangle\rangle$ such that F is a constant and F_{H_i} and F_{H_j} are both defined. Denote by A the partial structure B/r, modified by defining $F_A = \langle i, F_{H_i}\rangle/r$ for any constant F such that F_{H_i} is defined for at least one $i \in I$. For $i \in I$ define a mapping

μ_i of H_i into A by $\mu_i(a) = \langle i, a\rangle/r$. One can easily check that A together with μ is a coproduct of H in $\mathbf{Q}_\sigma$.

On the other hand, there is no coproduct of the empty family in $\mathbf{Q}_\sigma$, i.e., the category $\mathbf{Q}_\sigma$ has no initial object. The reason is just formal: we did not allow a partial structure to have the empty underlying set. In most situations it would be inconvenient to have to consider empty partial structures, but there are also situations, like here, where this lack causes a problem. For a category K define a new category K^{+0} obtained from K by adding a new object (we can call it the empty object and denote it by 0) in such a way that K is a full subcategory of K^{+0}, 0 is an initial object of K^{+0}, and there is no morphism of a K-object into 0.

8.1. THEOREM. *For a language σ, the category $\mathbf{Q}_\sigma^{+0}$ is both complete and cocomplete.*

PROOF. According to 2.3.1, it remains to prove that equalizers and coequalizers exist in $\mathbf{Q}_\sigma^{+0}$. Let A, B be two partial structures and $f : A \to B$ and $g : A \to B$ be two homomorphisms. Put $S = \{a \in A : f(a) = g(a)\}$. If S is nonempty, then there is a unique substructure C of A with the underlying set S, and C together with the identity is an equalizer of the pair f, g. If S is empty, then 0 together with the unique morphism of 0 into A is an equalizer of f, g.

Denote by r the congruence of B generated by the relation $\{\langle f(a), g(a)\rangle : a \in A\}$. It is easy to see that B/r together with the canonical homomorphism of B onto B/r is a coequalizer of the pair f, g. □

8.2. THEOREM. *A morphism $f : A \to B$ of $\mathbf{Q}_\sigma$ is a monomorphism if and only if f is injective. A morphism $f : A \to B$ of $\mathbf{Q}_\sigma$ is an epimorphism if and only if the range of f is a generating subset of B. Consequently, the category $\mathbf{Q}_\sigma$ is both locally and colocally small.*

PROOF. Both converse implications are easy. Let $f : A \to B$ be a monomorphism and suppose $f(a) = f(b)$ for two distinct elements a, b of A. Denote by C the discrete partial structure with the underlying set $\{a\}$ (no partial operations and no relations are defined in C), and define two homomorphisms g, h of C into A by $g(a) = a$ and $h(a) = b$. Then $fg = fh$, while $g \neq h$.

Now let $f : A \to B$ be an epimorphism and suppose that the range of f is not a generating subset of B, i.e., that there exists a proper substructure X of B containing the range of f. Put $Y = B \setminus X$. One can easily construct a partial structure C with the underlying set $X \cup (Y \times \{1\}) \cup (Y \times \{2\})$ and two distinct homomorphisms $g, h : B \to C$ such that $gf = hf$. □

9. ISP-closed classes as categories

By 6.3, every ISP-closed class of partial structures of a language σ is a reflective subcategory of the category $\mathbf{Q}_\sigma$.

9.1. THEOREM. *Let K be an ISP-closed class of partial structures of a language σ. The category K^{+0} is both complete and cocomplete.*

PROOF. It follows from 2.4.1 and 8.1. □

9.2. EXAMPLE. Consider the class K of partial algebras A of the signature consisting of a binary symbol $\cdot$ and two unary symbols α, β, satisfying the following conditions:

(1) the partial operations α and β are operations
(2) ab is defined in A if and only if $\alpha(a) = \beta(b)$
(3) if ab is defined, then $\alpha(ab) = \alpha(b)$ and $\beta(ab) = \beta(a)$
(4) $\alpha(\beta(a)) = \beta(a)$ for all $a \in A$
(5) $\beta(\alpha(a)) = \alpha(a)$
(6) $(ab)c = a(bc)$ whenever $\alpha(a) = \beta(b)$ and $\alpha(b) = \beta(c)$
(7) $a \cdot \alpha(a) = a$ for all $a \in A$
(8) $\beta(a) \cdot a = a$ for all $a \in A$

For every nonempty small category D we can define a partial algebra $A \in K$ with the underlying set D^m as follows: $\alpha(a)$ is the identical morphism of the beginning of a; $\beta(a)$ is the identical morphism of the end of a; ab in A is the same as ab in D. This mapping of the class of nonempty small categories onto the class K is almost a bijection. Since functors between small categories correspond precisely to homomorphisms between the corresponding partial algebras, and since K is (obviously) an ISP-closed class, it follows that the category of small categories and functors is both complete and cocomplete.

The proof of Theorem 9.1, based on 2.4.1, enables us to actually construct a limit of a diagram in a given ISP-closed class K. In the case of colimits, however, it is just existential. For the construction of a colimit in K, we need to construct a colimit in the category $\mathbf{Q}_\sigma$ (which may be easy; for example, we have already given the construction of a coproduct) and then to take a reflection in K, which may be a problem. So, the construction of a colimit is a particular problem for a particular ISP-closed class. In order to be able to construct colimits at least in the class of all structures of a given signature, we need to have a construction of a reflection of an arbitrary partial structure in this class. This can be done as follows.

Let A be a partial structure of a language σ. Define a chain $B_0 \subseteq B_1 \subseteq \dots$ in this way: $B_0 = A$; B_{i+1} is the union of B_i with the set of the finite sequences $(F, a_1, \dots, a_n)$ such that F is an n-ary operation symbol of σ, $a_1, \dots, a_n$ are elements of B_i and if $a_1, \dots, a_n \in B_0$ then the element $F_A(a_1, \dots, a_n)$ is not defined. Define a σ-structure B with the underlying set $\bigcup_{i=0}^{\infty} B_i$ in this way: for an n-ary operation symbol F put $F_B(a_1, \dots, a_n) = (F, a_1, \dots, a_n)$ unless $F_A(a_1, \dots, a_n)$ is defined, in which case put $F_B(a_1, \dots, a_n) = f_A(a_1, \dots, a_n)$; for a relation symbol R put $R_B = R_A$. One can easily check that this structure B, together with the identical homomorphism of A into B, is a reflection of A in the class of all σ-structures.

10. Terms

Let X be a given set (disjoint with $\mathbf{Dom}(\sigma)$). By a *term* (of the language σ) over X we mean a finite sequence of elements of $\mathbf{Dom}(\sigma) \cup X$ which can be obtained in a finite number of steps using the following two rules:

(1) every element of X is a term over X;
(2) if F is an n-ary operation symbol of σ and $t_1, \dots, t_n$ are terms over X, then the composition of the $n+1$ sequences $Ft_1 \dots t_n$ is a term over X.

If the set X is fixed, or clear from the context, then by a term we mean a term over X.

By the *length* of a term t we mean the length of the finite sequence t. The length of t will be denoted by $\lambda(t)$. So, $\lambda(x) = 1$ for $x \in X$, and $\lambda(Ft_1 \dots t_n) = 1 + \lambda(t_1) + \dots + \lambda(t_n)$. Clearly, $\lambda(t) \geq 1$ for all t; we have $\lambda(t) = 1$ if and only if either $t \in X$ or t is a constant.

10.1. Lemma. *If t is a term, then no proper beginning of the sequence t is a term.*

Proof. Suppose there are two terms t, u such that u is a proper beginning of t, and let t be the shortest term for which such a proper beginning u exists. Clearly, $t \notin X$ and $t = Ft_1 \dots t_n$ for some operation symbol F of arity $n \geq 1$ and some terms $t_1, \dots, t_n$. Also, $u = Fu_1 \dots u_n$ for some terms $u_1, \dots, u_n$. Since $t \neq u$, there exists an index i with $t_i \neq u_i$. Let i be the least index with $t_i \neq u_i$. Then either t_i is a proper beginning of u_i or u_i is a proper beginning of t_i. But both t_i and u_i are shorter than t, a contradiction by induction. $\square$

The following lemma says that every term can be read in only one way.

10.2. Lemma. *Let $Ft_1 \dots t_n = Gu_1 \dots u_m$, where F is an operation symbol of arity n, G is an operation symbol of arity m, and t_i and u_j are terms. Then $F = G$, $n = m$, and $t_1 = u_1, \dots, t_n = u_n$.*

Proof. Clearly, $F = G$ and hence $n = m$. Suppose there is an index i with $t_i \neq u_i$, and let i be the least index with this property. Then clearly either t_i is a proper beginning of u_i, or u_i is a proper beginning of t_i; we get a contradiction by Lemma 10.1. $\square$

Let X be arbitrary if σ contains constants, and nonempty if σ is without constants. Then the set T of terms over X is nonempty, and we can define a structure T with the underlying set T as follows: $R_T = 0$ for every relation symbol R; if F is an operation symbol of arity n then $F_T(t_1, \dots, t_n) = Ft_1 \dots t_n$. This structure is called *the structure of σ-terms* over X (the algebra of terms over X, if σ is a signature); it will be denoted by $\mathbf{T}_{X,\sigma}$ (or only $\mathbf{T}_X$).

Endomorphisms of the structure of terms are called its substitutions.

Let u, v be two terms. We write $u \leq v$ if there exists a substitution f such that $f(u)$ is a subterm of v. We write $u \sim v$ (and say that u, v are similar) if $u \leq v$ and $v \leq u$. We write $u < v$ if $u \leq v$ and $v \nleq u$.

For a term t we denote by $\mathbf{S}(t)$ the set of variables occurring in t.

11. Absolutely free algebras

An algebra is said to be *absolutely free* over X if it is free over X in the class of all algebras (of signature σ). An algebra is called absolutely free if it is absolutely free over some set X.

11.1. THEOREM. *Let X be nonempty if σ is without constants. Then the algebra $\mathbf{T}_X$ of terms over X is an absolutely free algebra over X.*

PROOF. Clearly, $\mathbf{T}_X$ is generated by X. Let A be an algebra and f be a mapping of X into A. For $t \in \mathbf{T}_X$, define $h(t) \in A$ by induction on $\lambda(t)$ as follows: $h(t) = f(t)$ for $t \in X$; $h(Ft_1 \dots t_n) = F_A(h(t_1), \dots, h(t_n))$. Clearly, h is a homomorphism of $\mathbf{T}_X$ into A extending f. $\square$

It follows that an algebra is absolutely free if and only if it is isomorphic to $\mathbf{T}_X$ for some X. Clearly, the set X is uniquely determined: it consists of the elements that cannot be expressed as $F_A(a_1, \dots, a_n)$ for any operation symbol F and any elements $a_1, \dots, a_n \in A$.

11.2. THEOREM. *An algebra A is absolutely free over a set X if and only if the following three conditions are satisfied:*

(1) *X is a set of generators of A;*
(2) *$F_A(a_1, \dots, a_n) \notin X$ for any $F, a_1, \dots, a_n$;*
(3) *$F_A(a_1, \dots, a_n) = G_A(b_1, \dots, b_m)$ implies $F = G$ and $a_1 = b_1, \dots, a_n = b_n$.*

PROOF. Clearly, the algebra of terms over X, and hence every absolutely free algebra over X, has the three properties. Let (1), (2) and (3) be satisfied for an algebra A. By 11.1, the identity on X can be extended to a homomorphism h of $\mathbf{T}_X$ into A. By (1), h is a homomorphism onto A. By (2) and (3), h is injective. $\square$

11.3. THEOREM. *An algebra A is absolutely free if and only if the following two conditions are satisfied:*

(1) *$F_A(a_1, \dots, a_n) = G_A(b_1, \dots, b_m)$ implies $F = G$ and $a_1 = b_1, \dots, a_n = b_n$;*
(2) *there is no infinite sequence $a_0, a_1, \dots$ of elements of A such that for any $i = 0, 1, \dots$, a_i can be expressed as $a_i = F_A(b_1, \dots, b_n)$ with $a_{i+1} \in \{b_1, \dots, b_n\}$.*

PROOF. The direct implication is clear: in the algebra of terms, there is no infinite sequence as in (2), since the term a_{i+1} would be shorter than a_i for any i. Conversely, let A be an algebra satisfying (1) and (2). Denote by X the set of all the elements of A that cannot be expressed as $F_A(a_1, \dots, a_n)$ for any F and any elements $a_1, \dots, a_n$. By 11.2, it is sufficient to show that A is generated by X. Suppose, on the contrary, that there exists an element $a \in A \setminus \mathbf{Sg}_A(X)$. Let us define an infinite sequence $a_0, a_1, \dots$ of elements of $A \setminus \mathbf{Sg}_A(X)$ as follows: $a_0 = a$; if $a_i \in A \setminus \mathbf{Sg}_A(X)$ has been already chosen, then $a_i = F_A(b_1, \dots, b_n)$ for some F and $b_1, \dots, b_n$; we cannot have

$b_j \in \mathbf{Sg}_A(X)$ for all j, so take one index j with $b_j \in A \smallsetminus \mathbf{Sg}_A(X)$ and put $a_{i+1} = b_j$. The infinite sequence $a_0, a_1, \dots$ contradicts (2). □

11.4. THEOREM. *A subalgebra of an absolutely free algebra is absolutely free. The direct product of a nonempty family of absolutely free algebras is absolutely free.*

PROOF. It follows from 11.3. □

12. Representation of lattices by subuniverses and congruences

Recall that an element a of a complete lattice L is said to be *compact* if for any subset S of L, $a \le \bigvee S$ implies $a \le \bigvee S'$ for some finite subset S' of S; by an *algebraic lattice* we mean a complete lattice L such that every element of L is the join of a set of compact elements of L.

12.1. THEOREM. *Let L be an algebraic lattice. The least element of L is compact. The join of any two compact elements of L is compact. Consequently, the set C of compact elements of L is a join-semilattice with a least element (with respect to the order relation of L restricted to C; this join-semilattice will be called the join-semilattice of compact elements of L).*

PROOF. It is easy. □

Let S be a join-semilattice with a least element o. By an ideal of S we mean a subset I of S such that $o \in I$, $x \le y \in I$ implies $x \in I$, and $x, y \in I$ implies $x \vee y \in I$. Clearly, the set of ideals of a join-semilattice with a least element is a complete lattice with respect to inclusion; it is called the lattice of ideals of S.

12.2. THEOREM. *Every algebraic lattice L is isomorphic to the lattice of ideals of some join-semilattice with a least element; namely, to the lattice of ideals of the join-semilattice of its compact elements.*

PROOF. Denote the join-semilattice of compact elements of L by C, and the lattice of ideals of C by K. For every $a \in L$ put $f(a) = \{x \in C : x \le a\}$, so that $f(a) \in K$. For every $I \in K$ denote by $g(I)$ the join of I in L. It is easy to check that f is an isomorphism of L onto K and g is the inverse isomorphism. □

12.3. THEOREM. *For every algebraic lattice L there exists a σ-algebra A (for some signature σ) such that L is isomorphic to the lattice of subuniverses of A. The signature can be chosen in such a way that it contains only some unary, one nullary and one binary operation symbols.*

PROOF. By 12.2, L is isomorphic to the lattice of ideals of a join-semilattice C with a least element o. Let σ be the signature containing one constant 0, one binary operation symbol $\vee$ and, for every pair a, b of elements of C such that $b < a$, a unary operation symbol $F_{a,b}$. Denote by A the σ-algebra with the underlying set C and operations defined in this way: $0_A = o$; $\vee_A = \vee_C$;

$F_{a,b}(a) = b$; $F_{a,b}(x) = x$ whenever $x \neq a$. Clearly, a subset of A is a subuniverse if and only if it is an ideal of C. □

12.4. THEOREM. *The following are equivalent for a lattice L:*

(1) *L is isomorphic to the lattice of subuniverses of some algebra of some countable signature*
(2) *L is isomorphic to the lattice of subuniverses of some groupoid*
(3) *L is isomorphic to the lattice of subuniverses of some commutative groupoid*
(4) *L is an algebraic lattice such that for every compact element a of L, the set of the compact elements in the principal ideal generated by a is countable*

PROOF. (4) implies (1): By 12.2, L is isomorphic to the lattice of ideals of a join-semilattice C with a least element o such that the principal ideal of every element of C is countable. For every $p \in C$ let $c^{(0)}, c^{(1)}, \dots$ be all elements of that principal ideal. Let σ be the signature containing one constant 0, one binary operation symbol $\vee$ and unary operation symbols $F_0, F_1, \dots$. Denote by A the σ-algebra with the underlying set C and operations defined in this way: $0_A = o$; $\vee_A = \vee_C$; $F_i(c) = c^{(i)}$. Clearly, a subset of A is a subuniverse if and only if it is an ideal of C.

(1) implies (2): It is sufficient to assume that L is isomorphic to the lattice of subuniverses of an algebra A of a countable signature σ without constants. Let $F_2, F_3, F_4, \dots$ be all the operation symbols of σ and $n_2, n_3, n_4, \dots$ be their arities. Denote by G the groupoid of terms over the set A. For every $t \in G$ and every positive integer n define an element $t^n \in G$ by $t^1 = t$ and $t^{n+1} = t^n t$. Define a groupoid H with the underlying set G and the basic binary operation $\circ$ in this way:

(1) if $a, b \in G$ and $a \neq b$ then $a^2 \circ b^2 = ab$
(2) if $a \in A$ then $a \circ a = a^2$
(3) if $a \in G \setminus A$ and $a_1, \dots, a_k$ are all the elements of A occurring in the term a and arranged into this finite sequence in the order of their first occurrences in A, then $a \circ a = a_1,\ a_1 \circ a = a_2,\ \dots,\ a_{k-1} \circ a = a_k,\ a_k \circ a = aa$
(4) if $a = ((((a_1 a_2) a_3) \dots) a_k)^m$ where $m \geq 2$, $k = n_m$ and $a_1, \dots, a_k \in A$ then $a \circ a_1 = F_m(a_1, \dots, a_k)$
(5) $a \circ b = a$ in all the remaining cases

For every subuniverse X of A denote by $z(X)$ the subuniverse of G generated by X. It is easy to see that z is an isomorphism of the lattice of subuniverses of A onto the lattice of subuniverses of H.

(2) implies (3): Let A be a groupoid, with the basic binary operation denoted by $g(x, y)$. Denote by G the groupoid of terms over the set A; for $t \in G$ and $n \geq 1$ define $t^n \in G$ as above. Let us take one fixed well ordering of G. Define a groupoid H with the underlying set G and the basic binary operation $\circ$ in this way:

(1) if $a \in A$ then $a \circ a = a^2$
(2) if $a \in G \setminus A$ and $a_1, \dots, a_n$ are all the elements of A occurring in the term a and ordered into this finite sequence according to the fixed well ordering of G, then $a \circ a = a_1,\ a_1 \circ a = a \circ a_1 = a_2,\ \dots,\ a_{n-1} \circ a = a \circ a_{n-1} = a_n,\ a_n \circ a = a \circ a_n = a^2$
(3) if $a \in G$ then $a^2 \circ (a^2)^2 = (a^2)^2 \circ a^2 = a^3$ and $a^2 \circ ((a^2)^2)^2 = ((a^2)^2)^2 \circ a^2 = a^4$
(4) if $a, b \in G$ then $a^2 \circ b^3 = b^3 \circ a^2 = ab$
(5) if $a, b \in A$ then $a^2 \circ b^4 = b^4 \circ a^2 = g(a, b)$
(6) in all the remaining cases let $a \circ b$ be the minimum of a, b with respect to the fixed well ordering

It is easy to see that H is a commutative groupoid and the mapping assigning to any subuniverse X of A the subuniverse of G generated by X is an isomorphism of the lattice of subuniverses of A onto the lattice of subuniverses of H.

(3) implies (4): This is clear. □

In the following we are going to prove a representation theorem for congruence lattices.

Let C be a join-semilattice with a least element o. By a C-graph we will mean an ordered pair $\langle X, h\rangle$ where X is a nonempty set and h is a mapping of a set of precisely 2-element subsets of X into C (write $h(x, y) = h(\{x, y\})$. By a stable mapping of a C-graph $\langle X, h\rangle$ into a C-graph $\langle X', h'\rangle$ we will mean a mapping f of X into X' such that whenever $h(a, b)$ is defined then either $f(a) = f(b)$ of $h'(f(a), f(b)) = h(a, b)$.

For every natural number n define a C-graph $\langle A_n, h_n\rangle$ in this way: $A_0 = \{1, 2\}$; $h_0(1, 2) = h_0(2, 1) = o$; A_{n+1} is the union of A_n with the set of all ordered quintuples $\langle a, b, p, q, i\rangle$ such that $a, b \in A_n,\ p, q \in C,\ i \in \{1, 2, 3\}$, $\{a, b\} \in \mathbf{Dom}(h_n)$ and $h_n(a, b) \le p \vee q$; let h_{n+1} be the extension of h_n by

$h_{n+1}(a, \langle a, b, p, q, 1\rangle) = p$,
$h_{n+1}(\langle a, b, p, q, 1\rangle, \langle a, b, p, q, 2\rangle) = q$,
$h_{n+1}(\langle a, b, p, q, 2\rangle, \langle a, b, p, q, 3\rangle) = p$,
$h_{n+1}(\langle a, b, p, q, 3\rangle, b) = q$.

Denote by A the union of the chain $A_0 \subseteq A_1 \subseteq \dots$ and by H the union of the chain $h_0 \subseteq h_1 \subseteq \dots$. Clearly, $\langle A, H\rangle$ is a C-graph.

Denote by S the set of stable mappings of $\langle A, H\rangle$ into itself. We can consider A as an algebra of a signature consisting of unary operation symbols only, such that the unary operations of A are precisely all the elements of S. We are going to show that the congruence lattice of A is isomorphic to the lattice of ideals of C.

Let $\{c, d\} \in \mathbf{Dom}(H)$ and let n be the least index with $c, d \in A_n$. We define a mapping $f_{c,d}$ of A_n into itself as follows. If $n = 0$, let $f_{c,d}$ be the identity on A_0. If $n > 0$ then, for some a, b, p, q, $\{c, d\}$ is one of the following four unordered pairs:

$\{a, \langle a, b, p, q, 1\rangle\}$,
$\{\langle a, b, p, q, 1\rangle, \langle a, b, p, q, 2\rangle\}$,

$\{\langle a,b,p,q,2\rangle,\langle a,b,p,q,3\rangle\}$,
$\{\langle a,b,p,q,3\rangle,b\}$.

In the first case put $f_{c,d}(\langle a,b,p,q,1\rangle)=f_{c,d}(\langle a,b,p,q,2\rangle)=\langle a,b,p,q,1\rangle$ and $f_{c,d}(y)=a$ for all the other elements $y\in A_n$. In the second case put $f_{c,d}(\langle a,b,p,q,2\rangle)=f_{c,d}(\langle a,b,p,q,3\rangle)=\langle a,b,p,q,2\rangle$ and $f_{c,d}(y)=\langle a,b,p,q,1\rangle$ for all the other elements $y\in A_n$. In the third case put $f_{c,d}(\langle a,b,p,q,1\rangle)=f_{c,d}(\langle a,b,p,q,2\rangle)=\langle a,b,p,q,2\rangle$ and $f_{c,d}(y)=\langle a,b,p,q,3\rangle$ for all the other elements $y\in A_n$. In the fourth case put $f_{c,d}(\langle a,b,p,q,2\rangle)=f_{c,d}(\langle a,b,p,q,3\rangle)=\langle a,b,p,q,3\rangle$ and $f_{c,d}(y)=b$ for all the other elements $y\in A_n$.

12.5. Lemma. *Let $\{c,d\}\in\mathbf{Dom}(H)$ and let n be the least index with $c,d\in A_n$. Then $f_{c,d}$ is a stable mapping of $\langle A_n,h_n\rangle$ into itself. It maps A_n onto the two-element set $\{c,d\}$. We have $f_{c,d}(c)=c$ and $f_{c,d}(d)=d$.*

Proof. It is easy. □

12.6. Lemma. *Let n,m be two natural numbers. Every stable mapping of $\langle A_n,h_n\rangle$ into $\langle A_m,h_m\rangle$ can be extended to a stable mapping of $\langle A,H\rangle$ into itself.*

Proof. Clearly, it is sufficient to prove that every stable mapping f of $\langle A_n,h_n\rangle$ into $\langle A_m,h_m\rangle$ can be extended to a stable mapping g of $\langle A_{n+1},h_{n+1}\rangle$ into $\langle A_{m+1},h_{m+1}\rangle$. For $x\in A_n$ put $g(x)=f(x)$. Let $x\in A_{n+1}\smallsetminus A_n$, so that $x=\langle a,b,p,q,i\rangle$ for some $a,b,p,q,1$. If $f(a)=f(b)$, put $g(x)=f(a)$. If $f(a)\neq f(b)$ then $h_n(a,b)=h_m(f(a),f(b))$ and thus $\langle f(a),f(b),p,q,i\rangle\in A_{m+1}$; put $g(x)=\langle f(a),f(b),p,q,i\rangle$. Clearly, g is a stable mapping. □

12.7. Lemma. *Let $H(a,b)=H(c,d)$. Then there exists a stable mapping f of $\langle A,H\rangle$ into itself such that $f(a)=c$ and $f(b)=d$.*

Proof. Let n be the least index such that $a,b\in A_n$ and let m be the least index such that $c,d\in A_m$. Denote by g the mapping with domain $\{a,b\}$, such that $g(a)=c$ and $g(b)=d$. By 12.5, $f_{a,b}$ is a stable mapping of $\langle A_n,h_n\rangle$ into itself, with the range $\{a,b\}$. Consequently, the composition $gf_{a,b}$ is a stable mapping of $\langle A_n,h_n\rangle$ into $\langle A_m,h_m\rangle$; it sends a to c and b to d. The rest follows from 12.6. □

12.8. Lemma. *The range of H is equal to C.*

Proof. Already the range of h_1 is equal to C. □

For any ideal I of C define a binary relation $F(I)$ on A as follows: $\langle a,b\rangle\in F(I)$ if and only if there exists a finite sequence $e_0,\dots,e_k$ such that $e_0=a$, $e_k=b$ and $H(e_{i-1},e_i)\in I$ for all $i=1,\dots,k$. Clearly, $F(I)$ is a congruence of A.

For every congruence E of A define a subset $G(E)$ of C as follows: $p\in G(E)$ if and only if $p=H(a,b)$ for some $\langle a,b\rangle\in E$ (such that $\{a,b\}\in\mathbf{Dom}(H)$).

12.9. Lemma. *Let E be a congruence of A. Then $G(E)$ is an ideal of C.*

Proof. Let $q, r \in G(E)$ and $p \le q \vee r$. We must show that $p \in G(E)$. We have $H(a_1, b_1) = q$ and $H(a_2, b_2) = r$ for some $\langle a_1, b_1 1\rangle \in E$ and $\langle a_2, b_2\rangle \in E$. By 12.8 we have $H(a, b) = p$ for some a, b. Put $e_0 = a$, $e_i = \langle a, b, q, r, i\rangle$ for $i = 1, 2, 3$ and $e_4 = b$. For every $i = 1, 2, 3, 4$ we have either $H(e_{i-1}, e_i) = q$ or $H(e_{i-1}, e_i) = r$, so that $\langle e_{i-1}, e_i\rangle \in E$ by 12.7, since E is a congruence. Hence $\langle a, b\rangle = \langle e_0, e_4\rangle \in E$ and thus $p \in G(E)$. $\square$

12.10. Lemma. *Let $e_0, \dots, e_k$ be a finite sequence of elements of A such that $e_0 = e_k$ and $\{e_{i-1}, e_i\} \in \mathbf{Dom}(H)$ for all $i = 1, \dots, k$. Then $H(e_0, e_1) \le H(e_1, e_2) \vee \dots \vee H(e_{k-1}, e_k)$.*

Proof. Suppose that $e_0, \dots, e_k$ is a sequence of minimal length for which the assertion fails. It is clear that k is not less than 3 and the elements $e_1, \dots, e_k$ are pairwise different. Let n be the least index such that the elements $e_0, \dots, e_k$ all belong to A_n. Clearly, $n \ne 0$. At least one of the elements $e_0, \dots, e_k$ does not belong to A_{n-1}; let us denote it by $\langle a, b, p, q, i\rangle$ and put $c_0 = a$, $c_1 = \langle a, b, p, q, 1\rangle$, $c_2 = \langle a, b, p, q, 2\rangle$, $c_3 = \langle a, b, p, q, 3\rangle$, $c_4 = b$. Clearly, either c_0, c_1, c_2, c_3, c_4 or c_4, c_3, c_2, c_1, c_0 is a connected part of $e_0, e_1, \dots, e_k, e_1, \dots, e_k$. If either e_0 or e_1 is one of the elements c_1, c_2, c_3 then $H(e_0, e_1)$ is either p or q; but each of p and q occurs twice among $H(c_0, c_1), H(c_1, c_2), H(c_2, c_3), H(c_3, c_4)$ and hence at least once among $H(e_1, e_2), H(e_2, e_3), \dots, H(e_{k-1}, e_k)$; hence the join of these $k - 1$ elements is above both p and q and hence above $H(e_0, e_1)$, a contradiction. It remains to consider the case when c_1, c_2, c_3 are all different from e_0, e_1. Then either c_0, c_1, c_2, c_3, c_4 or c_4, c_3, c_2, c_1, c_0 is a connected part of $e_1, \dots, e_k$. If we delete c_1, c_2, c_3 in the first case or c_3, c_2, c_1 in the second case from $e_0, e_1, \dots, e_k$, we get a shorter sequence again contradicting the assertion, which gives us a contradiction with the minimality of k. $\square$

12.11. Lemma. *Let I be an ideal of C. Then $G(F(I)) = I$.*

Proof. It follows from 12.8 that $I \subseteq G(F(I))$. Let $p \in G(F(I))$, so that $p = H(a, b)$ for some $\langle a, b\rangle \in F(I)$. There exists a finite sequence $e_0, \dots, e_k$ such that $e_0 = a$, $e_k = b$ and $H(e_{i-1}, e_i) \in I$ for all $i = 1, \dots, k$. It follows from 12.10 that $p \le H(e_0, e_1) \vee \dots \vee H(e_{k-1}, e_k)$, so that $p \in I$. $\square$

Let us say that a pair $\langle a, b\rangle \in A^2$ dominates over a pair $\langle c, d\rangle \in A^2$ if there exist a finite sequence $e_0, \dots, e_k$ of elements of A and a finite sequence $f_1, \dots, f_k$ of stable mappings of $\langle A, H\rangle$ into itself such that $e_0 = c$, $e_k = d$ and $f_i(a) = e_{i-1}$ and $f_i(b) = e_i$ for all $i = 1, \dots, k$. Since the composition of two stable mappings is stable, one can easily see that if $\langle a, b\rangle$ dominates over $\langle c, d\rangle$ and $\langle c, d\rangle$ dominates over $\langle e, f\rangle$ then $\langle a, b\rangle$ dominates over $\langle e, f\rangle$.

12.12. Lemma. *For every $c, d \in A$ there exists a finite sequence $e_0, \dots, e_k$ of elements of A such that $e_0 = c$, $e_k = d$ and for every $i = 1, \dots, k$, $\{e_{i-1}, e_i\} \in \mathbf{Dom}(H)$ and $\langle c, d\rangle$ dominates over $\langle e_{i-1}, e_i\rangle$.*

Proof. We are going to prove by induction on n that such a sequence exists whenever $c, d \in A_n$. If $\{c, d\} \in \mathbf{Dom}(H)$ then everything is clear, since the identity on A is stable. Let $\{c, d\} \notin \mathbf{Dom}(H)$. Clearly, $n \ne 0$. Let us first construct a finite sequence $c_0, \dots, c_r$ in this way: if $c \in A_{n-1}$, put $r = 0$, $c_0 = c$; let $c \notin A_{n-1}$, so that c is a quintuple $\langle a, b, p, q, j \rangle$; in the case $j = 1$ put $r = 1$, $c_0 = c$, $c_1 = a$; in the case $j = 3$ put $r = 1$, $c_0 = c$, $c_1 = b$; in the case $j = 2$ put $r = 2$, $c_0 = c$, $c_1 = \langle a, b, p, q, 1 \rangle$, $c_2 = a$. In each case, $c_0 = c$, $c_r \in A_{n-1}$, $\{c_{i-1}, c_i\} \in \mathbf{Dom}(H)$ and $\langle c, d \rangle$ dominates over $\langle c_{i-1}, c_i \rangle$ for all $i = 1, \dots, r$. In the case $c \in A_{n-1}$ it is clear; in the case $j = 1$ the mapping $f_{a,c}$ sends c to c and d to a (since $\{c, d\} \notin \mathbf{Dom}(H)$) and $f_{a,c}$ can be extended to a stable mapping of $\langle A, H \rangle$ into itself by 12.6; in the case $j = 3$ similarly $f_{b,c}$ sends c to c and d to b; in the case $j = 2$ the mapping f_{e_1,e_2} sends c to c and d to e_1 and the mapping f_{e_0,e_1} sends c to e_1 and d to a. Quite similarly we can construct a finite sequence $d_0, \dots, d_s$ such that $d_0 = d$, $d_s \in A_{n-1}$, $\{d_{i-1}, d_i\} \in \mathbf{Dom}(H)$ and such that $\langle c, d \rangle$ dominates over $\langle d_{i-1}, d_i \rangle$ for all $i = 1, \dots, s$. By the induction assumption applied to the elements c_r, d_s there exists a finite sequence $b_0, \dots, b_t$ such that $b_0 = c_r$, $b_t = d_s$, $\{b_{i-1}, b_i\} \in \mathbf{Dom}(H)$ and $\langle c_r, d_d \rangle$ dominates over $\langle b_{i-1}, b_i \rangle$ for all $i = 1, \dots, t$. Clearly $\langle c, d \rangle$ dominates over $\langle c_r, d_s \rangle$ and hence also over each $\langle b_{i-1}, b_i \rangle$. Now the sequence $c_0, \dots, c_r, b_1, \dots, b_t, d_{s-1}, \dots, d_0$ can be taken for $e_0, \dots, e_k$. □

12.13. LEMMA. *Let E be a congruence of A. Then $F(G(E)) = E$.*

PROOF. Let $\langle a, b \rangle \in F(G(E))$. There exists a finite sequence $e_0, \dots, e_k$ such that $e_0 = a$, $e_k = b$ and $H(e_{i-1}, e_i) \in G(E)$ for all $i = 1, \dots, k$. For every $i = 1, \dots, k$ there exists a pair $\langle c_i, d_i \rangle \in E$ such that $H(e_{i-1}, e_i) = H(c_i, d_i)$. By 12.7 there exists a stable mapping f_i of $\langle A, H \rangle$ into itself such that $f_i(c_i) = e_{i-1}$ and $f_i(d_i) = e_i$. Since E is a congruence, we get $\langle e_{i-1}, e_i \rangle \in E$ for all i, so that also $\langle a, b \rangle = \langle e_0, e_k \rangle \in E$.

In order to prove the converse, let $\langle a, b \rangle \in E$ and $a \neq b$. By 12.12 there exists a finite sequence $e_0, \dots, e_k$ such that $e_0 = a$, $e_k = b$, $\{e_{i-1}, e_i\} \in \mathbf{Dom}(H)$ and $\langle a, b \rangle$ dominates over $\langle e_{i-1}, e_i \rangle$ for all $i = 1, \dots, k$. Since E is a congruence and $\langle a, b \rangle \in E$, also $\langle e_{i-1}, e_i \rangle \in E$. Since, moreover, $\{e_{i-1}, e_i\}$ belongs to $\mathbf{Dom}(H)$, we have $\langle e_{i-1}, e_i \rangle \in F(G(E))$. But then $\langle a, b \rangle = \langle e_0, e_k \rangle \in F(G(E))$. □

12.14. THEOREM. *Every algebraic lattice is isomorphic to the congruence lattice of an algebra of a signature containing only unary operation symbols.*

PROOF. It follows from the above results. □

This result is due to Grätzer, Schmidt [63]; we have followed a more simple proof given by Pudlák [76].

12.15. LEMMA. *Let A be a nonempty set and n be a positive integer. Denote by G the groupoid with the underlying set A^n, with multiplication defined by $\langle a_1, \dots, a_n \rangle \langle b_1, \dots, b_n \rangle = \langle a_n, b_1, \dots, b_{n-1} \rangle$. The congruences of G are precisely all the relations r' that can be obtained from an equivalence r on A in this way: $\langle \langle a_1, \dots, a_n \rangle, \langle b_1, \dots, b_n \rangle \rangle \in r'$ if and only if $\langle a_i, b_i \rangle \in r$ for all $i = 1, \dots, n$.*

PROOF. It is easy. □

12.16. THEOREM. *For every algebra A of a finite signature σ there exists an algebra B with one binary and one unary operation such that the congruence lattice of B is isomorphic to the congruence lattice of A. One can require, moreover, the following:*

(1) *A is a subset of B*
(2) *whenever r is a congruence of A and R is the congruence of B corresponding to r under the isomorphism then $r = R \cap (A \times A)$*
(3) *if A is finite then B is finite*
(4) *if A is infinite then* $\mathbf{card}(A) = \mathbf{card}(B)$

PROOF. Let $f_1, \dots, f_k$ be all the basic operations of A and let $n_1, \dots, n_k$ be their arities. Denote by n the maximum of the numbers $k, n_1, \dots, n_k$. Let B be the algebra with one binary operation defined in the same way as in 12.15 and one unary operation g defined in this way: $g(\langle a_1, \dots, a_n \rangle) = \langle f_1(a_1, \dots, a_{n_1}), \dots, f_k(a_1, \dots, a_{n_k}), \dots, f_k(a_1, \dots, f_{n_k}) \rangle$. It is easy to check that the lattices $\mathbf{Con}(A)$ and $\mathbf{Con}(B)$ are isomorphic. $\square$

CHAPTER 4

LATTICES AND BOOLEAN ALGEBRAS

1. Modular and distributive lattices

A lattice L is said to be *modular* if $a \le c$ implies $(a \vee b) \wedge c = a \vee (b \wedge c)$ for all $a, b, c \in L$; this condition is equivalent to $a \wedge (b \vee (a \wedge c)) = (a \wedge b) \vee (a \wedge c)$ for all $a, b, c \in L$.

The lattice with five elements $0, a, b, c, 1$ and the only covering relations $0 < a < c < 1$ and $0 < b < 1$ will be denoted by $\mathbf{N}_5$. Clearly, $\mathbf{N}_5$ is non-modular.

1.1. Theorem. *A lattice is modular if and only if it does not contain a sublattice isomorphic to* $\mathbf{N}_5$*.*

Proof. The direct implication is clear. Now let A be a non-modular lattice. There exist elements $a, b, c \in L$ such that $a \le c$ and $(a \vee b) \wedge c \ne a \vee (b \wedge c)$. Clearly, $a < c$ and $a \vee (b \wedge c) < (a \vee b) \wedge c$. One can easily check that the elements $0' = b \wedge c$, $a' = a \vee (b \wedge c)$, $b' = b$, $c' = (a \vee b) \wedge c$ and $1' = a \vee b$ constitute a sublattice of A isomorphic to $\mathbf{N}_5$. $\square$

1.2. Corollary. *The dual of a modular lattice is a modular lattice.*

A lattice L is said to be *distributive* if it satisfies $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$ for all $a, b, c \in L$. Clearly, every distributive lattice is modular.

1.3. Theorem. *A lattice is distributive if and only if it satisfies* $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$ *for all* $a, b, c \in L$*.*

Proof. Direct implication: $a \vee (b \wedge c) = a \vee ((c \wedge a) \vee (c \wedge b)) = a \vee (c \wedge (a \vee b)) = ((a \vee b) \wedge a) \vee ((a \vee b) \wedge c) = (a \vee b) \wedge (a \vee c)$. The converse implication can be proved similarly. $\square$

1.4. Corollary. *The dual of a distributive lattice is a distributive lattice.*

The lattice with five elements $0, a, b, c, 1$ and the only covering relations $0 < a < 1$, $0 < b < 1$, $0 < c < 1$ will be denoted by $\mathbf{M}_5$. Clearly, $\mathbf{M}_5$ is not distributive.

1.5. Theorem. *A lattice is distributive if and only if it contains no sublattice isomorphic to either* $\mathbf{N}_5$ *or* $\mathbf{M}_5$*.*

Proof. The direct implication is clear. For the converse, by 1.1 it is sufficient to prove that if A is a modular but not distributive lattice then A contains a sublattice isomorphic to $\mathbf{M}_5$. There are elements $a, b, c \in L$ such that $(a \wedge b) \vee (a \wedge c) < a \wedge (b \vee c)$. It is easy to check that the elements $0' = (a \wedge b) \vee (a \wedge c) \vee (b \wedge c)$, $1' = (a \vee b) \wedge (a \vee c) \wedge (b \vee c)$, $a' = (a \wedge 1') \vee 0'$, $b' = (b \wedge 1') \vee 0'$ and $c' = (c \wedge 1') \vee 0'$ constitute a sublattice isomorphic to $\mathbf{M}_5$. □

By a *prime filter* of a lattice A we mean a filter U such that whenever $x, y \in A$ and $x \vee y \in U$ then either $x \in U$ or $y \in U$. Prime ideals are defined dually.

1.6. Theorem. *Let a, b be two elements of a distributive lattice A such that $a \not\leq b$. Then there exists a prime filter U of A such that $a \in U$ and $b \notin U$.*

Proof. It follows easily from Zorn's lemma that there exists a filter U maximal among those filters that contain a and do not contain b. Suppose that there are elements $x, y \notin U$ with $x \vee y \in U$. By the maximality of U, the filter generated by $U \cup \{x\}$ contains b, i.e., there exists an element $u \in U$ with $b \geq x \wedge u$. Similarly, there exists an element $v \in U$ with $b \geq y \wedge v$. Then $b \geq (x \wedge u) \vee (y \wedge v) = (x \vee y) \wedge (u \vee y) \wedge (x \vee v) \wedge (u \vee v) \in U$, a contradiction. □

1.7. Theorem. *A lattice is distributive if and only if it is isomorphic to a sublattice of the lattice of all subsets of some set X.*

Proof. Of course, the lattice of all subsets of X is distributive and a sublattice of a distributive lattice is also distributive. Conversely, let A be a distributive lattice. Denote by X the set of prime filters of A and define a mapping f of A into the lattice of all subsets of X by $f(a) = \{U \in X : a \in U\}$. One can easily check that f is a homomorphism; by 1.6, f is injective. □

1.8. Theorem. *The two-element lattice is (up to isomorphism) the only subdirectly irreducible distributive lattice.*

Proof. It follows from 1.7, since the lattice of all subsets of X is isomorphic to a direct power of the two-element lattice. □

By a *maximal filter* of a lattice A we mean a filter U that is maximal among the filters different from A. Maximal ideals are defined dually.

The *least element* of a lattice L is an element o such that $o \leq a$ for all $a \in L$. The *greatest element* is defined dually. Let L be a lattice with the least element o and the greatest element i. An element $b \in A$ is said to be a *complement* of an element $a \in L$ if $a \wedge b = o$ and $a \vee b = i$. By a *complemented lattice* we mean a lattice L with the least and the greatest elements, in which every element has at least one complement; if, moreover, every element of L has precisely one complement, we say that L is a *uniquely complemented lattice*. By a *relatively complemented lattice* we mean a lattice, every interval of which is complemented.

It follows from 1.5 that if A is a distributive lattice with the least and the greatest elements then every element of A has at most one complement.

1.9. Theorem. *Let A be a relatively complemented distributive lattice, F be a nonempty filter of A and a be an element of $A \setminus F$. Then there exists a maximal filter U of A such that $F \subseteq U$, $A \setminus U$ is a maximal ideal and $a \notin U$.*

PROOF. Let U be a filter maximal among those filters that contain F and do not contain a (its existence follows from Zorn's lemma). If $x, y \in A \smallsetminus U$ and $x \vee y \in U$ then it follows from the maximality of U that there are elements $u, v \in U$ with $a \geq x \wedge u$ and $a \geq y \wedge v$, so that $a \geq (x \wedge u) \vee (y \wedge v) = (x \vee y) \wedge (u \vee y) \wedge (x \vee v) \wedge (u \vee v) \in U$, a contradiction. Thus $A \smallsetminus U$ is an ideal. It remains to prove that U is a maximal filter (the maximality of the ideal $A \smallsetminus U$ will follow by duality). Let $b \in A \smallsetminus U$ and $c \in A$. There exists an element $d \in U$ with $d \geq b$. Denote by e the complement of b in the interval $[b \wedge c,\ d]$. Since $e \vee b = d \in U$ and $b \notin U$, we have $e \in U$. Then $c \geq b \wedge e$, so that c belongs to the filter generated by $U \cup \{b\}$; but c was an arbitrary element of A and thus the filter generated by $U \cup \{b\}$ (for an arbitrary $b \in A \smallsetminus U$) equals A. □

2. Boolean algebras

By a *Boolean algebra* we mean an algebra A of the signature $\{\wedge, \vee, ', 0, 1\}$ such that the reduct of A to $\{\wedge, \vee\}$ is a distributive lattice with the least element 0 and the greatest element 1 and such that for every $a \in A$, a' is the complement of a. One can easily prove that every Boolean algebra is uniquely determined by its underlying lattice (its reduct to $\{\wedge, \vee\}$) and that such reducts are precisely the complemented distributive lattices; these lattices are uniquely complemented. The class of Boolean algebras is equivalent with the class of complemented distributive lattices. Complemented distributive lattice are called *Boolean lattices*.

One can easily see that $(a \vee b)' = a' \wedge b'$ and $(a \wedge b)' = a' \vee b'$ for any elements a, b of a Boolean algebra. (These are called DeMorgan's Laws.) Clearly, every Boolean algebra is a relatively complemented lattice.

For every set X we define a Boolean algebra A, called the Boolean algebra of subsets of X, in this way: A is the set of all subsets of X; $0_A = 0$, $1_A = X$, $Y_1 \wedge Y_2 = Y_1 \cap Y_2$, $Y_1 \vee Y_2 = Y_1 \cup Y_2$ and $Y' = X \smallsetminus Y$ for $Y, Y_1, Y_2 \subseteq X$. For $X = 1$, this algebra is called the two-element Boolean algebra. It is easy to see that the Boolean algebra of subsets of X is isomorphic to the direct power B^X, where B is the two-element Boolean algebra.

For a Boolean algebra A and an element $a \in A$ we define an algebra $A \upharpoonright a$ of the same signature as follows: its underlying set is the interval $[0, a]$ of A; the operations corresponding to $\wedge, \vee, 0$ are defined in the same way as in A; the constant 1 is interpreted by a; and the unary operation is the operation $x \mapsto a \wedge x'$.

2.1. THEOREM. *Let A be a Boolean algebra and $a \in A$. Then $A \upharpoonright a$ is a Boolean algebra. The mapping $x \mapsto a \wedge x$ is a homomorphism of A onto $A \upharpoonright a$. The mapping $x \mapsto \langle a \wedge x, a' \wedge x \rangle$ is an isomorphism of A onto $(A \upharpoonright a) \times (A \upharpoonright a')$.*

PROOF. It is easy. □

2.2. THEOREM. *The two-element Boolean algebra is, up to isomorphism, the only nontrivial directly indecomposable Boolean algebra.*

Proof. It follows from 2.1. □

2.3. Theorem. *For every nonnegative integer n there exists, up to isomorphism, precisely one Boolean algebra of cardinality 2^n, namely, the Boolean algebra of subsets of an n-element set; and there are no other finite Boolean algebras.*

Proof. It follows from 2.2, since every finite algebra is isomorphic to a direct product of directly indecomposable algebras. □

2.4. Theorem. *Every Boolean algebra is isomorphic to a subdirect power of the two-element Boolean algebra. Consequently, every Boolean algebra is isomorphic to a subalgebra of the Boolean algebra of subsets of some set.*

Proof. It follows from 2.2. □

2.5. Theorem. *The congruence lattice of a Boolean algebra A is isomorphic to the lattice of ideals of A which is isomorphic to the lattice of filters of A. For a congruence r, the corresponding ideal is the set $\{x \in A : \langle 0, x\rangle \in r\}$. For an ideal I, the corresponding congruence r is defined by $\langle x, y\rangle \in r$ if and only if $(x \wedge y') \vee (x' \wedge y) \in I$. The filter corresponding to I is the set $\{x' : x \in I\}$.*

Proof. It is easy. □

Let X be a nonempty subset of a Boolean algebra A. The ideal generated by X (the intersection of all ideals containing X) can be described as the set of all the elements a for which there exist some elements $x_1, \dots, x_n \in X$ (for some $n \geq 1$) with $a \leq x_1 \vee \dots \vee x_n$. Similarly, the filter generated by X is the set $\{a \in A : a \geq x_1 \wedge \dots \wedge x_n \text{ for some } x_1, \dots, x_n \in X\}$.

By an *ultrafilter* of a Boolean algebra A we mean a maximal filter of A. By 1.9, if F is a filter of a Boolean algebra A and $a \in A$ is an element not belonging to F then there exists an ultrafilter U of A such that $F \subseteq U$ and $a \notin F$.

2.6. Theorem. *A filter F of a Boolean algebra A is an ultrafilter if and only if for every $a \in A$, precisely one of the elements a and a' belongs to F.*

Proof. Let F be an ultrafilter. Let $a \in A$. If a, a' both belong to F then $0 = a \wedge a' \in F$ and hence $F = A$, a contradiction. Suppose that neither a nor a' belongs to F. By the maximality of F, the filter generated by $F \cup \{a\}$ equals A, which means that $0 = f_1 \wedge a$ for some $f_1 \in F$. Similarly, the filter generated by $F \cup \{a'\}$ equals A and $0 = f_2 \wedge a'$ for some $f_2 \in F$. Put $f = f_1 \wedge f_2$, so that $f \in F$. We have $0 = (f \wedge a) \vee (f \wedge a') = f \wedge (a \vee a') = f$, a contradiction. The converse implication is clear. □

Clearly, a principal filter of A is an ultrafilter if and only if it is generated by an atom of A, i.e., by an element that is minimal among the elements of $A \setminus \{0\}$. Every filter of a finite Boolean algebra is principal.

3. Boolean rings

By a *Boolean ring* we mean a ring R satisfying $aa = a$ for all $a \in R$.

3.1. THEOREM. *Every Boolean ring R is commutative and satisfies $a+a = 0$ for all $a \in R$. The class of Boolean algebras is equivalent with the class of Boolean rings: given a Boolean algebra, the corresponding Boolean ring is defined by*

$$a + b = (a \wedge b') \vee (a' \wedge b),$$
$$ab = a \wedge b,$$
$$-a = a;$$

and given a Boolean ring, the corresponding Boolean algebra is defined by

$$a \wedge b = ab,$$
$$a \vee b = a + b + ab,$$
$$a' = 1 + a.$$

PROOF. Let R be a Boolean ring. For $a \in R$ we have $a + a = (a + a)^2 = aa + aa + aa + aa = a + a + a + a$, from which we get $a + a = 0$. For $a, b \in R$ we have $a + b = (a + b)^2 = aa + ab + ba + bb = a + ab + ba + bb$, so that $ab + ba = 0$ and hence $ab = -ba = ba$. One can easily check the rest. □

4. Boolean spaces

By a *topology* on a set A we mean a set T of subsets of A satisfying the following three conditions:

(1) $0 \in T$ and $A \in T$
(2) if $X, Y \in T$ then $X \cap Y \in T$
(3) for every subset S of T, $\bigcup(S) \in T$

By a *topological space* we mean an ordered pair $\langle A, T\rangle$ such that T is a topology on A. When there is no confusion, we often forget to mention T and say that A is a topological space.

Let $A = \langle A, T\rangle$ be a topological space. The elements of T are called *open subsets* of A. By a *closed subset* of A we mean a subset $X \subseteq A$ such that $A \setminus X$ is open. Thus 0 and A are both open and closed; the intersection of finitely many open subsets is open; the union of any set of open subsets is open; the union of finitely many closed subsets is closed; the intersection of any nonempty set of closed subsets is closed. By a *clopen subset* of A we mean a subset that is both open and closed. Clearly, the set of clopen subsets of any topological space is a Boolean algebra with respect to inclusion.

Let $\langle A, T\rangle$ be a topological space. For a set $X \subseteq A$, the intersection of all closed subsets containing X is called the *closure* of X and is denoted by $\bar{X}$; it is the smallest closed subset containing X. The union of all open subsets contained in X is called the *interior* of X; it is the largest open subset contained in X. A subset X of A is called *dense* if its closure is A. A subset X of A is called *nowhere-dense* if the interior of the closure of X is empty.

Let $A = \langle A, T \rangle$ and $B = \langle B, S \rangle$ be two topological spaces. By a *continuous mapping* of $\langle A, T \rangle$ into $\langle B, S \rangle$ we mean a mapping f of A into B such that for any open subset X of B, the set f^{-1}''X is open in A. By a *homeomorphism* of $\langle A, T \rangle$ onto $\langle B, S \rangle$ we mean a bijection f of A onto B such that f is a continuous mapping of A into B and f^{-1} is a continuous mapping of B into A.

A topological space A is said to be a *Hausdorff space* if for any $a, b \in A$ with $a \neq b$ there exist two open subsets X, Y such that $a \in X$, $b \in Y$ and $X \cap Y$ is empty.

A topological space A is said to be *compact* if for every nonempty set S of closed subsets of A such that the intersection of S is empty there exists a finite nonempty subset U of S such that the intersection of U is empty. Equivalently, A is compact if and only if for every set S of open subsets of A with $\bigcup(S) = A$ there exists a finite subset U of S with $\bigcup(U) = A$.

4.1. LEMMA. *Let A be a compact Hausdorff space.*

(1) *For every open subset X of A and every element $a \in X$ there exists an open subset Y such that $a \in Y \subseteq X$ and the closure of Y is contained in X.*
(2) *If X is a union of countably many nowhere-dense subsets of A then the interior of X is empty..*

PROOF. (1) For every $b \in A \smallsetminus X$ choose two open subsets M_b and N_b such that $a \in M_b$, $b \in N_b$ and $M_b \cap N_b = 0$. The set $\{A \smallsetminus X\} \cup \{A \smallsetminus N_b : b \in A \smallsetminus X\}$ is a set of closed subsets with empty intersection. It follows from the compactness of A that there exist finitely many elements $b_1, \dots, b_n \in A \smallsetminus X$ with $(A \smallsetminus X) \cap (A \smallsetminus N_{b_1}) \cap \dots \cap (A \smallsetminus N_{b_n}) = 0$, i.e., $A \smallsetminus X \subseteq N_{b_1} \cup \dots \cup N_{b_n}$. The set $Y = M_{b_1} \cap \dots \cap M_{b_n}$ is open and contains a; it is contained in the closed subset $A \smallsetminus (N_{b_1} \cup \dots \cup N_{b_n})$ of X.

(2) Let $X = X_1 \cup X_2 \cup \dots$ where each X_i is nowhere-dense. Suppose that there exists a nonempty open subset $Y_0 \subseteq X$. Since X_1 is nowhere-dense, there exists an element $a_1 \in Y_0 \smallsetminus \bar{X}_1$; by (1) there exists an open subset Y_1 such that $a_1 \in Y_1$ and $\bar{Y}_1 \subseteq Y_0 \smallsetminus \bar{X}_1$. Since X_2 is nowhere-dense, there exists an element $a_2 \in Y_1 \smallsetminus \bar{X}_2$; by (1) there exists an open subset Y_2 such that $a_2 \in Y_2$ and $\bar{Y}_2 \subseteq Y_1 \smallsetminus \bar{X}_2$. If we continue in this way, we find an infinite sequence $Y_0, Y_1, Y_2, \dots$ of nonempty open subsets such that $\bar{Y}_n \subseteq \bar{Y}_{n-1} \smallsetminus \bar{X}_n$. In particular, $\bar{Y}_n \subseteq \bar{Y}_{n-1}$. By the compactness of A, the intersection of this chain is nonempty. Take an element a of this intersection. Clearly, a does not belong to any X_n and hence $a \notin X$; but $a \in \bar{Y}_1 \subseteq Y_0 \subseteq X$, a contradiction. $\square$

By a *Boolean space* we mean a compact Hausdorff space B such that every open subset of B is a union of a set of clopen subsets of B.

Let A be a Boolean algebra. We denote by A^* the set of ultrafilters of A. For $a \in A$ put $N_a = \{U \in A^* : a \in U\}$. Clearly, $N_a \cup N_b = N_{a \wedge b}$, $N_a \cap N_b = N_{a \vee b}$ and $A^* \smallsetminus N_a = N_{a'}$. It follows that A^* is a topological space with respect to the topology defined in this way: a subset of A^* is open if and only if it is a union $\bigcup\{N_a : a \in S\}$ for some subset S of A. We consider A^* as a topological space with respect to this topology.

For a topological space B denote by B^* the Boolean algebra of clopen subsets of B.

4.2. THEOREM.

(1) *Let A be a Boolean algebra. Then A^* is a Boolean space; the sets N_a, with $a \in A$, are precisely all the clopen subsets of A^*; the mapping $a \mapsto N_a$ is an isomorphism of A onto A^{**}.*
(2) *Let B be a Boolean space. Then B^* is a Boolean algebra and the mapping $x \mapsto \{X \in B^* : x \in X\}$ is a homeomorphism of B onto B^{**}.*

PROOF. It is easy. □

For a Boolean algebra A, the space A^* is called the Boolean space of A. The correspondence between Boolean algebras and Boolean spaces described in 4.2 is called the Stone duality.

4.3. THEOREM. *Let A_1, A_2 be Boolean algebras and B_1, B_2 be Boolean spaces.*

(1) *If f is a homomorphism of A_1 into A_2 then the mapping $f^* : A_2^* \to A_1^*$ defined by $f^*(U) = f^{-1}(U)$ is a a continuous mapping; if f is injective then f^* is surjective and if f is surjective then f^* is injective.*
(2) *If f is a continuous mapping of B_1 into B_2 then the mapping $f^* : B_2^* \to B_1^*$ defined by $f^*(x) = f^{-1}(x)$ is a homomorphism; if f is injective then f^* is surjective and if f is surjective then f^* is injective.*

PROOF. It is easy. □

Under the Stone duality, elements of a Boolean algebra correspond to clopen subsets; ideals correspond to open subsets and filters to closed subsets; the direct product of two Boolean algebras A_1, A_2 corresponds to the disjoint union of the Boolean spaces A_1^*, A_2^* (where open sets are unions of an open subset of A_1^* with an open subset of A_2^*); the free product of two Boolean algebras corresponds to the product of their Boolean spaces, with the product topology defined in the obvious way.

4.4. THEOREM. *Let A be a Boolean algebra. Let $a \in A$, $a \neq 0_A$ and for every positive integer n let E_n be a subset of A having the join a_n in A. Then there exists an ultrafilter U of A such that $a \in U$ and for all positive integers n, if $a_n \in U$ then $U \cap E_n$ is nonempty.*

PROOF. For every n denote by Y_n the set of all $U \in A^*$ such that $a_n \in U$ and U is disjoint with E_n. Suppose that Y_n has nonempty interior. Then there exists an element $b \neq 0_A$ of A such that $N_b \subseteq Y_n$. Then $b \leq a_n$ and $b \leq x'$ for all $x \in E_n$, so that $x \leq b'$ for all $x \in E_n$ and thus $a_n \leq b'$; from $b \leq a_n \leq b'$ and $b \neq 0_A$ we get a contradiction. This shows that for every n, Y_n is nowhere-dense. By 4.1, the union $Y = Y_1 \cup Y_2 \cup \dots$ has empty interior, so that N_a is not contained in Y. But this means the existence of an ultrafilter with the desired property. $\square$

5. Boolean products

For two mappings f, g with the same domain I, the set $\{i \in I : f(i) = g(i)\}$ will be denoted by $\mathbf{e}(f = g)$.

Let A_x ($x \in X$, X nonempty) be a family of σ-algebras. By a *Boolean product* of this family we mean a subdirect product A such that there is a Boolean space topology on X with the following two properties:

(1) for any $a, b \in A$, the set $\mathbf{e}(a = b)$ is clopen
(2) if $a, b \in A$ and Y is a clopen subset of X then $(a \restriction Y) \cup (b \restriction (X \smallsetminus Y)) \in A$ (the patchwork property)

This notion was introduced and studied in Foster [53] and [53a].

5.1. Theorem. *Let A be a nontrivial algebra such that the set L of finitely generated congruences of A is a distributive and relatively complemented sublattice of $\mathbf{Con}(A)$ and $r \circ s = s \circ r$ for all $r, s \in L$. Denote by M_0 the set of maximal congruences of A and put $M = M_0 \cup \{A^2\}$. Then A is isomorphic to a Boolean product of the algebras A/r with $r \in M$. Consequently, A is isomorphic to a subdirect product of its simple factors.*

Proof. Denote by X_0 the set of all maximal ideals of L and put $X = X_0 \cup \{L\}$. Observe that the mapping $I \mapsto \bigcup(I)$ is a bijection of X onto M. For $r \in L$ put $C_r = \{I \in X : r \in I\}$ and $D_r = X \smallsetminus C_r = \{I \in X : r \notin I\}$. One can easily check that for $r, s \in L$ we have

$$C_r \cup C_s = C_{r \cap s},\ C_r \cap C_s = C_{r \vee r},\ D_r \cup D_s = D_{r \vee s},\ D_r \cap D_s = D_{r \wedge s},$$
$$C_r \cup D_s = C_{r-s},\ C_r \cap D_s = D_{s-r}$$

where $r - s$ denotes the complement of s in the interval $[\mathbf{id}_A, r \vee s]$. It follows that the set of arbitrary unions of subsets of $\{C_r : r \in L\} \cup \{D_r : r \in L\}$ is a topology on X. Clearly, the sets C_r and D_r ($r \in L$) are all clopen. If $I, J \in X$ and $I \neq J$ then there exists an r such that either $r \in I \smallsetminus J$ or $r \in J \smallsetminus I$; in the first case $I \in C_r$ and $J \in D_r$, while in the second case $I \in D_r$ and $J \in C_r$; since C_r, D_r are disjoint, we see that X is a Hausdorff space.

In order to prove that X is a Boolean space, it remains to show that X is compact. Let $X = \bigcup_{r \in K_1} C_r \cup \bigcup_{s \in K_2} D_s$ where K_1, K_2 are two subsets of L. Since $L \in X$, K_1 is nonempty. Take one fixed congruence $t \in K_1$. For $r \in K_1$ define $r' \in L$ by $C_r \cap D_t = D_{r'}$ (i.e., put $r' = t - r$) and for $s \in K_2$ define $s' \in L$ by $D_s \cap D_t = D_{s'}$ (i.e., put $s' = s \wedge t$). We have $D_t = X \cap D_t = \bigcup_{r \in K_1} D_{r'} \cup \bigcup_{s \in K_2} D_{s'}$. The ideal generated by all r' with $r \in K_1$ and all s' with $s \in K_2$ contains t, since otherwise (by 1.9) it would be contained in a maximal ideal I such that $t \notin I$, and we would have $I \in D_t \smallsetminus (\bigcup_{r \in K_1} D_{r'} \cup \bigcup_{s \in K_2} D_{s'})$ which is impossible. Hence $t \subseteq r'_1 \vee \cdots \vee r'_k \vee s'_1 \vee \cdots \vee s'_m$ for some $r_i \in K_1$ and $s_j \in K_2$. We get $D_t \subseteq D_{r'_1} \cup \cdots \cup D_{r'_k} \cup D_{s'_1} \cup \cdots \cup D_{s'_m}$ and $X = C_t \cup D_t = C_t \cup D_{r'_1} \cup \cdots \cup D_{r'_k} \cup D_{s'_1} \cup \cdots \cup D_{s'_m} \subseteq C_t \cup C_{r_1} \cup \cdots \cup C_{r_k} \cup D_{s_1} \cup \cdots \cup D_{s'_m}$.

It follows that the sets C_r and D_r with $r \in L$ are the only clopen subsets of X.

Define a mapping f of A into the direct product $\prod_{I\in X} A/(\bigcup I)$ by $f(a) = a/(\bigcup I)$. Clearly, f is a homomorphism. If a, b are two distinct elements of A then the filter $\{r \in L : \langle a, b\rangle \in r\}$ is contained in a maximal filter F and $I = L \smallsetminus F$ is a maximal ideal; we have $\langle a, b\rangle \notin \bigcup I$. Hence the intersection of all congruences $\bigcup I$, with $I \in X$, is the identity and f is an isomorphism of A onto a subdirect product of $A/(\bigcup I)$ $(I \in X)$.

For $a, b \in A$ we have $\mathbf{e}(f(a) = f(b)) = C_{\mathbf{Cg}(a,b)}$, a clopen set. It remains to prove the patchwork property. Let $a, b \in A$ and $r \in L$; we need to show that $(f(a) \restriction C_r) \cup (f(b) \restriction D_r) \in f(A)$. Put $s = \mathbf{Cg}(a, b)$. We have $r \vee s = r \vee (s - r) = r \circ (s - r) = (s - r) \circ r$, so that there is an element $c \in A$ with $\langle a, c\rangle \in r$ and $\langle c, b\rangle \in s - r$. We have $\mathbf{e}(f(a) = f(c)) \supseteq C_r$ and $\mathbf{e}(f(c) = f(b)) \supseteq C_{s-r} = C_s \cup D_r \supseteq D_r$, so that $(f(a) \restriction C_r) \cup (f(b) \restriction D_r) = f(c)$. □

Let A be an algebra and B be a Boolean algebra. Denote by $A[B]^*$ the set of all continuous mappings of the Boolean space B^* into the discrete topological space A (discrete means that all subsets are open). Equivalently, $f \in A[B]^*$ if and only if f is a mapping of B^* into A and $f^{-1}(a)$ is open for any $a \in A$; actually, the set $f^{-1}(a)$ is clopen. It follows from the compactness of B^* that the range of any function $f \in A[B]^*$ is finite. Now it is easy to see that $A[B]^*$ is a subuniverse of the direct power A^{B^*}; it is a subdirect power. The subalgebra $A[B]^*$ of A^{B^*} is called the *Boolean power* of A by the Boolean algebra B.

5.2. THEOREM. *Let A be an algebra and B be a Boolean algebra. For a subset S of A^{B^*} we have $S = A[B]^*$ if and only if the following three conditions are satisfied:*

(1) *all constant maps of B^* into A are in S*
(2) *for $f, g \in S$, the set $\mathbf{e}(f = g)$ is clopen*
(3) *for $f, g \in S$ and Y a clopen subset of B^*, $(f \restriction Y) \cup (g \restriction (B^* \smallsetminus Y) \in S$*

Consequently, the Boolean power $A[B]^*$ is a Boolean product (of algebras equal to A).

PROOF. Let $S = A[B]^*$. (1) is clear. For $f, g \in S$, the set $\mathbf{e}(f = g)$ is the union of the clopen sets $f^{-1}(a) \cap g^{-1}(a)$ with a ranging over A, and only finitely many of these intersections are nonempty. For $h = (f \restriction Y) \cup (g \restriction (B^* \smallsetminus Y)$ where $f, g \in S$ and Y is clopen, we have $h^{-1}(a) = (f^{-1}(a) \cap Y) \cup (g^{-1}(a) \cap (B^* \smallsetminus Y))$ which is a clopen set.

Let the three conditions be satisfied. If $f \in S$ then for any $a \in A$ we have $f^{-1}(a) = \mathbf{e}(f = c_a)$ where c_a is the constant map with range $\{a\}$, so that $f^{-1}(a)$ is clopen by (2) and $f \in A[B]^*$. If $f \in A[B]^*$ then it follows from (3) that $f = \bigcup_{a\in A}(c_a \restriction \mathbf{e}(f = c_a)) \in S$. □

5.3. THEOREM. *Let A, A_1, A_2 be σ-algebras and B_1, B_2 be Boolean algebras.*

(1) *Where B is the two-element Boolean algebra, we have $A[B^n]^* = A^n$*
(2) $A[B_1 \times B_2]^* \simeq A[B_1]^* \times A[B_2]^*$

(3) $(A_1 \times A_2)[B]^* \simeq A_1[B]^* \times A_2[B]^*$

Proof. It is easy. $\square$

CHAPTER 5

MODEL THEORY

1. Formulas

Let σ be a given language.

By a *formula* of the language σ we mean any word over the infinite alphabet, consisting of countably many variables, the operation and relation symbols of σ and the symbols $\approx, \neg, \sqcap, \sqcup, \rightarrow, \forall, \exists, (,)$, that can be obtained by finitely many applications of the following rules:

(1) If u, v are two terms of (the underlying signature of) σ, then $u \approx v$ is a formula (these can be identified with equations);
(2) If R is a relation symbol of arity n and $t_1, \dots, t_n$ are terms, then $R(t_1 \dots t_n)$ is a formula (these are called *atomic formulas*);
(3) If f, g are two formulas and x is a variable, then

$$(\neg f), \quad (f \sqcap g), \quad (f \sqcup g), \quad (f \rightarrow g), \quad (\forall x f), \quad (\exists x g)$$

are also formulas.

The symbols $\neg, \sqcap, \sqcup, \rightarrow, \forall, \exists$ are called the *negation, conjunction, disjunction, implication, universal quantifier* and *existential quantifier*, respectively. In particular formulas, parentheses will be omitted at places where this does not cause any confusion. We consider $f \leftrightarrow g$ an abbreviation for $(f \rightarrow g) \sqcap (g \rightarrow f)$.

Formulas considered here are objects of mathematics, while those in Chapter 1 are at the level of metamathematics.

By an *interpretation* in a structure A we mean a homomorphism of the algebra of terms into the underlying algebra of A. Given an interpretation h in A, a variable x and an element $a \in A$, we denote by $h_{x:a}$ the unique interpretation in A such that $h_{x:a}(x) = a$ and $h_{x:a}(y) = h(y)$ for all variables $y \neq x$.

By induction on the length of a formula f, we define the meaning of the phrase 'f is satisfied in a structure A under an interpretation h', as follows:

(1) An equation $u \approx v$ is satisfied in A under h if $h(u) = h(v)$;
(2) An atomic formula $R(t_1 \dots t_n)$ is satisfied in A under h if the n-tuple $(h(t_1), \dots, h(t_n))$ belongs to R_A;
(3) $\neg f$ is satisfied in A under h if f is not satisfied in A under h;
(4) $f \sqcap g$ is satisfied in A under h if both f and g are satisfied in A under h;
(5) $f \sqcup g$ is satisfied in A under h if at least one of the formulas, either f or g, is satisfied in A under h;

(6) $f \to g$ is satisfied in A under h if either f is not, or g is satisfied in A under h;
(7) $\forall x f$ is satisfied in A under h if for every element $a \in A$, f is satisfied in A under $h_{x:a}$;
(8) $\exists x f$ is satisfied in A under h if there exists an element $a \in A$ such that f is satisfied in A under $h_{x:a}$.

For any formula f we define a finite set of variables, called the *free variables* in f, as follows:

(1) If f is either an equation or an atomic formula, then a variable is free in f if and only if it occurs in f;
(2) If $f = \neg g$, then a variable is free in f if and only if it is free in g;
(3) If f is either $g_1 \sqcap g_2$ or $g_1 \sqcup g_2$ or $g_1 \to g_2$, then a variable is free in f if and only if it is free in either g_1 or g_2;
(4) If f is either $\forall x g$ or $\exists x g$, then a variable is free in f if and only if it is free in g and different from x.

One can easily prove that if h_1 and h_2 are two interpretations in A such that $h_1(x) = h_2(x)$ for all variables x free in f, then f is satisfied in A under h_1 if and only if it is satisfied in A under h_2.

We say that a formula is *satisfied* in A if it is satisfied in A under any interpretation. By a *tautology* we mean any formula which is satisfied in all structures (of the given language). Two formulas f, g are said to be *equivalent* if $f \leftrightarrow$ is a tautology.

Clearly, $f \sqcup g$ is equivalent to $\neg(\neg f \sqcup \neg g)$, $f \to g$ is equivalent to $\neg(f \sqcap \neg g)$, and $\forall x f$ is equivalent to $\neg(\exists x(\neg f))$. So, if we want to prove by induction that all formulas have a given property, and if it is clear that the property is preserved under equivalence of formulas, then it is sufficient to perform the induction steps for $\neg, \sqcap, \exists$ only.

By a *sentence* we mean a formula without free variables. The *closure* of a formula f is the sentence $\forall x_1 \dots \forall x_n f$, where $x_1, \dots, x_n$ are all the variables free in f. Clearly, a formula is satisfied in A if and only if its closure is satisfied in A.

For a formula f and a substitution s (i.e., an endomorphism of the structure of terms) we define a formula $s(f)$ as follows:

(1) If f is an equation $u \approx v$, then $s(f)$ is the equation $s(u) \approx s(v)$;
(2) If f is an atomic formula $R(t_1 \dots t_n)$, then $s(f)$ is the atomic formula $R(s(t_1) \ \dots \ s(t_n))$;
(3) If f is either $\neg g$ or $g_1 \sqcap g_2$ or $g_1 \sqcup g_2$ or $g_1 \to g_2$, then $s(f)$ is either $\neg s(g)$ or $s(g_1) \sqcap s(g_2)$ or $s(g_1) \sqcup s(g_2)$ or $s(g_1) \to s(g_2)$, respectively;
(4) If f is either $\forall x g$ or $\exists x g$, then $s(f)$ is either $\forall x s'(g)$ or $\exists x s'(g)$, respectively, where s' is the substitution such that $s'(x) = x$ and $s'(y) = s(y)$ for all variables $y \neq x$.

Also, we define what we mean by saying that a substitution is *good* for a given formula:

(1) If f is either an equation or an atomic formula, then every substitution is good for f;
(2) A substitution is good for $\neg f$ if and only if it is good for f;
(3) A substitution is good for $f \sqcap g$ (or $f \sqcup g$, or $f \to g$) if and only if it is good for both f and g;
(4) A substitution s is good for $\forall x f$ (or $\exists x f$) if and only if s' is good for f and x does not occur in $s(y)$ for any variable y free in f, where s' is the substitution such that $s'(x) = x$ and $s'(y) = s(y)$ for all variables $y \neq x$.

One can easily prove that if s is a substitution good for a formula f and if h is an interpretation in A, then $s(f)$ is satisfied in A under h if and only if f is satisfied in A under hs.

Let f be a formula and x be a variable. Take the first variable y different from x and not occurring in f. The substitution s, such that $s(x) = y$ and $s(z) = z$ for all variables $z \neq x$, is good for f; the formula $\forall x \forall y((f \sqcap s(f)) \to x \approx y)$ is denoted by $\exists{*}xf$. The formula $(\exists xf) \sqcap (\exists{*}xf)$ is denoted by $\exists!xf$.

2. Theories

By a *theory* (of a given language) we mean an arbitrary set of formulas (of the given language). These formulas are called *axioms* of the theory.

By a *model* of a theory T we mean any structure in which all the axioms of T are satisfied. The class of all models of a theory T is denoted by $\mathbf{Mod}(T)$. A theory is said to be *consistent* if it has at least one model; in the opposite case, it is called *inconsistent*.

By a *consequence* of a theory T we mean any formula which is satisfied in all models of T. Instead of saying that f is a consequence of T, we also write $T \vdash f$.

2.1. Theorem. *The following are equivalent for a theory T:*

(1) *T is inconsistent;*
(2) *$T \vdash f$ for every formula f;*
(3) *There exists a formula f such that both $T \vdash f$ and $T \vdash \neg f$.*

Proof. It is easy. □

3. Ultraproducts

By a filter (or ultrafilter) over a set I we mean a filter (or ultrafilter, resp.) of the Boolean algebra of all subsets of I.

Let H be a family of structures over a set I and let U be a filter over I. Define a relation $\sim$ on the product ΠH in this way: $f \sim g$ if and only if $\{i \in I : f(i) = g(i)\} \in U$. It is easy to see that $\sim$ is a congruence of ΠH; it is called the congruence induced by U. Let us define a structure A with the underlying set $\Pi H/{\sim}$ as follows: the operations are those of the structure $\Pi H/{\sim}$; for a relation symbol R of arity n, let $\langle f_1/{\sim}, \dots, f_n/{\sim} \rangle \in R_A$ if and only if $\{i \in I : \langle f_1(i), \dots, f_n(i) \rangle \in R_{H_i}\} \in U$. The structure A is denoted by

$\Pi_U H$ and is called the *reduced product* of H through U; if U is an ultrafilter over I, it is called the *ultraproduct* of H through U. If $H_i = A$ for all $i \in I$, then $\Pi_U H$ is called the *ultrapower* of A over U.

3.1. LEMMA. *Let H be a family of structures over a set I and let U be an ultrafilter over I. For $i \in I$ denote by p_i the projection of ΠH onto H_i. A formula f is satisfied in the ultraproduct $\Pi_U H$ under an interpretation h if and only if the set of the indexes i, such that f is satisfied in H_i under $p_i h$, belongs to I.*

PROOF. It is easy by induction on the length of f. □

3.2. THEOREM. *Let H be a family of structures over a set I and let U be an ultrafilter over I. A formula f is satisfied in the ultraproduct $\Pi_U H$ if and only if the set of the indexes i, such that f is satisfied in H_i, belongs to I.*

PROOF. It follows from 3.1; consider the closure of f. □

3.3. THEOREM. *Let H be a family of structures over a set I and let H be the principal ultrafilter over I generated by $\{i_0\}$, for an element $i_0 \in I$. Then $\Pi_U H$ is isomorphic to H_{i_0}. In particular, if I is finite, then every ultraproduct of H is isomorphic to H_i for some $i \in I$.*

PROOF. It is easy. □

3.4. THEOREM. *Every structure A is isomorphic to a substructure of an ultraproduct of its finitely generated substructures.*

PROOF. Denote by I the set of all nonempty finite subsets of A. For $i \in I$ denote by A_i the substructure of A generated by i and denote by J_i the set of all $j \in I$ for which $i \subseteq j$. Clearly, $J_i \cap J_j = J_{i \cup j}$ for $i, j \in I$ and so there exists an ultrafilter U over I such that $J_i \in U$ for all $i \in I$. Denote by B the product of the family A_i $(i \in I)$, by $\sim$ the congruence of B induced by U and by C the ultraproduct $B/\sim$. Define a mapping h of A into C as follows: if $a \in A$ then $h(a) = g/\sim$ where g is any element of B such that $g(i) = a$ whenever $a \in i$. It is easy to check that h is an isomorphism of A onto a substructure of C. □

4. Elementary substructures and diagrams

By an *elementary substructure* of a structure A we mean a substructure B such that for any formula f and any interpretation h in B, f is satisfied in A under h if and only if f is satisfied in B under h. We also say that A is an *elementary extension* of B.

For every positive integer n one can easily construct a formula which is satisfied in a structure A if and only if $\mathbf{card}(A) = n$. Consequently, a finite structure has no elementary substructure except itself, and also has no elementary extension except itself.

By an *elementary embedding* of A into B we mean an isomorphism of A onto an elementary substructure of B.

4.1. EXAMPLE. Let A be a structure and U be an ultrafilter over a nonempty set I. We obtain an elementary embedding of A into its ultrapower over I if we assign to any element $a \in A$ the element $p/\sim$, where $p(i) = a$ for all $i \in I$.

4.2. LEMMA. *Let f be a one-to-one mapping of a structure A into a structure B. Then f is an elementary embedding if and only if for any formula f and any interpretation h in A, if f is satisfied in A under h then f is satisfied in B under fh.*

PROOF. If f is not satisfied in A under h then $\neg f$ is, so that $\neg f$ is satisfied in B under fh and f is not. □

4.3. LEMMA. *A substructure B of a structure A is an elementary substructure if and only if for any formula f, any variable x and any interpretation h in B, if $\exists x f$ is satisfied in A under h then there exists an element $b \in B$ such that f is satisfied in A under $h_{x:b}$.*

PROOF. The direct implication is clear. For the converse, we are going to prove by induction on the length of a formula f that for any interpretation h in B, f is satisfied in A under h if and only if f is satisfied in B under h. If f is an equation or an atomic formula, it follows from the fact that B is a substructure. The steps corresponding to $\neg$ and $\sqcap$ are clear, so it remains to consider the step corresponding to $\exists$. If $\exists x f$ is satisfied in A under h then, according to the assumption, there exists an element $b \in B$ such that f is satisfied in A under $h_{x:b}$, which means by the induction assumption that f is satisfied in B under $h_{x:b}$, i.e., $\exists x f$ is satisfied in B under h. If $\exists x f$ is satisfied in B under h, then f is satisfied in B under $h_{x:b}$ for some $b \in B$, so that f is satisfied in A under $h_{x:b}$ by the induction assumption. □

4.4. THEOREM. *Let A be a structure and S be an infinite subset of A such that $\mathbf{card}(S) \geq \mathbf{card}(\sigma)$. Then A has an elementary substructure B such that $S \subseteq B$ and $\mathbf{card}(B) = \mathbf{card}(S)$.*

PROOF. Let us take a well ordering of the set A. Define an infinite sequence $S_0, S_1, \dots$ of subsets of A as follows: $S_0 = S$; S_{n+1} is the set of the elements $b \in A$ for which there exist a formula f, a variable x and an interpretation h in A mapping all variables into S_n, such that b is the least element (with respect to the well ordering) with the property that f is satisfied in A under $h_{x:b}$. The union of this chain of subsets is a substructure of cardinality $\mathbf{card}(S)$, and 4.3 can be used to prove that this is an elementary substructure. □

4.5. THEOREM. *Let a structure A be the union of a set S of its substructures, such that if $B, C \in S$ then either B is an elementary substructure of C or C is an elementary substructure of B. Then A is an elementary substructure of every $B \in S$.*

PROOF. It is easy to prove by induction on the length of a formula f that if $B \in S$ and h is an interpretation in B, then f is satisfied in B under h if and only if it is satisfied in A under h. □

Let A be a structure of a language σ. Denote by $\sigma + A$ the language obtained from σ by adding a new constant c_a for any element $a \in A$, and let A' be the structure of the language $\sigma + A$ obtained from A by adding $(c_a)_{A'} = a$ for all $a \in A$. By the *full diagram* of A we mean the set of the sentences of the language $\sigma + A$ that are satisfied in A'. By the *diagram* of A we mean the subset of the full diagram, consisting of the following sentences:

(1) whenever a, b are two different elements of A, then the formula $\neg(c_a \approx c_b)$ belongs to the diagram;
(2) whenever $F_A(a_1, \dots, a_n) = a$ for an operation symbol F of σ, then the formula $F(c_{a_1}, \dots, c_{a_n}) \approx c_a$ belongs to the diagram;
(3) whenever $\langle a_1, \dots, a_n \rangle \in R_A$ for a relation symbol R of σ, then the formula $R(c_{a_1}, \dots, c_{a_n})$ belongs to the diagram;
(4) whenever $\langle a_1, \dots, a_n \rangle \notin R_A$ for a relation symbol R of σ, then the formula $\neg R(c_{a_1}, \dots, c_{a_n})$ belongs to the diagram.

4.6. Theorem. *Let A be a structure. Models of the full diagram of A are precisely the structures of the language $\sigma + A$ that are isomorphic to an elementary extension of A'. Models of the diagram of A are precisely the structures of the language $\sigma + A$ containing a substructure isomorphic to A'.*

Proof. It is easy. □

5. Elementary equivalence

Two structures A and B (of the same language) are said to be *elementarily equivalent* if any sentence is satisfied in A if and only if it is satisfied in B. (It would be sufficient to say that whenever a sentence is satisfied in A, then it is satisfied in B.)

Isomorphic structures are elementarily equivalent. An elementary substructure of A is elementarily equivalent with A. Every ultrapower of A is elementarily equivalent with A. If A, B are elementarily equivalent and A is finite, then A, B are isomorphic.

5.1. Theorem. *Two structures A and B are elementarily equivalent if and only if A is isomorphic to an elementary substructure of an ultrapower of B.*

Proof. We only need to prove the direct implication. If f is a sentence of the language $\sigma + A$ and d is a mapping of A into B, denote by $f_{:d}$ the sentence of the language $\sigma + B$ obtained from f by replacing all occurrences of c_a, for any $a \in A$, with $c_{d(a)}$. Denote by I the full diagram of A. Let f be a sentence from I and let $a_1, \dots, a_n$ be all the elements of A such that c_a occurs in f. Take pairwise different variables $x_1, \dots, x_n$ not occurring in f and denote by f' the formula obtained from f by replacing c_{a_i} with x_i. Then $\exists x_1 \dots \exists x_n f'$ is a sentence satisfied in A; since A, B are elementarily equivalent, this sentence is satisfied in B, which means that there exists a mapping d of A into B such that the sentence $f_{:d}$ belongs to the full diagram of B. For every sentence $f \in I$ take one such mapping d and denote it by d_f.

For every $f \in I$ denote by I_f the set of the sentences $g \in I$ such that $f_{:d_g}$ belongs to the full diagram of B. Since $f \in I_f$, the set I_f is nonempty. Since $I_{f \sqcap g} \subseteq I_f \cap I_g$, there exists an ultrafilter U over I such that $I_f \in U$ for all $f \in I$. Denote by C the ultrapower of B over U and define a mapping p of A into B^I by $p(a)(f) = d_f(a)$ for $a \in A$ and $f \in I$.

Let f be a formula satisfied in A under an interpretation h. In order to prove that the mapping $a \mapsto p(a)/\sim$ is an elementary embedding of A into C, it is sufficient to prove (acording to 4.2) that f is satisfied in C under $x \mapsto ph(x)/\sim$. Denote by s the substitution such that $s(x) = c_{h(x)}$ for all variables x. We have $s(f) \in I$, so that $I_{s(f)} \in U$. This means that U contains the set of the sentences $g \in I$ for which $s(f)_{:d_g}$ belongs to the full diagram of B, i.e., U contains the set of the sentences $g \in I$ such that f is satisfied in B under $d_g h$. Since $d_g h(x) = ph(x)(g)$, it follows by 3.1 that f is satisfied in C under $x \mapsto ph(x)/\sim$. $\square$

6. Compactness theorem and its consequences

6.1. THEOREM. (Compactness theorem) *A theory T is consistent if and only if every finite subset of T is consistent.*

PROOF. The direct implication is clear. Let us prove the converse. Denote by I the set of all finite subsets of T. For every $S \in I$ take one model A_S of S. For $S \in I$ denote by I_S the set of the finite subsets $H \in I$ such that every formula from S is satisfied in A_H. Since $S \in I_S$, the sets I_S are nonempty. Moreover, we have $I_{S_1 \cup S_2} \subseteq I_{S_1} \cap I_{S_2}$, and so there exists an ultrafilter U over I such that $I_S \in U$ for all $S \in I$. Denote by A the ultraproduct of the family A_S $(S \in I)$ through U. If $f \in T$, then $I_{\{f\}} \in U$, so that U contains the set of the subsets $S \in I$ such that f is satisfied in A_S; hence by 3.2, f is satisfied in A. This means that A is a model of T. $\square$

6.2. THEOREM. *Let T be a theory and f be a formula. Then $T \vdash f$ if and only if there exists a finite subset T' of T such that $T' \vdash f$.*

PROOF. It follows from 6.1 and the obvious fact that if f is a sentence, then f is not a consequence of T if and only if the theory $T \cup \{\neg f\}$ is consistent. $\square$

6.3. THEOREM. *Let T be a theory such that for every positive integer n there exists a model of T of cardinality $\geq n$. Then for every infinite cardinal number $c \geq \mathbf{card}(T)$ there exists a model of T of cardinality c.*

PROOF. Let C be a set of constants not belonging to the language σ, such that $\mathbf{card}(C) = c$. Denote by T' the theory obtained from T by extending the language by the elements of C, and the axioms by the sentences $\neg(a \approx b)$ for any pair a, b of distinct elements of C. It follows from the assumption by 6.1 that T' is consistent. Let A' be a model of T', and A be the structure of the language σ obtained from A by forgetting the new constants. Clearly, $\mathbf{card}(A) \geq c$ and A is a model of T. By 4.4, A has an elementary substructure of the cardinality c. $\square$

7. Syntactic approach

Where f, g are formulas, x, y are variables, u, v, w, t, u_i, v_i are terms, F is an operation symbol of arity $n \geq 1$ and R is a relation symbol of arity n, each of the following formulas is called a *logical axiom*:

(a1) $f \to (g \to f)$,
(a2) $f \to f$,
(a3) $(f \to g) \to ((f \to (g \to h)) \to (f \to h))$,
(a4) $(f \to g) \to ((g \to h) \to (f \to h))$,
(a5) $(f \to g) \to ((h \to k) \to (f \sqcap h \to g \sqcap k))$,
(a6) $(f \to g) \to ((h \to k) \to (f \sqcup h \to g \sqcup k))$,
(a7) $(f \to g) \to (\neg g \to \neg f)$,
(a8) $(f \sqcap g) \to f$,
(a9) $(f \sqcap g) \to g$,
(a10) $f \to (g \to (f \sqcap g))$,
(a11) $f \sqcap g \to g \sqcap f$,
(a12) $f \sqcup g \to g \sqcup f$,
(a13) $(f \sqcap g) \sqcap h \leftrightarrow f \sqcap (g \sqcap h)$,
(a14) $(f \sqcup g) \sqcup h \leftrightarrow f \sqcup (g \sqcup h)$,
(a15) $((f \sqcap g) \sqcup f) \leftrightarrow f$,
(a16) $((f \sqcup g) \sqcap f) \leftrightarrow f$,
(a17) $(f \sqcup g) \sqcap h \leftrightarrow (f \sqcap h) \sqcup (g \sqcap h)$,
(a18) $(f \sqcap \neg f) \to g$,
(a19) $f \sqcup \neg f$,
(a20) $(f \to g) \leftrightarrow (\neg f \sqcup g)$,
(a21) $((f \leftrightarrow g) \sqcap (h \leftrightarrow k)) \to ((f \to h) \leftrightarrow (g \to k))$,
(a22) $(\neg f \to (g \sqcap \neg g)) \to f$,
(a23) $u \approx u$,
(a24) $u \approx v \to v \approx u$,
(a25) $(u \approx v \sqcap v \approx w) \to u \approx w$,
(a26) $(u_1 \approx v_1 \sqcap \cdots \sqcap u_n \approx v_n) \to F u_1 \dots u_n \approx F v_1 \dots v_n$,
(a27) $(u_1 \approx v_1 \sqcap \cdots \sqcap u_n \approx v_n) \to (R u_1 \dots u_n \leftrightarrow R v_1 \dots v_n)$,
(a28) $(\forall x (f \to g)) \to (f \to \forall x g)$ where f is a sentence,
(a29) $f[x : t] \to \exists x f$ where $[x : t]$ is a substitution good for f,
(a30) $(\forall x (f \to g)) \to ((\exists x f) \to g)$ where x does not occur in g,
(a31) $(\exists x f) \to \exists y f[x : y]$ where y does not occur in f,
(a32) $(\forall x (f \leftrightarrow g)) \to ((\exists x f) \leftrightarrow \exists y g[x : y])$ where either $y = x$ or y does not occur in g,
(a33) $(\forall x (f \leftrightarrow g)) \to ((\forall x f) \leftrightarrow \forall y g[x : y])$ where either $y = x$ or y does not occur in g,
(a34) $\forall x f \leftrightarrow \neg \exists \neg f$,
(a35) $(\forall x f) \to f$.

By a *proof in a theory T* we mean a finite sequence $f_1, \dots, f_n$ of formulas such that for every $i = 1, \dots, n$ at least one of the following four cases takes place:

(b1) f_i is a logical axiom;
(b2) $f_i \in T$;
(b3) there are indexes $j, k \in \{1, \dots, i-1\}$ such that f_k is the formula $f_j \to f_i$;
(b4) there are an index $j \in \{1, \dots, i-1\}$ and a variable x such that f_i is the formula $\forall x f_j$.

We say that a formula f is *provable* in T if it is the last member of a proof in T.

Let us remark that the system of logical axioms, given by the list (a1) through (a35), is not independent. It is a good excercise to delete a lot of the items in such a way that provability remains unchanged.

7.1. Lemma. *Let T be a theory, f be a sentence and g be a formula. Then g is provable in $T \cup \{f\}$ if and only if $f \to g$ is provable in T.*

Proof. Let $g_1, \dots, g_n$ be a proof in T, where $g_n = g$. Let us prove by induction on $i = 1, \dots, n$ that $f \to g_i$ is provable in T. If g_i is either a logical axiom or an element of T, it follows from (a1) using (b3); for $g_i = f$ it follows from (a2). If g_k is the formula $g_j \to g_i$ for some $j, k < i$, apply (b3) twice on an axiom of the form (a3). Let g_i be the formula $\forall x g_j$ for some $j < i$. By induction, $f \to g_j$ is provable in T. By (b4), $\forall x(f \to g_j)$ is provable in T. Now (a28) and (b4) give $f \to g_j$.

Conversely, if $f_1, \dots, f_m$ is a proof in T such that f_m is the formula $f \to g$, then $f_1, \dots, f_m, f, g$ is a proof in $T \cup \{f\}$ with the last member g. $\square$

For every theory T define an algebra A of signature $\{\wedge, \vee, ', 0, 1\}$ in this way: The underlying set of A is the set of all formulas (of the given language); $f \wedge g = f \sqcap g$; $f \vee g = f \sqcup g$; $f' = \neg f$; 0_A is the formula $\neg x \approx x$ and 1_A is the formula $x \approx x$ (where x is a variable). Define a binary relation $\sim$ on A as follows: $f \sim g$ if and only if both $f \to g$ and $g \to f$ are provable in T.

7.2. Lemma. *The relation $\sim$ is a congruence of A and the factor $A/{\sim}$ is a Boolean algebra.*

Proof. Clearly, the relation is symmetric. Its reflexivity and transitivity follow from (a2) and (a4). Then it follows from (a5), (a6) and (a7) that $\sim$ is a congruence. It follows from (a8), (a9) and (a10) that a formula $f \sqcap g$ is provable in T if and only if both f and g are provable in T. In particular, we have $f \sim g$ if and only if $f \leftrightarrow g$ is provable in T. Now the axioms (a11) through (a19) are transcripts of the defining equations for Boolean algebras. $\square$

The Boolean algebra $B = A/{\sim}$ is called the *Lindenbaum algebra* of the theory T. It is easy to see that for two formulas f and g, $f/{\sim} \le g/{\sim}$ if and only if $f \to g$ is provable in T. A formula f is provable in T if and only if $f/{\sim} = 1_B$.

7.3. Lemma. *Let T be a theory and $A/{\sim}$ be the Lindenbaum algebra of T; let f be a formula and x be a variable. Then $(\exists x f)/{\sim}$ is the join of the elements*

$(f[x : t])/{\sim}$ in $A/{\sim}$, where t runs over the terms such that the substitution $[x : t]$ is good for f.

Proof. It follows from (a29) that $(\exists x f)/{\sim}$ is an upper bound of the set. Let $g/{\sim}$ be any upper bound of the set. Take a variable y such that y occurs in neither f nor g. Since $g/{\sim}$ is an upper bound, $f[x : y] \to g$ is provable in T. Now (b4) gives $\forall y(f[x : y] \to g)$ and (a30) gives $(\exists y f[x : y]) \to g$. From this, using (a31), we get $(\exists x f)/{\sim} \le (\exists y f[x : y])/{\sim} \le g/{\sim}$. □

By an *adequate ultrafilter* of the Lindenbaum algebra $A/{\sim}$ we mean an ultrafilter U such that for any $(\exists x f)/{\sim} \in U$ there exists a term t such that $[x : t]$ is good for f and $(f[x : t])/{\sim} \in U$.

Let T be a theory and U be an adequate ultrafilter of the corresponding Lindenbaum algebra $A/{\sim}$. Define a relation r on the algebra $\mathbf{T}$ of terms (over the set of variables) by $\langle u, v\rangle \in r$ if and only if $(u \approx v)/{\sim} \in U$. It follows from (a23) through (a26) that r is a congruence of T. Define a structure C as follows: its underlying algebra is the algebra $\mathbf{T}/r$; for an n-ary relation symbol R we have $\langle u_1/r, \dots, u_n/r\rangle \in R_C$ if and only if $(Ru_1 \dots u_n)/{\sim} \in U$. Correctness of this definition follows from (a27).

7.4. Lemma. *Denote by H the canonical homomorphism $t \mapsto t/{\sim}$ of $\mathbf{T}$ onto $\mathbf{T}/{\sim}$. A formula f is satisfied in C under the interpretation H if and only if $f/{\sim} \in U$.*

Proof. Let us fix an ordering $x_0, x_1, \dots$ of the set of variables. For every formula h and every positive integer n define a formula $c_n(h)$ in this way: $c_n(h) = h$ for h atomic; $c_n(h_1 \sqcap h_2) = c_n(h_1) \sqcap c_n(h_2)$; similarly for the symbols $\sqcup, \neg$ and $\to$; if $h = Qxh'$ where Q is a quantifier, then $c_n(h) = Qx_m(c_n(h')[x : x_m])$ where m is the least index such that $n \le m$ and x_m does not occur in $c_n(h')$. Clearly, h is equivalent with $c_n(h)$. Using (a5), (a6), (a7), (a21), (a32) and (a33) one can prove by induction for any formula h that $h \leftrightarrow c_n(h)$ is provable in T.

We will prove the lemma by double induction: by induction on the number $d(f)$ of occurrences of $\forall$ in f, and for a given $d(f)$ by induction on the number of occurrences of $\sqcap, \sqcup, \neg, \to, \exists$ in f. If none of these symbols occurs in f, then f is atomic and the statement follows from the construction of C. If f is either $g \sqcap h$ or $g \sqcup h$ or $\neg g$ for some formulas g and h, the proof is easy: it only uses the fact that U is an ultrafilter. If f is $g \to h$, we can apply (a20). If f is $\forall x g$, use the double induction and apply (a34). It remains to consider the case when f is $\exists x g$.

Let f be satisfied in C under H, so that there exists a term t such that g is satisfied in C under $H_{x:t/r}$. Clearly, there exists an n such that the substitution $[x : t]$ is good for the formula $\bar{g} = c_n(g)$. Then $\bar{g}[x : t]$ is satisfied in C under H. By induction, $(\bar{g}[x : t])/{\sim} \in U$. By 7.3, $(\exists x \bar{g})/{\sim} \in U$. Since $g \leftrightarrow \bar{g}$ is provable in T, we have $g/{\sim} = \bar{g}/{\sim}$. Now $f/{\sim} = (\exists x \bar{g})/{\sim} \in U$ by (a31).

Conversely, let $f/{\sim} \in U$. Since U is adequate, there exists a term t such that $[x : t]$ is good for g and $(g[x : t])/{\sim} \in U$. By induction, $g[x : t]$ is satisfied in C under H. From this is tollows that g is satisfied in C under $H_{x:t/r}$, so that f is satisfied in C under H. □

7.5. Lemma. *C is a model of T.*

Proof. Let $f \in T$. It follows from (b4) that the closure g of f is provable in T, so that $g/\sim = 1_{A/\sim} \in U$. It follows by 7.4 that g is satisfied in C. But then, f is satisfied in C. □

7.6. Theorem. *Let T be a theory. A formula f is a consequence of T if and only if it is provable in T.*

Proof. Clearly, it is sufficient to consider the case when f is a sentence. The direct implication is clear. Let f be a consequence of T. By 6.2 there exists a finite subset T' of T such that f is a consequence of T'. We can consider T' as a theory of a finite sublanguage σ' of σ. Suppose that f is not provable in T. Then f is not provable in T'. So, using (a22) it follows from 7.1 that no formula $g \sqcap \neg g$ is provable in the theory $T' \cup \{\neg f\}$. This means that the Lindenbaum algebra $A/\sim$ of $T' \cup \{\neg f\}$ has at least two elements. Since this is a theory of a finite language, the Lindenbaum algebra is countable and it follows from 4.4.4 that it has an adequate ultrafilter U. The corresponding structure C is a model of $T' \cup \{\neg f\}$ by 7.5; but it also satisfies f, and we get a contradiction. □

8. Complete theories

A theory T is said to be *complete* if it is consistent and for every sentence f of the language σ, either f or $\neg f$ is a consequence of T.

8.1. Theorem. *Let A be a structure. Then the set of all sentences that are satisfied in A is a complete theory.*

Proof. It is evident. □

8.2. Theorem. *Every consistent theory is contained in a complete theory.*

Proof. If T is a consistent theory, then T has a model A. According to 8.1, the set of the sentences satisfied in A is a complete theory. □

Let c be a cardinal number. A theory T is said to be *categorical in the cardinality* c if every two models of T of the cardinality c are isomorphic.

8.3. Theorem. *Let T be a consistent theory and let T be categorical in some infinite cardinality $c \geq \mathbf{card}(\sigma)$. Then T is complete.*

Proof. Suppose that there is a sentence f such that neither f nor $\neg f$ is a consequence of T. Then both $T \cup \{f\}$ and $T \cup \{\neg f\}$ are consistent. According to 6.3, each of these two theories has a model of cardinality c. Since T is categorical in c, the two models are isomorphic, a contradiction. □

8.4. Example. The theory of dense linearly ordered sets without extreme elements is categorical in the countable cardinality. (A linearly ordered set A is said to be dense if for every $a, b \in A$ with $a < b$ there exists an element $c \in A$ with $a < c < b$.) The theory of Boolean algebras without atoms is categorical in the countable cardinality.

9. Axiomatizable classes

A class of structures is said to be *axiomatizable* if it is the class of all models of some theory.

9.1. Theorem. *A class of structures is axiomatizable if and only if it is closed under ultraproducts, isomorphic images and elementary substructures.*

Proof. The direct implication is clear. Let K be a class of structures closed under ultraproducts, isomorphic images and elementary substructures. It follows from 5.1 that K is closed under elementary equivalence. Denote by T the set of the sentences that are satisfied in every structure from K. Since $K \subseteq \mathbf{Mod}(T)$ is evident, it remains to prove $\mathbf{Mod}(T) \subseteq K$. Let $A \in \mathbf{Mod}(T)$. Denote by I the set of the sentences satisfied in A. For every $f \in I$ there exists a structure $A_f \in K$ such that f is satisfied in A_f. (In the opposite case $\neg f$ would be satisfied in every structure from K, so that $\neg f \in T$ and $\neg f$ would be satisfied in A.) For every $f \in I$ denote by I_f the set of all $g \in I$ such that f is satisfied in A_g. The set I_f is nonempty, since $f \in I_f$. Since $I_{f \sqcap g} \subseteq I_f \cap I_g$, there exists an ultrafilter U over I such that $I_f \in U$ for all $f \in I$. The ultraproduct B of the family A_f $(f \in I)$ over U belongs to K, since K is closed under ultraproducts. It remains to prove that the structures A and B are elementarily equivalent, i.e., that every sentence $f \in I$ is satisfied in B. But this follows from the fact that the set of all $g \in I$ such that f is satisfied in A_g belongs to U. □

A class of structures is said to be *finitely axiomatizable*, or also *elementary*, if it is the class of all models of a theory with finitely many axioms; in that case, it is the class of all models of a theory with a single axiom.

9.2. Theorem. *A class K of structures is finitely axiomatizable if and only if both K and the complement of K in the class of all structures of the language σ are axiomatizable.*

Proof. If K is axiomatizable by a sentence f, then the complement is axiomatizable by $\neg f$. Let T_1 and T_2 be two theories such that $\mathbf{Mod}(T_1) = K$ and $\mathbf{Mod}(T_2)$ is the complement of K; we can assume that T_1 and T_2 are sets of sentences. The theory $T_1 \cup T_2$ is inconsistent, so that (according to the Compactness theorem) it contains an inconsistent finite subset; in particular, there exists a finite subset $\{f_1, \dots, f_n\}$ of T_2 such that the theory $T_1 \cup \{f_1, \dots, f_n\}$ is inconsistent. Clearly, K is the class of all odels of the sentence $\neg(f_1 \sqcap \dots \sqcap f_n)$. □

10. Universal classes

Let A be a structure of the language σ; let S be a nonempty finite subset of A and τ be a finite sublanguage of σ. We denote by $A \restriction (S, \tau)$ the reduct of the partial structure $A \restriction S$ to τ.

Let A be a structure and K be a class of structures of language σ. We say that A is *locally embeddable* into K if for every nonempty finite subset S of A and every finite sublanguage τ of σ there exists a structure $B \in K$ such that $A \restriction (S, \tau)$ is isomorphic to $B \restriction (S', \tau)$ for some subset S' of B.

By a *universal formula* we mean a formula containing no quantifiers.

10.1. Theorem. *The following are equivalent for a class K of structures of the language σ:*

(1) *K is axiomatizable and closed under substructures;*
(2) *K is the class of all models of a theory, all the axioms of which are universal formulas;*
(3) *every σ-structure that is locally embeddable into K belongs to K;*
(4) *K is closed under substructures, ultraproducts and isomorphic images.*

Proof. The equivalence of (1) with (4) follows from 9.1.

(1) implies (2): Let $K = \mathbf{Mod}(T)$ be closed under substructures. Denote by Z the set of the universal formulas satisfied in all structures from K. It is sufficient to prove that every structure $A \in \mathbf{Mod}(Z)$ is isomorphic to a substructure of a structure belonging to K; according to 4.6, we need to prove that the union of T with the diagram of A is a consistent theory of the language $\sigma + A$. Suppose that this theory is inconsistent. According to 6.1, there exists a finite subset $\{f_1, \dots, f_n\}$ of the diagram of A such that the theory $T \cup \{f_1, \dots, f_n\}$ is inconsistent. Put $f = f_1 \sqcap \dots \sqcap f_n$ and let $c_{a_1}, \dots, c_{a_m}$ be all the constants occurring in f and not belonging to σ. Take pairwise different variables $x_1, \dots, x_m$ not occurring in f and denote by g the formula obtained from f by replacing each c_{a_i} with x_i. If the formula $\exists x_1 \dots \exists x_m g$ is satisfied in a structure $B \in K$, then g is satisfied in B under an interpretation h and the structure C of the language $\sigma + A$, such that B is a reduct of C and $(c_{a_i})_C = h(x_i)$ for all i, is a model of the inconsistent theory $T \cup \{f\}$. Hence $\neg \exists x_1 \dots \exists x_m g$ is satisfied in every structure $B \in K$. But then the universal formula $\neg g$ is satisfied in every structure from K, so that it belongs to Z and is satisfied in A, a contradiction.

(2) implies (3): Let $K = \mathbf{Mod}(T)$ where T is a set of universal formulas, and let A be locally embeddable into K. It is sufficient to prove that every formula f is satisfied in A under an arbitrary interpretation h. Denote by S the set of the elements $h(t)$, where t runs over all subterms of f (it should be clear what do we mean by a subterm of a formula), and denote by τ the sublanguage of σ consisting of the symbols occurring in f. There exists a structure $B \in K$ such that $A \restriction (S, \tau)$ is isomorphic to $B \restriction (S', \tau)$ for a subset S' of B. Since f is a universal formula satisfied in B, it is easy to see that f is satisfied in A under h.

(3) implies (1): Clearly, K is closed under substructures. We are going to prove that $K = \mathbf{Mod}(T)$, where T is the set of the universal formulas satisfied in all structures from K. Let $A \in \mathbf{Mod}(T)$. In order to prove $A \in K$, it is enough to show that for any nonempty finite subset $S = \{a_1, \dots, a_n\}$ of A and any finite sublanguage τ of σ, there is a structure $B \in K$ such that $A \restriction (S, \tau)$ is isomorphic to $B \restriction (S', \tau)$ for a subset S' of B. Suppose that there is no such B in K. Denote by M the set of the formulas from the diagram of A that do not contain other operation and relation symbols than those belonging to $\tau \cup \{c_{a_1}, \dots, c_{a_n}\}$. Take pairwise different variables $x_1, \dots, x_n$ and denote by $f_1, \dots, f_m$ all the τ-formulas obtained from formulas belonging to M by replacing the constants c_{a_i} with x_i. Put $f = f_1 \sqcap \dots \sqcap f_m$. Clearly, the sentence $\exists x_1 \dots \exists x_n f$ is satisfied in a σ-structure B if and only if $A \restriction (S, \tau)$ is isomorphic to $B \restriction (S', \tau)$ for a subset S' of B. So, this sentence is not satisfied in any structure from K. But then the universal formula $\neg f$ is satisfied in all structures from K, so that it belongs to T and is satisfied in A. But f is satisfied in A under an interpretation, a contradiction. □

A class of structures is said to be *universal* if it satisfies any of the equivalent conditions of Theorem 10.1.

10.2. THEOREM. *Let K be a universal class of structures. A structure A belongs to K if and only if every finitely generated substructure of A belongs to K.*

PROOF. It is easy. □

11. Quasivarieties

By a *quasiequation* we mean a formula of the form $(f_1 \sqcap \dots \sqcap f_n) \to f$ where $n \geq 0$ and $f_1, \dots, f_n, f$ are atomic formulas. (For $n = 0$ the quasiequation is just f.)

By a *quasivariety* we mean the class of all models of a theory, all the axioms of which are quasiequations.

For a class K of structures denote by $\mathbf{P}_R(K)$ the class of reduced products and by $\mathbf{P}_U(K)$ the class of ultraproducts of arbitrary families of structures from K.

11.1. THEOREM. *The following are equivalent for a class K of structures of the language σ:*

(1) *K is a quasivariety;*
(2) *K is axiomatizable and closed under substructures and direct products;*
(3) *K is closed under substructures, direct products and ultraproducts;*
(4) *K is closed under substructures and reduced products;*
(5) *K is universal and closed under direct products of finitely many structures;*
(6) *K is closed under products of finitely many algebras and every structure that is locally embeddable into K belongs to K.*

The class $\mathbf{ISPP}_U(K) = \mathbf{ISP}_R(K)$ *is the quasivariety generated by* K.

Proof. The implications (1)$\rightarrow$(2)$\rightarrow$(3)$\rightarrow$(5)$\leftrightarrow$(6) are clear or follow from 10.1. Let us prove that (5) implies (1). Denote by T the set of the universal formulas satisfied in all structures from K, so that $K = \mathbf{Mod}(T)$. Denote by Y the set of the formulas of the form $f_1 \sqcup \dots \sqcup f_n$ ($n \geq 1$) belonging to T and satisfying the following two conditions:

(i) there exists a number $p \in \{0, \dots, n\}$ such that $f_1, \dots, f_p$ are atomic formulas and $f_{p+1}, \dots, f_n$ are negations of atomic formulas;

(ii) if $n \geq 2$ then the formula $g_i = f_1 \sqcup \dots \sqcup f_{i-1} \sqcup f_{i+1} \sqcup \dots \sqcup f_n$ does not belong to T for any $i \in \{1, \dots, n\}$.

It is easy to prove by induction that every universal formula is equivalent to a formula of this form, so that $K = \mathbf{Mod}(Y)$.

Let $f = f_1 \sqcup \dots \sqcup f_n \in Y$, and let p and g_i be as in (i) and (ii). Suppose $p \geq 2$. For every $i = 1, \dots, p$ there exist a structure $A_i \in K$ and an interpretation h_i in A_i such that g_i is not satisfied in A_i under h_i. Put $A = A_1 \times \dots \times A_p$ and define an interpretation h in A by $h(t)(i) = h_i(t)$. Since $A \in K$, f is satisfied in A under h; hence there exists an $i \in \{1, \dots, n\}$ such that f_i is satisfied in A under h. If $i \leq p$, then f_i is satisfied in A_j under h_j for any $j \in \{1, \dots, p\}$; for $j \neq i$ it follows that g_j is satisfied in A_j under h_j, but this contradicts the choice of A_j. Hence $i \geq p + 1$. Since f_i is a negation of an atomic formula and f_i is satisfied in A under h, there exists an index $j \in \{1, \dots, p\}$ such that f_i is satisfied in A_j under h_j; but then g_j is satisfied in A_j under h_j, a contradiction.

We have proved $p \leq 1$. We cannot have $p = 0$, since f is satisfied in the product of the empty family of structures. Hence $p = 1$ and f is equivalent to $(f_2 \sqcap \dots \sqcap f_n) \rightarrow f_1$.

It remains to prove the last statement. Let us first prove that $\mathbf{P}_R\mathbf{P}_R(K) \subseteq \mathbf{IP}_R(K)$. Let $A \in \mathbf{P}_R\mathbf{P}_R(K)$. There exist a set J, a family I_j ($j \in J$) of pairwise disjoint sets, structures A_i ($i \in I = \bigcup\{I_j : j \in J\}$), a filter U over J and filters U_j over I_j such that $A = \Pi_{j \in J}((\Pi_{i \in I_j} A_i)/\!\sim U_j)/\!\sim_U$ (here $\sim_U$ denotes the congruence induced by U). Denote by F the set of the subsets S of I such that $\{j \in J : S \cap I_j \in U_j\} \in U$. Clearly, F is a filter over I. Define a mapping h of $\Pi_{i \in I} A_i$ into A by $h(a) = b_a/\!\sim_U$ where $b_a(j) = (a \restriction I_j)/\!\sim_{U_j}$ for all $j \in J$. One can easily check that h is a surjective homomorphism, that the kernel of h is the congruence $\sim_F$ and that the corresponding bijection of $(\Pi_{i \in I} A_i)/\!\sim_F$ onto A is an isomorphism.

Next let us prove that $\mathbf{P}_R(K) \subseteq \mathbf{ISPP}_U(K)$. Let A_i ($i \in I$) be a family of structures from K and F be a filter over I. Denote by J the set of all the ultrafilters U over I such that $F \subseteq U$. Define a mapping h of $(\Pi_{i \in I} A_i)/\!\sim_F$ into $\Pi_{U \in J}((\Pi_{i \in I} A_i)/\!\sim_U)$ by $h(a/\!\sim_F)(U) = a/\!\sim_U$. Since (as it is easy to see) F is the intersection of the ultrafilters from J, the mapping h is injective. One can easily check that h is an isomorphism.

Clearly, $\mathbf{P}_R\mathbf{S}(K) \subseteq \mathbf{ISP}_R(K)$.

We have $\mathbf{ISP}_R(K) \subseteq \mathbf{ISPP}_U(K) \subseteq \mathbf{ISP}_R\mathbf{P}_R(K) \subseteq \mathbf{ISP}_R(K)$, from which it follows that $\mathbf{ISPP}_U(K) = \mathbf{ISP}_R(K)$. The rest is clear. $\square$

11.2. Theorem. *For a finite structure A, the quasivariety generated by A equals $\mathbf{ISP}(A)$.*

Proof. It follows from 11.1. $\square$

CHAPTER 6

VARIETIES

1. Terms: Syntactic notions

The notion of a *subterm* of a term u can be defined by induction on $\lambda(t)$ as follows: if $t \in X$, then u is a subterm of t if and only if $u = t$; if $t = Ft_1 \dots t_n$, then u is a subterm of t if and only if either $u = t$ or u is a subterm of one of the terms $t_1, \dots, t_n$. Clearly, every term has only finitely many subterms. Instead of saying that u is a subterm of t, we will often write $u \subseteq t$; we hope that this can cause no confusion.

The set of elements of X that occur in t (i.e., are subterms of t,) will be denoted by $\mathbf{S}(t)$ and called the *support* of t. It is always a finite subset of X.

By an *elementary address* we mean an ordered pair $\langle F, i\rangle$ where $F \in \sigma$ is an operation symbol of a positive arity n and $i \in \{1, \dots, n\}$. By an *address* we mean a finite (possibly empty) sequence of elementary addresses. Any two addresses can be concatenated to form a new address. We say that an address a is an *initial segment* of an address b, and that b is an *extension* of a, is $b = ca$ for some address c. Two addresses are said to be *incomparable* if neither is an extension of the other.

Let a be an address. For some terms t, we are going to define a subterm $t[a]$ of t, called the subterm of t at address a, in the following way. If $a = \emptyset$, then $t[a] = t$ for any term t. If $a = \langle F, i\rangle b$ for an elementary address $\langle F, i\rangle$ and some address b, then $t[a]$ is defined if and only if $t = Ft_1 \dots t_n$ for some terms $t_1, \dots, t_n$ and $t_i[b]$ is defined; if so, put $t[a] = t_i[b]$. If $t[a] = u$, we say that a is an *occurrence* of u in t; it is easy to prove that u is a subterm of t if and only if it has at least one occurrence in t. For a given term t, the set of occurrences of subterms in t will be denoted by $\mathbf{O}(t)$. This set is always finite; its maximal elements (with respect to the ordering by extension) are just the occurrences of elements of X and constants in t. We denote by $\mathbf{O}_X(t)$ the set of occurrences of elements of X in t. For two terms t and u, we denote by $|t|_u$ the number of occurrences of u in t.

Let a be an occurrence of a subterm u in a term t, and let s be a term. Then there is a unique term r such that $r[a] = s$ and $r[b] = t[b]$ for every address b which is incomparable with a. This term r is called the term obtained from t by *replacing* u with s at the address a. We denote r by $t(a : u \to s)$.

A *substitution* can be most easily defined as an endomorphism of the term algebra. If $x_1, \dots, x_n$ are pairwise different elements of X and $u_1, \dots, u_n$ are any terms, then we denote by $[x_1 : u_1, \dots, x_n : u_n]$ the substitution f with

$f(x_i) = u_i$ for $i = 1, \dots, n$ and $f(x) = x$ for $x \in X \smallsetminus \{x_1, \dots, x_n\}$; the term $f(t)$ will be then denoted by $t[x_1 : u_1, \dots, x_n : u_n]$.

Sometimes, a term t will be denoted by something like $t(x_1, \dots, x_n)$. We will mean that t is a term, $x_1, \dots, x_n$ are pairwise different elements of X and $\{x_1, \dots, x_n\} \subseteq \mathbf{S}(t)$. If this notation for t is used, then for any n-tuple $u_1, \dots, u_n$, $t(u_1, \dots, u_n)$ will stand for $t[x_1 : u_1, \dots, x_n : u_n]$.

By a *substitution instance* of a term u we mean any term $f(u)$, where f is a substitution.

It is easy to characterize automorphisms of the algebra $\mathbf{T}_X$: they are just the extensions of a permutation of X to a substitution. (The proof is obvious.)

Given two terms u and v, we write $u \leq v$ if $v = Lf(u)$ for a lift L and a substitution f. This is a quasiordering on the set of terms. Two terms u, v are called (*literally*) *similar* if $u \leq v$ and $v \leq u$; we then write $u \sim v$. Also, $u \sim v$ if and only if $v = \alpha(u)$ for an automorphism α of the term algebra. Factored through this equivalence, the set of terms becomes a partially ordered set every principal ideal of which is finite. We write $u < v$ if $u \leq v$ and $v \not\leq u$.

Two finite sequences $u_1, \dots, u_n$ and $v_1, \dots, v_m$ of terms are called similar if $n = m$ and there exists an automorphism α of $\mathbf{T}_X$ with $\alpha(u_i) = v_i$ for $i = 1, \dots, n$.

1.1. LEMMA. *Two finite sequences $u_1, \dots, u_n$ and $v_1, \dots, v_n$ of terms are similar if and only if there exist substitutions f and g such that $f(u_i) = v_i$ and $g(v_i) = u_i$ for $i = 1, \dots, n$.*

PROOF. It is easy. $\square$

2. The Galois correspondence

In the following let $\mathbf{X}$ be a fixed countably infinite set; its elements will be called *variables*. By a term we will mean a term over $\mathbf{X}$. Let the signature σ be fixed.

By an *equation* we will mean an ordered pair of terms. Sometimes an equation $\langle u, v \rangle$ will be denoted by $u \approx v$. The terms u and v are called the *left side* and the *right side* of $\langle u, v \rangle$, respectively.

An equation $\langle u, v \rangle$ is said to be *satisfied* in an algebra A if $f(u) = f(v)$ for any homomorphism f of the term algebra into A.

For a class C of algebras, let $\mathbf{Eq}(C)$ denote the set of the equations that are satisfied in every algebra from C. This set of equations is called the *equational theory of* C. By an equational theory we mean a set of equations E such that $E = \mathbf{Eq}(C)$ for a class of algebras C.

Let E be a set of equations. An algebra A is said to be a *model* of E if every equation from E is satisfied in A. The class of all models of a set of equations E is denoted by $\mathbf{Mod}(E)$. By a *variety* we will mean a class C such that $C = \mathbf{Mod}(E)$ for a set of equations E.

The facts collected in the following theorem are often expressed by saying that the operators $\mathbf{Eq}$ and $\mathbf{Mod}$ form a Galois correspondence between sets of equations and classes of algebras.

2.1. THEOREM. *Let E_1, E_2 and E be sets of equations and C_1, C_2 and C be classes of algebras. Then:*

(1) $C_1 \subseteq C_2$ *implies* $\mathbf{Eq}(C_1) \supseteq \mathbf{Eq}(C_2)$;
(2) $E_1 \subseteq E_2$ *implies* $\mathbf{Mod}(E_1) \supseteq \mathbf{Mod}(E_2)$;
(3) $C \subseteq \mathbf{Mod}(\mathbf{Eq}(C))$;
(4) $E \subseteq \mathbf{Eq}(\mathbf{Mod}(E))$;
(5) $\mathbf{Eq}(\mathbf{Mod}(\mathbf{Eq}(C))) = \mathbf{Eq}(C)$;
(6) $\mathbf{Mod}(\mathbf{Eq}(\mathbf{Mod}(E))) = \mathbf{Mod}(E)$.

PROOF. It is easy. In fact, (5) and (6) are consequences of (1)–(4). $\square$

2.2. THEOREM. *A set E of equations is an equational theory if and only if it is a fully invariant congruence of the term algebra, i.e., a congruence of the term algebra such that $\langle u, v\rangle \in E$ implies $\langle f(u), f(v)\rangle \in E$ for any substitution f.*

PROOF. The direct implication is easy to prove. Let E be a fully invariant congruence of the term algebra. Put $A = \mathbf{T_X}/E$ and denote by p the canonical projection of $\mathbf{T_X}$ onto A. If h is a homomorphism of $\mathbf{T_X}$ into A, then $h = pf$ for a substitution f; if $\langle u, v\rangle \in E$, then $\langle f(u), f(v)\rangle \in E$, so that $pf(u) = pf(v)$, i.e., $h(u) = h(v)$. We get $A \in \mathbf{Mod}(E)$.

Let $\langle u, v\rangle \in \mathbf{Eq}(\mathbf{Mod}(E))$. Since $A \in \mathbf{Mod}(E)$, $\langle u, v\rangle$ is satisfied in A. Since p is a homomorphism of $\mathbf{T_X}$ into A, by definition we get $p(u) = p(v)$, i.e., $\langle u, v\rangle \in E$. Hence $\mathbf{Eq}(\mathbf{Mod}(E)) \subseteq E$. By Theorem 2.1 we get $E = \mathbf{Eq}(\mathbf{Mod}(E))$ and E is an equational theory. $\square$

2.3. THEOREM. (Birkhoff [35]) *A class of algebras is a variety if and only if it is HSP-closed.*

PROOF. The direct implication is easy to prove. Let C be closed under homomorphic images, subalgebras and direct products and let $A \in \mathbf{Mod}(\mathbf{Eq}(C))$. By Theorem 2.1 it is sufficient to prove $A \in C$.

Since C is closed under subalgebras and direct products, there exists a free algebra B in C over the set A. Denote by $\mathbf{T}$ the algebra of terms over A, by g the unique homomorphism of $\mathbf{T}$ onto B extending $\mathbf{id}_A$, and by h the unique homomorphism of $\mathbf{T}$ onto A extending $\mathbf{id}_A$.

Let $\langle a, b\rangle \in \mathbf{ker}(g)$, i.e., $g(a) = g(b)$. Since $\mathbf{X}$ is infinite and there are only finitely many elements of A occurring in either a or b, there are two terms u, v over the set $\mathbf{X}$ such that the equation $\langle u, v\rangle$ behaves similarly as the pair $\langle a, b\rangle$ in the following sense: $\langle u, v\rangle$ is satisfied in an algebra U if and only if $s(a) = s(b)$ for every homomorphism s of $\mathbf{T}$ into U. If $U \in C$ and s is a homomorphism of $\mathbf{T}$ into U, then $s = fg$ for a homomorphism f of B into U, so that $s(a) = fg(a) = fg(b) = s(b)$. Consequently, $\langle u, v\rangle$ is satisfied in every algebra $U \in C$, and we have $\langle u, v\rangle \in \mathbf{Eq}(C)$. Since $A \in \mathbf{Mod}(\mathbf{Eq}(C))$, it follows that $\langle u, v\rangle$ is satisfied in A. Hence $h(a) = h(b)$.

We have proved that $\mathbf{ker}(g) \subseteq \mathbf{ker}(h)$. But then A is a homomorphic image of B. Since $B \in C$ and C is closed under homomorphic images, we get $A \in C$. $\square$

Although it is not legitimate to speak about lattices of proper classes, the collection of varieties is a 'lattice' in a sense, according to Theorem 2.1. It follows from Theorem 2.1 that the set of equational theories (of the given signature σ) is a complete lattice with respect to inclusion, and that this lattice is antiisomorphic to the 'lattice' of varieties. The lattice of equational theories will be denoted by $\mathbf{L}_\sigma$ (or just $\mathbf{L}$).

The least equational theory of signature σ will be denoted by $\mathbf{id}_\sigma$; it consists of the equations $\langle u, v\rangle$ with $u = v$. The corresponding variety is the variety of all σ-algebras. The largest equational theory of signature σ is the set T_σ^2 of all σ-equations. It will be called the *trivial* equational theory, because it corresponds to the trivial variety of one-element algebras. Both $\mathbf{id}_\sigma$ and T_σ^2 are called *extreme*.

3. Derivations, consequences and bases

Let E be a set of equations. The least equational theory containing E (its existence being clear) will be denoted by $\mathbf{Eq}(E)$ and its elements will be called *consequences* of E. We write $E \vdash \langle u, v\rangle$ if $\langle u, v\rangle$ is a consequence of E.

By a *base* for an equational theory E we mean any subset B of E such that $E = \mathbf{Eq}(B)$; we also say that E is generated by B. An equational theory is called *finitely based* if it has a finite base; it is called *one-based* if it has a base consisting of a single equation.

Both extreme equational theories are one-based: $\langle x, x\rangle$ is a base for $\mathbf{id}_\sigma$, and $\langle x, y\rangle$ is a base for T_σ^2, where x and y are two distinct variables.

An equation $\langle r, s\rangle$ is said to be an *immediate consequence* of an equation $\langle u, v\rangle$ if there exist a substitution f and an address a in r such that $r[a] = f(u)$ and $s = r[a : f(u) \to f(v)]$. (Less formally: if s can be obtained from r by replacing one occurrence of a subterm $f(u)$, for a substitution f, with $f(v)$.)

Let B be a set of equations. By a *derivation* based on B we mean a finite sequence $u_0, \dots, u_k$ ($k \geq 0$) of terms such that for any $i \in \{1, \dots, k\}$, either $\langle u_{i-1}, u_i\rangle$ or $\langle u_i, u_{i-1}\rangle$ is an immediate consequence of an equation from B. By a derivation of an equation $\langle u, v\rangle$ from B we mean a derivation $u_0, \dots, u_k$ based on B, such that $u_0 = u$ and $u_k = v$.

3.1. THEOREM. *We have $B \vdash \langle u, v\rangle$ if and only if there exists a derivation of $\langle u, v\rangle$ from B.*

PROOF. Denote by E the set of the equations $\langle u, v\rangle$ such that there exists a derivation of $\langle u, v\rangle$ from B. Using 2.2, it is easy to see that E is an equational theory, and that E is the least equational theory containing B. $\square$

4. Term operations and polynomials

Let A be an algebra and k be a nonnegative integer, such that $k > 0$ if the signature contains no constants. The direct product A^{A^k} is called the algebra of k-ary operations on A. (Its elements are just the k-ary operations on A.) For $i \in \{1, \dots, k\}$, the k-ary operation e_i on A, defined by $e_i(a_1, \dots, a_k) = a_i$, is called the i-th k-ary *trivial operation* on A. The subalgebra of the algebra of k-ary operations on A generated by the set of k-ary trivial operations on A is called the *algebra of k-ary term operations* of A, and its elements are called the k-ary term operations of A.

Consider the term algebra T over a fixed generating set $\{x_1, \dots, x_k\}$ of k elements. The mapping $x_i \to e_i$ can be uniquely extended to a homomorphism h of T onto the algebra of k-ary term operations of A. If $h(t) = f$, then we say that f is the k-ary term operation of A represented by a term t; this operation f is denoted by t^A.

4.1. THEOREM. *Let A be an algebra and X be a subset of A. An element $a \in A$ belongs to $\mathbf{Sg}(X)$ if and only if there exists a k-ary term operation f of A (for some $k \geq 0$) such that $a = f(a_1, \dots, a_k)$ for some $a_1, \dots, a_k \in X$.*

PROOF. It is easy. □

4.2. THEOREM. *Let A be a nontrivial algebra and let k be a nonnegative integer, such that $k > 0$ if the signature contains no constants. Then the algebra of k-ary term operations of A is a free algebra in $\mathbf{HSP}(A)$ over the k-element set of the trivial k-ary operations on A.*

PROOF. Denote by V the class of all algebras B such that every mapping of the set of trivial k-ary operations on A into B can be extended to a homomorphism of the algebra A^{A^k} into B. It is easy to see that V is a variety, and it remains to check that the algebra A belongs to V. Let f be a mapping of the set of trivial k-ary operations $\{e_1, \dots, e_k\}$ into A. For any k-ary operation g on A put $h(g) = g(f(e_1), \dots, f(e_k))$. Then h is a homomorphism of A^{A^k} into A and h extends the mapping f. □

Let A be an algebra and h be an n-ary operation on A. We say that h *preserves subuniverses* of A if for every subuniverse S of A and any elements $a_1, \dots, a_n \in S$, $h(a_1, \dots, a_n) \in S$. We say that h *preserves endomorphisms* of A if for every endomorphism f of A and any elements $a_1, \dots, a_n \in A$, $f(h(a_1, \dots, a_n)) = h(f(a_1), \dots, f(a_n))$. We say that h *preserves congruences* of A if for every congruence r of A and any $\langle a_1, b_1 \rangle \in r$, $\dots$, $\langle a_n, b_n \rangle \in r$, $\langle h(a_1, \dots, a_n), h(b_1, \dots, b_n) \rangle \in r$.

4.3. THEOREM. *Every term operation of an algebra A preserves subuniverses, endomorphisms and congruences of A.*

PROOF. It is easy. □

An algebra A is said to be *free in itself* over a set S if S is a generating subset of A and every mapping of S into A can be extended to an endomorphism of A. (In other words, A is free in $\{A\}$ over S.)

4.4. THEOREM. *Let A be an algebra free in itself over a set S and n be a positive integer such that $n \leq \mathbf{card}(S)$. Then an n-ary operation on A is a term operation of A if and only if it preserves endomorphisms of A.*

PROOF. The direct implication follows from 4.3. Let h be an n-ary operation of A preserving endomorphisms of A. Take pairwise different elements $x_1, \dots, x_n \in S$. If U is a subuniverse of A and $a_1, \dots, a_n \in U$, then we can take an endomorphism f such that $f(x_i) = a_i$ for all i and $f(x) = a_1$ for $x \in S \smallsetminus \{x_1, \dots, x_n\}$; since the range of f is contained in U, we have $h(a_1, \dots, a_n) = h(f(x_1), \dots, f(x_n)) = f(h(x_1, \dots, x_n)) \in U$. So, the operation h preserves subuniverses. Since $h(x_1, \dots, x_n)$ belongs to the subuniverse generated by $x_1, \dots, x_n$, by 4.1 there exists an n-ary term operation g of A such that $h(x_1, \dots, x_n) = g(x_1, \dots, x_n)$. For any $a_1, \dots, a_n \in A$ we can define an endomorphism f as above; then

$$\begin{aligned} h(a_1, \dots, a_n) &= h(f(x_1), \dots, f(x_n)) = f(h(x_1, \dots, x_n)) \\ &= f(g(x_1, \dots, x_n)) = g(f(x_1), \dots, f(x_n)) = g(a_1, \dots, a_n), \end{aligned}$$

so that $h = g$. □

Let $k \geq 1$. By a k-ary *polynomial* of an algebra A we mean a k-ary operation f on A for which there exist a number $m \geq k$, an m-ary term operation g of A and elements $c_{k+1}, \dots, c_m \in A$ such that $f(x_1, \dots, x_k) = g(x_1, \dots, x_k, c_{k+1}, \dots, c_m)$ for all $x_1, \dots, x_k \in A$. One can easily prove that the set of k-ary polynomials of A is just the subuniverse of A^{A^k} generated by the operations e_i (as above) together with all constant k-ary operations on A.

By a *elementary unary polynomial* of A we mean a mapping f of A into A for which there exist a term $t(x_1, \dots, x_n)$ ($n \geq 1$) with precisely one occurrence of x_1 and elements $c_2, \dots, c_n \in A$ such that $f(a) = t^A(a, c_2, \dots, c_n)$ for all $a \in A$.

4.5. THEOREM. (Mal'cev [54]) *Let A be an algebra and r be a nonempty binary relation on A. Then $\mathbf{Cg}_A(r)$ is the transitive closure of the set of all pairs of the form $\langle f(a), f(b) \rangle$ where f is an elementary unary polynomial of A and $\langle a, b \rangle \in r \cup r^{-1}$.*

PROOF. It is easy. □

An equivalent formulation is this: a pair $\langle a, b \rangle$ belongs to $\mathbf{Cg}(r)$ if and only if there exists a finite sequence $a_0, a_1, \dots, a_n$ such that $a_0 = a$, $a_n = b$ and such that for every $i = 1, \dots, n$ there are a unary polynomial p_i of A and a pair $\langle c_i, d_i \rangle \in r$ with $\{a_{i-1}, a_i\} = \{p_i(c_i), p_i(d_i)\}$. Such a finite sequence $a_0, \dots, a_n$ is called a *Mal'cev chain* from a to b with respect to r.

Two algebras (of possibly different signatures) are said to be *term equivalent* if they have the same term operations of positive arities. They are said to be *polynomially equivalent* if they have the same polynomials.

5. Locally finite and finitely generated varieties

An algebra A is said to be *locally finite* if every finitely generated subuniverse of A is finite. A class of algebras is said to be locally finite if it contains only locally finite algebras.

5.1. Theorem. *A variety V is locally finite if and only if any free algebra in V over a finite set is finite.*

Proof. Every finitely generated algebra in V is a homomorphic image of a free V-algebra over a finite set. □

A variety V is said to be *finitely generated* if $V = \mathbf{HSP}(A)$ for a finite algebra A.

A variety generated by a finite set of finite algebras is finitely generated: if it is generated by $A_1, \dots, A_k$, then it is generated by $A_1 \times \dots \times A_k$.

5.2. Theorem. *Every finitely generated variety is locally finite.*

Proof. It follows from 4.2 and 5.1. □

6. Subdirectly irreducible algebras in varieties

Since every variety is generated (as a variety) by the class of its subdirectly irreducible algebras, it is clear that subdirectly irreducible algebras must play a very important role in the investigation of varieties.

A variety V is called *residually small* if there exists a cardinal number κ such that $|A| < \kappa$ for all subdirectly irreducible algebras of V. Equivalently stated, a variety V is residually small if and only if there exists a set S such that any subdirectly irreducible algebra from V belongs to S. A variety is called *residually large* if it is not residually small. It is called *residually finite* if all its subdirectly irreducible members are finite. It is called *residually very finite* if there exists a positive integer n such that all its subdirectly irreducible members have less than n elements.

6.1. Theorem. *Let V be a locally finite variety containing at least one infinite subdirectly irreducible algebra. Then, for any positive integer n, V contains a finite subdirectly irreducible algebra of cardinality at least n.*

Proof. Let A be an infinite subdirectly irreducible algebra in V. Denote by K the class of finite subdirectly irreducible algebras from V and suppose that the cardinalities of all algebras in K are bounded by some positive integer n. According to 5.3.4, A can be embedded into an ultraproduct of its finitely generated subalgebras; since V is locally finite, it means that A can be embedded into an ultraproduct of its finite subalgebras. Each of these finite subalgebras is isomorphic to a subdirect product of algebras from K. Hence $A \in \mathbf{ISP}_U\mathbf{SP}(K)$. From this it follows by 5.11.1 that $A \in \mathbf{ISPP}_U(K)$. But an ultraproduct of algebras of cardinality at most n is of cardinality at most n, so A can be embedded into a direct product of finite algebras, a contradiction. □

7. Minimal varieties

A variety W is said to *cover* a variety V if V is properly contained in W and there is no variety properly contained in W and properly containing V. A variety is said to be *minimal* if it covers the trivial variety.

7.1. THEOREM. *Let V be a finitely based variety. Then for every variety W, properly containing V, there exists a variety covering V and contained in W.*

PROOF. Formulate this in terms of equational theories (the two lattices are antiisomorphic) and use Zorn's lemma. □

7.2. COROLLARY. *For every nontrivial variety V there exists a minimal variety contained in V.*

7.3. EXAMPLE. (1) The variety of groupoids satisfying $xy \approx x$ is minimal. This follows from the description of the corresponding equational theory E. We have $\langle u, v\rangle \in E$ if and only if the terms u, v have the same leftmost variable.

(2) Similarly, the variety of groupoids satisfying $xy \approx y$ is minimal.

(3) The variety of groupoids satisfying $xy \approx zu$ is minimal. The corresponding equational theory E can be described as follows: $\langle u, v\rangle \in E$ if and only if either $u = v$ or neither u nor v is a variable.

(4) The variety of semilattices is minimal. The corresponding equational theory E can be described as follows: $\langle u, v\rangle \in E$ if and only if $\mathbf{S}(u) = \mathbf{S}(v)$. ($\mathbf{S}(u)$ is the set of variables occurring in u.)

(5) For every prime number p, the variety of commutative semigroups satisfying $x^p y \approx yx^p \approx y$ (these are commutative groups satisfying $x^p = 1$) is minimal. The corresponding equational theory E can be described as follows: $\langle u, v\rangle \in E$ if and only if for every variable x, the number of occurrences of x in u is congruent to the number of occurrences of x in v modulo p.

All these are examples of minimal varieties of semigroups. It can be proved that there are no other minimal varieties of semigroups. This collection is countably infinite.

7.4. EXAMPLE. The variety of Boolean algebras is minimal. The variety of distributive lattices is minimal. Since the two-element lattice belongs to every nontrivial variety of lattices, the variety of distributive lattices is the only minimal variety of lattices.

7.5. THEOREM. *There are 2^ω minimal varieties of commutative groupoids.*

PROOF. Define terms $t_1, t_2, \dots$ by $t_1 = xx \cdot x$ and $t_{n+1} = xx \cdot t_n$. Denote by t'_n the term $f(t_n)$, where f is the substitution sending x to y. For every infinite sequence $e = (e_1, e_2, \dots)$ of elements of $\{0, 1\}$ denote by V_e the variety of commutative groupoids satisfying

$$\begin{aligned}
&xx \approx yy,\\
&xt_n \approx x \text{ for all } n \text{ with } e_n = 0,\\
&xt_n \approx yt'_n \text{ for all } n \text{ with } e_n = 1.
\end{aligned}$$

V_e is nontrivial for any e, because it contains the infinite groupoid with underlying set $\{a_0, a_1, a_2, \dots\}$ and multiplication defined as follows: $a_i a_i = a_0$ for all i; $a_i a_0 = a_0 a_i = a_{i+1}$ for all $i > 0$; if i, j are two distinct positive integers and $k = \min(i, j)$, then in the case $e_{|i-j|} = 0$ put $a_i a_j = a_k$, while in the case

$e_{|i-j|} = 1$ put $a_i a_j = a_0$. According to 7.2, for every e there exists a minimal variety M_e contained in V_e. If $e \neq f$, then $V_e \cap V_f$ is the trivial variety, so that $M_e \neq M_f$. □

7.6. THEOREM. *Every locally finite variety has only finitely many minimal subvarieties.*

PROOF. Let V be a locally finite variety. Every minimal subvariety M of V is uniquely determined by any of its nontrivial algebras; in particular, it is uniquely determined by its two-generated free algebra. But the two-generated free algebra of M is a homomorphic image of the two-generated free algebra of V, and the two-generated free algebra of V has only finitely many nonisomorphic homomorphic images (because it is finite). □

7.7. THEOREM. *For a signature containing at least one symbol of positive arity, the lattice of varieties of that signature has no coatoms.*

PROOF. It is easy. □

8. Regular equations

An equation $\langle u, v\rangle$ is called *regular* if $\mathbf{S}(u) = \mathbf{S}(v)$. An equational theory is called regular if it contains regular equations only. Clearly, the set of all regular equations is an equational theory; it is the largest regular equational theory. In the signature of groupoids, this largest regular equational theory is based on the three equations

(1) $x(yz) \approx (xy)z$,
(2) $xy \approx yx$,
(3) $xx \approx x$

and the corresponding variety is the variety of semilattices.

8.1. THEOREM. *Let E be a nonregular equational theory and $\langle u, v\rangle$ be an arbitrary nonregular equation from E. Denote by E_0 the set of all regular equations from E. Then $E = \mathbf{Eq}(E_0 \cup \{\langle u, v\rangle\})$.*

PROOF. (This theorem belongs to a Russian mathematician, whose name is forgotten and the reference is lost. He proved it around 1950.) It is sufficient to assume that there is a variable $x \in \mathbf{S}(u) \smallsetminus \mathbf{S}(v)$. Let $\langle s, t\rangle$ be a nonregular equation from E. We need to prove that $\langle s, t\rangle$ is a consequence of $E_0 \cup \{\langle u, v\rangle\}$.

Suppose first that σ contains no operation symbols of arity > 1. If there is a variable $y \in \mathbf{S}(s)$, then the equation $s \approx s[y : u[x : y]]$ belongs to E_0, the equation $s[y : u[x : y]] \approx s[y : u[x : t]]$ is a consequence of $\langle u, v\rangle$ and the equation $s[y : u[x : t]] \approx t$ belongs to E_0, so that the equation $s \approx t$ is a consequence of $E_0 \cup \{\langle u, v\rangle\}$. If $\mathbf{S}(s)$ is empty, then $\mathbf{S}(t)$ is nonempty, so that, according to the previous argument, $t \approx s$ is a consequence of $E_0 \cup \{\langle u, v\rangle\}$; but $s \approx t$ is a consequence of $t \approx s$.

Next consider the case when σ contains an at least binary operation symbol F. Let $a = a(x, z_1, \dots, z_n)$ and take two distinct variables y, z different from x.

Put

$$\begin{aligned} b &= u(y, x, \dots, x), \\ a &= Fbx \dots x, \end{aligned}$$

so that $\mathbf{S}(a) = \{x, y\}$ (let us write $a = a(x, y)$) and the nonregular equation $a \approx a(x, z)$ is a consequence of $E_0 \cup \{\langle u, v \rangle\}$. Put $\mathbf{S}(s) \smallsetminus \mathbf{S}(t) = \{x_1, \dots, x_m\}$ and $\mathbf{S}(t) \smallsetminus \mathbf{S}(s) = \{y_1, \dots, y_k\}$. If $m \geq 1$, then the equation

$$s \approx s[x_1 : a(x_1, x_1), \dots, x_m : a(x_m, x_m)]$$

belongs to E_0, the equation

$$s[x_1 : a(x_1, x_1), \dots, x_m : a(x_m, x_m)] \approx s[x_1 : a(x_1, t), \dots, x_m : a(x_m, t)]$$

is a consequence of $a \approx a(x, z)$, the equation

$$s[x_1 : a(x_1, t), \dots, x_m : a(x_m, t)] \approx s[x_1 : a(t, x_1), \dots, x_m : a(t, x_m)]$$

belongs to E_0, the equation

$$s[x_1 : a(t, x_1), \dots, x_m : a(t, x_m)] \approx s[x_1 : a(t, t), \dots, x_m : a(t, t)]$$

is a consequence of $a \approx a(x, z)$ and the equation

$$s[x_1 : a(t, t), \dots, x_m : a(t, t)] \approx t$$

belongs to E_0. In total, $zs \approx t$ is a consequence of $E_0 \cup \{\langle u, v \rangle\}$. If $m = 0$, then $k = 1$ and we can proceed similarly. □

A variety is said to be regular if its equational theory is regular. By the *regularization* of a variety V we mean the variety based on all regular equations satisfied in V. It follows from 8.1 that the regularization of any non-regular variety V is a cover of V in the lattice of varieties of σ-algebras.

In the following we are going to describe a general construction of algebras in the regularization of a given variety of idempotent algebras. (An algebra is said to be *idempotent* if it satisfies $F(x, \dots, x) \approx x$ for every symbol F of σ.) We assume that σ is a signature without constants.

Let S be a join-semilattice, considered as a small category: its objects are elements of S, and morphisms are pairs $\langle s_1, s_2 \rangle \in S^2$ such that $s_1 \leq s_2$. Let H be a functor of H into the category of σ-algebras. That means, for every $s \in S$ there is a σ-algebra $H(s)$ and for every pair $s_1 \leq s_2$ there is a homomorphism $h_{s_1,s_2} : H_{s_1} \to H_{s_2}$ such that $H_{s,s}$ is the identity on $H(s)$ and whenever $s_1 \leq s_2 \leq s_3$ then $H_{s_1,s_3} = H_{s_2,s_3} H_{s_1,s_2}$. Suppose, moreover, that the algebras $H(s)$ ($s \in S$) are pairwise disjoint. Then we can define a σ-algebra A with the underlying set $\bigcup_{s \in S} H(s)$ as follows: if F is an n-ary operation symbol and $a_i \in H(s_i)$ for $i = 1, \dots, n$ then $F_A(a_1, \dots, a_n) = F_{H(s)}(H_{s_1,s}(a_1), \dots, H_{s_n,s}(a_n))$ where $s = s_1 \vee \dots \vee s_n$. The algebra A defined in this way is called the *Płonka sum* of the functor H (or just of the family of algebras $H(s)$, $s \in S$).

By a *partition operation* on a set A we mean a binary operation $\circ$ such that

(1) A is a left normal band with respect to $\circ$, i.e., an idempotent semigroup such that $x \circ y \circ z = x \circ z \circ y$ for all $x, y, z \in A$

(2) $F(x_1, \dots, x_n) \circ y = F(x_1 \circ y, \dots, x_n \circ y)$ and $y \circ F(x_1, \dots, x_n) = y \circ x_1 \circ \dots \circ x_n$ for any n-ary symbol F of σ and any $x_1, \dots, x_n, y \in A$

8.2. THEOREM. *Let σ be a signature without constants.*

Let A be the Płonka sum of a functor H of a join-semilattice S into the category of σ-algebras. Then the binary operation $\circ$ defined on A by $a \circ b = H_{s_1, s_1 \vee s_2}(a)$ for $a \in H(s_1)$ and $b \in H(s_2)$, is a partition operation on A.

Conversely, let $\circ$ be a partition operation on a σ-algebra A. Define an equivalence r on A by $\langle a, b \rangle \in r$ if and only if $a \circ b = a$ and $b \circ a = b$, so that r is a congruence of A all the blocks of which are subalgebras of A. Put $S = A/r$ and define a join-semilattice ordering $\leq$ on S by $a/r \leq b/r$ if $b \circ a = b$. Then A is the Płonka sum of the functor H of S into the category of σ-algebras defined by $H(a/r) = a/r$ and $H_{a/r, b/r}(x) = x \circ b$ whenever $a/r \leq b/r$ and $x \in a/r$.

PROOF. Let A be the Płonka sum. It is easy to see that A is a left normal band with respect to $\circ$. For $x_i \in H(s_i)$, $y \in H(t)$ and $s = s_1 \vee \dots \vee s_n$ we have

$$\begin{aligned} F_A(x_1, \dots, x_n) \circ y &= F_{H(s)}(H_{s_1,s}(x_1), \dots, H_{s_n,s}(x_n)) \circ y \\ &= H_{s, s \vee t}(F_{H(s)}(H_{s_1,s}(x_1), \dots, H_{s_n,s}(x_n)) \\ &= F_{H(s \vee t)}(H_{s_1, s \vee t}(x_1), \dots, H_{s_n, s \vee t}(x_n)) \end{aligned}$$

and

$$\begin{aligned} F_A(x_1 \circ y, \dots, x_n \circ y) &= F_A(H_{s_1, s_1 \vee t}(x_1), \dots, H_{s_n, s_n \vee t}(x_n)) \\ &= F_{H(s \vee t)}(H_{s_1, s \vee t}(x_1), \dots, H_{s_n, s \vee t}(x_n)). \end{aligned}$$

Similarly one can prove $y \circ F(x_1, \dots, x_n) = y \circ x_1 \circ \dots x_n$.

Now let $\circ$ be a partition operation on A and let S and H be defined as above. It is easy to see that $\circ$ is a congruence of the left normal band. It is also a congruence of the σ-algebra A: if $\langle a_i, b_i \rangle \in r$ for $i = 1, \dots, n$ then

$$\begin{aligned} F(a_1, \dots, a_n) \circ F(b_1, \dots, b_n) &= F(a_1, \dots, a_n) \circ F(a_1, \dots, a_n) \circ b_1 \circ \dots \circ b_n \\ &= F(a_1, \dots, a_n) \circ a_1 \circ \dots \circ a_n \circ b_1 \circ \dots \circ b_n \\ &= F(a_1, \dots, a_n) \circ (a_1 \circ b_1) \circ \dots \circ (a_n \circ b_n) \\ &= F(a_1, \dots, a_n) \circ a_1 \circ \dots \circ a_n \\ &= F(a_1, \dots, a_n) \circ F(a_1, \dots, a_n) = F(a_1, \dots, a_n) \end{aligned}$$

and similarly $F(b_1, \dots, b_n) \circ F(a_1, \dots, a_n) = F(b_1, \dots, b_n)$. Clearly, $S = A/r$ is a join-semilattice with respect to $\leq$; we have $a/r \vee b/r = (a \circ b)/r$. One can easily check that $\langle F(a_1, \dots, a_n), a_1 \circ \dots \circ a_n \rangle \in r$, that the blocks of r are subalgebras of A and that H is a correctly defined functor. Denote by A^* the Płonka sum. For $a_1, \dots, a_n$ we have (where $a = a_1 \circ \dots \circ a_n$ and

$s = a_1/r \vee \cdots \vee a_n/r = a/r$)

$$\begin{aligned} F_{A^*}(a_1, \dots, a_n) &= F_s(H_{a_1/r,s}(a_1), \dots, H_{a_n/r,s}(a_n)) \\ &= F_s(a_1 \circ a, \dots, a_n \circ a) \\ &= F_A(a_1, \dots, a_n) \circ a = F_A(a_1, \dots, a_n). \end{aligned}$$

Consequently, $A^* = A$. □

8.3. THEOREM. *Let σ be a signature without constants, S be a nontrivial join-semilattice, H be a functor of S into the category of σ-algebras and A be its Płonka sum. An equation is satisfied in A if and only if it is regular and is satisfied in all the algebras $H(s)$, $s \in S$.*

PROOF. Let $\langle u, v \rangle$ be satisfied in A. There exist elements $s, t \in S$ with $s < t$. For a variable $x \in \mathbf{S}(u)$ let f_x be the homomorphism of the term algebra into A sending x to t and every other variable to s. We have $f_x(u) \in H(t)$, so that $f_x(v) \in H(t)$ and consequently $x \in \mathbf{S}(v)$. Similarly $\mathbf{S}(v) \subseteq \mathbf{S}(u)$, so that $\langle u, v \rangle$ is regular. Clearly, $\langle u, v \rangle$ is satisfied in all the subalgebras $H(s)$ of A.

Conversely, let $\langle u, v \rangle$ be a regular equation satisfied in all $H(s)$. Denote by $x_1, \dots, x_n$ the elements of $\mathbf{S}(u) = \mathbf{S}(v)$. Let f be a homomorphism of the term algebra into A. For $i = 1, \dots, n$ denote by s_i the element of S with $f(x_i) \in H(s_i)$. Put $s = s_1 \vee \cdots \vee s_n$ and define a homomorphism g of the term algebra into $H(s)$ by $g(x_i) = H_{s_i,s} f(x_i)$. Then $f(u) = g(u)$, $f(v) = g(v)$ and $g(u) = g(v)$, so that $f(u) = f(v)$. □

8.4. THEOREM. *Let V be a non-regular variety of idempotent algebras of a signature σ without constants and containing at least one at least binary symbol; let W be the regularization of V. Then every algebra from W is a Płonka sum of some algebras from V.*

PROOF. Let $A \in W$. There is an equation $\langle u, v \rangle$ satisfied in V such that $y \in \mathbf{S}(u) \smallsetminus \mathbf{S}(v)$ for some variable y; since σ contains an at least binary symbol, we can also assume that y is not the only variable occurring in u. Take a variable $x \neq y$ and denote by $w = w(x, y)$ the term obtained from u by the substitution sending y to y and every other variable to x. Then $w(x, y) \approx x$ is satisfied in V and hence $w(x, y)$ represents a partition operation, which will be denoted by $\circ$, on every algebra from V. The equations in the definition of a partition operation were all regular. So, $\circ$ is a partition operation on every algebra of W; in particular, $\circ$ is a partition operation on A. By 8.2, A is the Płonka sum of its subalgebras that are blocks of the congruence r (where $\langle a, b \rangle \in r$ means $a \circ b = a$ and $b \circ a = b$) and satisfy $w(x, y) \approx x$. Let $\langle p, q \rangle$ be an arbitrary equation satisfied in V. Put $\mathbf{S}(p) \smallsetminus \mathbf{S}(q) = \{x_1, \dots, x_n\}$ and $\mathbf{S}(q) \smallsetminus \mathbf{S}(p) = \{y_1, \dots, y_m\}$. The equation $p \circ y_1 \circ \cdots \circ y_m \approx q \circ x_1 \circ \cdots \circ x_n$ is regular and satisfied in V, so that it is satisfied in W and consequently in A. It follows that the blocks of r satisfy $p \approx q$. Consequently, all these blocks belong to V. □

9. Poor signatures

A signature σ is called *poor* if it contains nothing else than at most one unary operation symbol, and a set of constants.

9.1. THEOREM. *Let σ contain no other symbols than constants. Then $\mathbf{L}_\sigma$ is isomorphic to the partition lattice of σ with a new largest element added.*

PROOF. In this case, the nontrivial equational theories are just the equivalences on σ. □

9.2. THEOREM. *Let σ be a finite poor signature. Then every equational theory of signature σ is finitely based.*

PROOF. It is sufficient to consider the case when σ actually contains a unary symbol F. Let E be an equational theory. Let x and y be two distinct variables.

If E contains the equation $\langle F^k(x), F^k(y)\rangle$ for some nonnegative integer k, let k be the least such integer and put $B_1 = \{\langle F^k(x), F^k(y)\rangle\}$; otherwise, put $B_1 = \emptyset$.

If E contains the equation $\langle F^n(x), F^m(x)\rangle$ for some pair $\langle n, m\rangle$ of integers with $0 \leq n < m$, let $\langle n, m\rangle$ be the least such pair with respect to the lexicographic ordering of ordered pairs of nonnegative integers, and put $B_2 = \{\langle F^n(x), F^m(x)\rangle\}$; otherwise, put $B_2 = \emptyset$.

Let $c, d \in \sigma$ be two constants. If E contains the equation $\langle F^i(c), F^j(d)\rangle$ for some pair $\langle i, j\rangle$ of nonnegative integers such that $i < j$ if $c = d$, let $\langle i, j\rangle$ be the least such pair in the lexicographic ordering and put $B_{c,d} = \{\langle F^i(c), F^j(d)\rangle\}$; otherwise, put $B_{c,d} = \emptyset$.

It is not difficult to prove that the set $B_1 \cup B_2 \cup \bigcup_{c,d} B_{c,d}$ is a finite base for E. □

9.3. EXAMPLE. Let $\sigma = \{F\}$ for a unary symbol F. The lattice $\mathbf{L}_\sigma$, with its least element removed, is isomorphic to the dual of the direct product of two lattices: the lattice of nonnegative integers with respect to the usual ordering, and the lattice of nonnegative integers with respect to the ordering given by the divisibility relation on positive integers and setting 0 to be the least element. Consequently, the lattice is countably infinite and distributive. For $d > 0$, the equational theory corresponding to $\langle n, d\rangle$ is based on $\langle F^n(x), F^{n+d}(x)\rangle$. For $n = 0$, the equational theory corresponding to $\langle n, d\rangle$ is based on $\langle F^n(x), F^n(y)\rangle$. This can be also left to the reader as an easy excercise.

A complete description of the lattice $\mathbf{L}_\sigma$ for any poor signature σ is given in J. Ježek [69]. It follows from the description that if $\sigma = \{F\} \cup C$ for a unary symbol F and a set of constants C, then $\mathbf{L}_\sigma$ is distributive for $|C| \leq 1$ and nonmodular for $|C| \geq 2$.

10. Equivalent varieties

10.1. THEOREM. *Let ε be an equivalence between two varieties of algebras K and L. Then:*

(1) *For $A, B \in K$, A is a subalgebra of B if and only if $\varepsilon(A)$ is a subalgebra of $\varepsilon(B)$;*
(2) *For $A \in K$, a nonempty subset of A is a subuniverse of A if and only if it is a subuniverse of $\varepsilon(B)$;*
(3) *For a family H of algebras from K, $\varepsilon(\Pi H) = \Pi\varepsilon(H)$.*
(4) *For $A \in K$, a subset of $A \times A$ is a congruence of A if and only if it is a congruence of $\varepsilon(A)$.*
(5) *For $A \in K$ and a nonempty subset S of A, A is K-free over S if and only if $\varepsilon(A)$ is L-free over S.*

Proof. (1) and (2) follow from 3.3.1. (3) follows from 3.5.1. (4) follows-from 3.5.2, and (5) is clear. □

10.2. Theorem. *Let A be an algebra of signature σ and B be an algebra of signature τ with the same underlying set, such that the two algebras have the same term operations of positive arities. Then the varieties* $\mathbf{HSP}(A)$ *and* $\mathbf{HSP}(B)$ *are equivalent.*

Proof. Put $K = \mathbf{HSP}(A)$ and $L = \mathbf{HSP}(B)$. For a nonempty set S and integers $1 \le i \le n$ denote by $e_{i,n,S}$ the i-th trivial operation of arity n on S. For any algebra C denote by $H_n(A)$ the algebra of n-ary term operations of C. For an algebra $C \in K$ denote by $\varphi_{n,C}$ the unique homomorphism of $H_n(A)$ into $H_n(C)$ such that $\varphi_{n,C}(e_{i,n,A}) = e_{i,n,A}$ for all $i = 1, \dots, n$; its existence follows from 4.2. For $C \in K$ define an algebra $\varepsilon(C)$ of signature τ, with the same underlying set, by $F_{\varepsilon(C)} = \varphi_{n,C}(F_B)$ for any n-ary symbol F of τ.

For every algebra $D \in L$ we can define homomorphisms $\psi_{n,D} : H_n(B) \to H_n(D)$, and we can define a mapping ε' of L into the class of algebra of signature σ in a similar way. Clearly, $\varepsilon(A) = B$ and $\varepsilon'(B) = A$.

Let $n \ge 1$. One can easily see that the set of the n-ary term operations h of A such that

$$f(\varphi_{n,C}(h)(a_1, \dots, a_n)) = \varphi_{n,D}(h)(f(a_1), \dots, f(a_n))$$

for every homomorphism $f : C \to D$ and every n-tuple $a_1, \dots, a_n$ of elements of C, where $C, D \in K$, is a subuniverse of $H_n(A)$ containing the trivial operations, so that it equals $H_n(A)$. From this it follows that if f is a homomorphism of an algebra $C \in K$ into an algebra $D \in K$, then f is also a homomorphism of $\varepsilon(C)$ into $\varepsilon(D)$. Similarly, the mapping ε' preserves homomorphisms in the same sense.

It follows from 3.3.1 and 3.5.1 that ε preserves subalgebras and products. In particular, $\varepsilon(H_n(A)) = H_n(B)$ for all n.

Let $C \in K$. If D is a finitely generated subalgebra of $\varepsilon(C)$, then we can take a positive integer n and a homomorphism of the free algebra $H_n(A)$ into C, mapping the generators of $H_n(A)$ onto a generating subset of D. Then f is a homomorphism of $H_n(B)$ onto D, so that $D \in L$. Since every finitely generated subalgebra of $\varepsilon(C)$ belongs to L, we get $\varepsilon(C) \in L$. Hence ε is a mapping of K into L. Similarly, ε' is a mapping of L into K.

Let $C \in K$; let F be an n-ary symbil of σ and $a_1, \dots, a_n$ be elements of C. There is a homomorphism $f : H_n(A) \to C$ with $f(e_i) = a_i$ for all i (here $e_i = e_{i,n,A}$). Since f is also a homomorphism of $H_n(A) = \varepsilon'\varepsilon(H_n(A))$ into $\varepsilon'\varepsilon(C)$, we have

$$\begin{aligned} F_{\varepsilon'\varepsilon(C)}(a_1, \dots, a_n) &= F_{\varepsilon'\varepsilon(C)}(f(e_1), \dots, f(e_n)) \\ &= f(F_{H_n(A)}(e_1, \dots, e_n)) = F_C(f(e_1), \dots, f(e_n)) \\ &= F_C(a_1, \dots, a_n). \end{aligned}$$

Hence $\varepsilon'\varepsilon(C) = C$. Similarly, $\varepsilon\varepsilon'$ is an identity on L. □

10.3. THEOREM. *Let K be a variety of algebras of signature σ and L be a variety of algebras of signature τ. Let $A \in K$ and $B \in L$ be two algebras with the same underlying sets, such that A is K-free and B is L-free over the same infinite subset. Suppose that A and B have the same endomorphisms. Then K and L are equivalent varieties.*

PROOF. It follows from 4.4 and 10.2. □

11. Independent varieties

Let $V_1, \dots, V_n$ $(n \geq 1)$ be varieties of σ-algebras. We say that $V_1, \dots, V_n$ are *independent* if there exists a term $t(x_1, \dots, x_n)$ such that for $i = 1, \dots, n$, V_i satisfies $t(x_1, \dots, x_n) \approx x_i$.

11.1. THEOREM. *Let $V_1, \dots, V_n$ be varieties and V be the variety generated by $V_1 \cup \dots \cup V_n$. Then $V_1, \dots, V_n$ are independent if and only if the following two conditions are satisfied:*

(1) *Every algebra in V is isomorphic to a product $A_1 \times \dots \times A_n$ for some algebras $A_i \in V_i$*
(2) *If $A = A_1 \times \dots \times A_n$ and $B = B_1 \times \dots \times B_n$ where $A_i, B_i \in V_i$ then a mapping $f : A \to B$ is a homomorphism if and only if there are homomorphisms $f_i : A_i \to B_i$ $(i = 1, \dots, n)$ such that $f(\langle x_1, \dots, x_n \rangle) = \langle f_1(x_1), \dots, f_n(x_n) \rangle$ for all $\langle x_1, \dots, x_n \rangle \in A$*

PROOF. Let $V_1, \dots, V_n$ be independent with respect to a term $t(x_1, \dots, x_n)$.

Claim 1. If $A_i \in V_i$ and B is a subalgebra of $A_1 \times \dots \times A_n$ then $B = B_1 \times \dots \times B_n$ for some subalgebras B_i of A_i. Denote by B_i the image of B under the i-th projection. Clearly, B is contained in $B_1 \times \dots \times B_n$. If $\langle b_1, \dots, b_n \rangle \in B_1 \times \dots \times B_n$ then for every i there exists an element $\langle c_{i,1}, \dots, c_{i,n} \rangle$ of B with $b_i = c_{i,i}$; we have $\langle b_1, \dots, b_n \rangle = t(\langle c_{1,1}, \dots, c_{1,n} \rangle, \dots, \langle c_{n,1}, \dots, c_{n,n} \rangle) \in B$.

Claim 2. If $A_i \in V_i$ and r is a congruence of $A_1 \times \dots \times A_n$ then there exist congruences r_i of A_i such that $\langle \langle x_1, \dots, x_n \rangle, \langle y_1, \dots, y_n \rangle \rangle \in r$ if and only if $\langle x_i, y_i \rangle \in r_i$ for all i. This follows from Claim 1, since r can be viewed as a subalgebra of $A_1^2 \times \dots \times A_n^2$.

It follows from these two claims that the class of the algebras isomorphic to $A_1 \times \dots \times A_n$ for some $A_i \in V_i$ is HSP-closed, so that it equals V. We have proved (1), and (2) is also a consequence of Claim 1 since a homomorphism of $A_1 \times \cdots \times A_n$ into $B_1 \times \cdots \times B_n$ is a subalgebra of $(B_1 \times A_1) \times \cdots \times (A_n \times B_n)$.

Now suppose that (1) and (2) are satisfied. For $i = 1, \dots, n$ let A_i be a V_i-free algebra over $\{x_1, \dots, x_n\}$. (We can assume that all the varieties are nontrivial; if some of them are trivial, just delete them.) It is easy to check that the algebra $A = A_1 \times \cdots \times A_n$ is generated by (and V-free over) $\{\langle x_1, \dots, x_1\rangle, \dots, \langle x_n, \dots, x_n\rangle\}$. Consequently, there exists a term $t(x_1, \dots, x_n)$ such that $\langle x_1, \dots, x_n\rangle = t(\langle x_1, \dots, x_1\rangle, \dots, \langle x_n, \dots, x_n\rangle)$. Clearly, V_i satisfies $t(x_1, \dots, x_n) \approx x_i$. □

11.2. EXAMPLE. The variety V of *rectangular bands*, i.e., idempotent semigroups satisfying $xyz \approx xz$, is generated by its two independent subvarieties: the subvariety determined by $xy \approx x$ and the subvariety determined by $xy \approx y$. It follows that for every rectangular band A there exist two nonempty sets B, C such that A is isomorphic to the rectangular band with the underlying set $B \times C$ and multiplication $\langle b_1, c_1\rangle\langle b_2, c_2\rangle = \langle b_1, c_2\rangle$.

Let A be an algebra of signature σ and n be a positive integer. Let τ be an extension of σ by two operation symbols: an n-ary symbol D and a unary symbol U. We define a τ-algebra $A^{[n]}$ in this way: its reduct to σ is the algebra A^n;

$$D(\langle a_{1,1}, \dots, a_{1,n}\rangle, \dots, \langle a_{n,1}, \dots, a_{n,n}\rangle) = \langle a_{1,1}, a_{2,2}, \dots, a_{n,n}\rangle;$$
$$U(\langle a_1, \dots, a_n\rangle) = \langle a_2, \dots, a_n, a_1\rangle.$$

For a variety V of σ-algebras we denote by $V^{[n]}$ the class of the algebras $A^{[n]}$ with $A \in V$.

11.3. THEOREM. *Let V be a variety of σ-algebras. Then $V^{[n]}$ is a variety. It is generated by its subvarieties V_i $(i = 1, \dots, n)$ determined by $D(x_1, \dots, x_n) \approx x_i$. The varieties $V_1, \dots, V_n$ are independent and each of them is equivalent with V.*

PROOF. One can check that any homomorphism $f : A_1 \times \cdots \times A_n \to B_1 \times \cdots \times B_n$ (for $A_i, B_i \in V$) is of the form $f(\langle a_1, \dots, a_n\rangle) = \langle f_1(x_1), \dots, f_n(x_n)\rangle$ for some homomorphisms $f_i : A_i \to B_i$. Then it is easy to conclude that $V^{[n]}$ is a variety. □

12. The existence of covers

12.1. THEOREM. (Trakhtman [74]) *Let A, B be two equational theories such that $A \subset B$. Denote by S the set of the terms u for which there exists a term v with $\langle u, v\rangle \in B \setminus A$. Suppose that there exists a term $t \in S$ such that whenever $\langle t, u\rangle \in A$ and $v < u$ then $v \notin S$. Then there exists an equational theory containing A and covered by B in the lattice of equational theories.*

PROOF. Let us take one term t as above. Denote by Q the set of the terms u for which there exists a term v with $u \sim v$ and $\langle t, v\rangle \in A$. Clearly, Q is the set of the terms u for which there exists a term w with $t \sim w$ and $\langle w, u\rangle \in A$. The following facts are easy to see:

(1) $t \in Q$ and $Q \subseteq S$
(2) if $u \in Q$ and $u \sim v$ then $v \in Q$
(3) if $u \in Q$ and $\langle u, v\rangle \in A$ then $v \in Q$
(4) if $u, v \in Q$ then there exists a term w with $\langle u, w\rangle \in A$ and $v \sim w$
(5) if $u \in S$ and $f(u) \in Q$ for a substitution f then $u \sim f(u)$

Define a binary relation L on the algebra of terms by $\langle u, v\rangle \in L$ if and only if either $\langle u, v\rangle \in A$ or $u, v \notin Q$ and $\langle u, v\rangle \in B$. Clearly, $A \subseteq L \subset B$. It is easy to check that L is an equivalence. Let us prove that L is a congruence of the algebra of terms. Let F be an n-ary operation symbol in the signature and let $\langle u_i, v_i\rangle \in L$ for $i = 1, \dots, n$. If $\langle u_i, v_i\rangle \in A$ for all i then $\langle F(u_1, \dots, u_n), F(v_1, \dots, v_n)\rangle \in A \subseteq L$. Let $\langle u_j, v_j\rangle \notin A$ for some j. Then $u_j, v_j \in S$. Since $u_j < F(u_1, \dots, u_n)$ and $v_j < F(v_1, \dots, v_n)$, we have $F(u_1, \dots, u_n) \notin Q$ and $F(v_1, \dots, v_n) \notin Q$. Since $\langle F(u_1, \dots, u_n), F(v_1, \dots, v_n)\rangle \in B$, we get $\langle F(u_1, \dots, u_n), F(v_1, \dots, v_n)\rangle \in L$.

Let us prove that L is fully invariant. Let $\langle u, v\rangle \in L$ and let f be a substitution. Suppose $\langle f(u), f(v)\rangle \notin L$. We have $\langle u, v\rangle \notin A$, since in the opposite case we would have $\langle f(u), f(v)\rangle \in A \subseteq L$. Since $\langle u, v\rangle \in B$, we get $u, v \in S$. If $u \in Q$ then $\langle u, v\rangle \in L$ implies $\langle u, v\rangle \in A$, a contradiction. Hence $u \notin Q$ and similarly $v \notin Q$. By (5) we get $f(u), f(v) \notin Q$. Since $\langle f(u), f(v)\rangle \in B$, we get $\langle f(u), f(v)\rangle \in L$.

We have proved that L is an equational theory. Let C be an equational theory such that $L \subseteq C \subset B$. We are going to prove that if $u \notin Q$ and $v \in Q$ then $\langle u, v\rangle \notin C$. Suppose $\langle u, v\rangle \in C$. Since $C \neq B$, there exists a pair $\langle a, b\rangle \in B \smallsetminus C$. At least one of the terms a, b belongs to Q, since otherwise we would have $\langle a, b\rangle \in L \subseteq C$. We can suppose without loss of generality that $a \in Q$. There exists a substitution f with $\langle f(a), v\rangle \in A$. We have $\langle u, f(a)\rangle \in C$ and $\langle f(a), f(b)\rangle \in B$, so that $\langle u, f(b)\rangle \in B$. If $f(b) \notin Q$ then $\langle u, f(b)\rangle \in L$ and then $\langle u, f(b)\rangle \in C$, so that $\langle f(a), f(b)\rangle \in C$ and consequently $\langle a, b\rangle \in C$, a contradiction. Hence $f(b) \in Q$. There exists a substitution g such that $\langle g(v), f(b)\rangle \in A$. We have $\langle g(u), g(v)\rangle \in C$; since

$$u \;\; C \;\; v \;\; A \;\; f(a) \;\; B \;\; f(b) \;\; A \;\; g(v) \;\; C \;\; g(u)$$

and $u, g(u) \notin Q$, we get $\langle u, g(u)\rangle \in C$. Hence

$$f(a) \;\; A \;\; v \;\; C \;\; u \;\; C \;\; g(u) \;\; C \;\; g(v) \;\; A \;\; f(b),$$

so that $\langle a, b\rangle \in C$, a contradiction.

Again, let C be an equational theory with $L \subseteq C \subset B$. We are going to prove that if $u, v \in Q$ and $\langle u, v\rangle \in C$ then $\mathbf{S}(u) = \mathbf{S}(v)$. Suppose, on the contraty, that (for example) $x \in \mathbf{S}(v) \smallsetminus \mathbf{S}(u)$ for a variable x. If we substite w in v for x where w is an arbitrary term with $w > x$ (such a term w exists if the signature is nonempty; if it is empty, then everything is clear) we get a term v' such that $\langle u, v'\rangle \in C$ and $v < v'$, so that $v' \notin Q$, a contradiction.

Denote by R the set of the terms that are similar to t and contain the same variables as t; so, R is finite. We are going to prove that if C, D are two equational theories containing L, properly contained in B and coinciding on $R \times R$ then $C = D$. Let u, v be two terms. If $u, v \notin Q$ then $\langle u, v\rangle \in C$ if and only if $\langle u, v\rangle \in B$ if and only if $\langle u, v\rangle \in D$. If one of the terms u, v belongs to and the other one does not belong to Q then $\langle u, v\rangle$ belongs to neither C nor D. Let $u, v \in Q$. Since u, v contain the same variables, there exist two automorphisms f, g of the algebra of terms such that f maps $\mathbf{S}(t)$ onto itself, $\langle g(t), u\rangle \in A$ and $\langle g(f(t)), v\rangle \in A$. We have $\langle u, v\rangle \in C$ if and only if $\langle t, f(t)\rangle \in C$ and $\langle u, v\rangle \in D$ if and only if $\langle t, f(t)\rangle \in D$. But $t, f(t) \in R$, so that $\langle t, f(t)\rangle \in C$ if and only if $\langle t, f(t)\rangle \in D$.

It follows that there are only finitely many equational theories C such that $L \subseteq C \subset B$. Among them, there must be a maximal one. $\square$

An equation $\langle u, v\rangle$ is said to be *balanced* if it satisfies the following two conditions:

(1) for every variable x, the number of occurrences of x in u is the same as the number of occurrences of x in v
(2) for every at most unary operation symbol F of σ, the number of occurrences of F in u is the same as the number of occurrences of F in v

Clearly, the set of balanced equations is an equational theory. An equational theory is said to be balanced if it contains only balanced equations. A variety is said to be balanced if its equational theory is balanced.

12.2. THEOREM. *Let K be a balanced variety and L be a proper subvariety of K. Then L has a cover in the lattice of subvarieties of K.*

PROOF. Let A and B be the equational theories of K and L, respectively. For every term u denote by n_u the sum of the number of occurrences of variables in u and the number of occurrences of at most unary operation symbols in u. Let S be as in 12.1 and denote by n the minimum of the numbers n_u for $u \in S$. Let $t \in S$ be a term such that $n_t = n$ and whenever $t' \in S$ and $n_{t'} = n$ then $\mathbf{card}(\mathbf{S}(t')) \leq \mathbf{card}(\mathbf{S}(t))$. Clearly, the assumptions of 12.1 are satisfied with respect to this term t. $\square$

12.3. COROLLARY. *Every variety different from the variety of all σ-algebras has a cover in the lattice of varieties of σ-algebras.*

CHAPTER 7

MAL'CEV TYPE THEOREMS

1. Permutable congruences

The composition $r \circ s$ of two binary relations r, s is defined as follows: $\langle a,b\rangle \in r \circ s$ iff there is an element c with $\langle a,c\rangle \in r$ and $\langle c,b\rangle \in s$. If r, s are two equivalences on a given set, then $r \circ s$ is not necessarily an equivalence.

1.1. THEOREM. *Let r, s be two equivalences on a set A such that $r \circ s = s \circ r$. Then $r \circ s$ is an equivalence on A; it is just the join of r, s in the lattice of equivalences on A.*

PROOF. It is easy. $\square$

An algebra is said to have *permutable congruences* if $r \circ s = s \circ r$ for any pair r, s of congruences of A. A variety V is said to have permutable congruences (or to be congruence permutable) if every algebra in V has permutable congruences.

1.2. THEOREM. *Let A be an algebra with permutable congruences. Then the congruence lattice of A is modular.*

PROOF. Let r, s, t be three congruences of A such that $r \subseteq t$. In order to prove the modularity of $\mathbf{Con}(A)$, we need to show that $(r \vee s) \wedge t = r \vee (s \wedge t)$. It is sufficient to prove $(r \vee s) \wedge t \subseteq r \vee (s \wedge t)$, since the converse inclusion is true in any lattice. By 1.1, this translates to $(r \circ s) \cap t \subseteq r \circ (s \cap t)$.

Let $\langle a,b\rangle \in (r \circ s) \cap t$. We have $\langle a,b\rangle \in t$, $\langle a,c\rangle \in r$ and $\langle c,b\rangle \in s$ for some element c. Since $r \subseteq t$, we have $\langle a,c\rangle \in t$. Hence $\langle b,c\rangle \in t$ by transitivity, and we get $\langle b,c\rangle \in s \cap t$. Together with $\langle a,c\rangle \in r$, this gives $\langle a,b\rangle \in r \circ (s \cap t)$. $\square$

1.3. LEMMA. *Let V be a variety and F be a free algebra in V over a finite set Y of variables; let $x_1, \dots, x_k, y_1, \dots, y_k \in Y$, where $y_1, \dots, y_k$ are pairwise different. Denote by h the homomorphism of $\mathbf{T}_X$ onto F extending the identity on X. Let $u, v \in \mathbf{T}_X$ be two terms not containing any of the variables $y_1, \dots, y_k$. If $\langle h(u), h(v)\rangle$ belongs to the congruence of F generated by $\{\langle x_1, y_1\rangle, \dots, \langle x_k, y_k\rangle\}$, then the equation $\langle u, v\rangle$ is satisfied in V.*

PROOF. Let $A \in V$ and $p : \mathbf{T}_X \to A$ be a homomorphism. Since F is free, there exists a homomorphism $q : F \to A$ such that $p = qh$. Denote by f the endomorphism of $\mathbf{T}_X$ such that $f(y_i) = x_i$ and $f(z) = z$ for $z \in X \setminus \{y_1, \dots, y_k\}$. Since u, v do not contain $y_1, \dots, y_k$, we have $f(u) = u$ and $f(v) = v$. Denote by g the endomorphism of F such that $g(y_i) = x_i$ and

$g(z) = z$ for $z \in X \setminus \{y_1, \dots, y_k\}$. We have $hf = gh$, since these two homomorphisms coincide on the generating subset X of $\mathbf{T}_X$. Since $\langle h(u), h(v)\rangle \in \mathbf{Cg}_F\{\langle x_1, y_1\rangle, \dots, \langle x_k, y_k\rangle\} \subseteq \mathbf{ker}(g)$, we have $gh(u) = gh(v)$. Hence

$$p(u) = pf(u) = qhf(u) = qgh(u) = qgh(v) = qhf(v) = pf(v) = p(v).$$

$\square$

1.4. THEOREM. (Mal'cev [54]) *Let V be a variety. Then V is congruence permutable if and only if there exists a term $t = t(x, y, z)$ in three variables x, y, z such that V satisfies*

$$t(x, y, y) \approx x \quad \text{and} \quad t(x, x, y) \approx y.$$

A nontrivial variety is congruence permutable if and only if its free algebra of rank 3 has permutable congruences.

PROOF. Let V be a nontrivial congruence permutable variety. Denote by F the free V-algebra over x, y, z, by T the algebra of terms over x, y, z and by h the homomorphism of T onto F acting as the identity on x, y, z. Denote by r the congruence of F generated by $\langle x, y\rangle$ and by s the congruence of F generated by $\langle y, z\rangle$. Since $\langle x, z\rangle \in r \circ s = s \circ r$, there is an element $d \in F$ with $\langle x, d\rangle \in s$ and $\langle d, y\rangle \in r$. Take an arbitrary term $t \in T$ such that $h(t) = d$. We have $\langle h(t(x, y, y)), h(t)\rangle \in s$ and $\langle h(t), h(x)\rangle \in s$, so that $\langle h((t(x, y, y)), h(x)\rangle \in \mathbf{Cg}(y, z)$. It follows by 1.3 that the equation $t(x, y, y) \approx x$ is satisfied in V. Similarly, $t(x, x, y) \approx y$ is satisfied.

Conversely, suppose that there exists a term t with the property stated above. Let $A \in V$, let r, s be two congruences of A and let $\langle a, b\rangle \in r \circ s$. In order to prove $r \circ s \subseteq s \circ r$, we need to show that $\langle a, b\rangle \in s \circ r$. There is an element c with $\langle a, c\rangle \in r$ and $\langle c, b\rangle \in s$. Put $d = t(a, c, b)$. Since $\langle c, b\rangle \in s$, we have $\langle t(a, c, c), t(a, b, c)\rangle \in s$, i.e., $\langle a, d\rangle \in s$. Since $\langle b, a\rangle \in r$, we have $\langle t(a, b, c), t(a, a, c)\rangle \in r$, i.e., $\langle d, c\rangle \in r$. Hence $\langle a, c\rangle \in s \circ r$. $\square$

Any term t, satisfying the equations in Theorem 1.4, is called a *Mal'cev term* for the given variety V.

1.5. EXAMPLE. For the variety of groups, $xy^{-1}z$ is a Mal'cev term.

For the variety of quasigroups, both $(x/(y\backslash y)) \cdot (y\backslash z)$ and $((xy)/x)\backslash(xz)$ are Mal'cev terms.

A binary relation r on an algebra A is said to have the *substitution property* if $\langle a_1, b_1\rangle \in r, \dots, \langle a_n, b_n\rangle \in r$ imply $\langle F_A(a_1, \dots, a_n), F_A(b_1, \dots, b_n)\rangle \in r$ for any n-ary operation symbol F from the given signature.

1.6. THEOREM. *Let V be a congruence permutable variety and $A \in V$. Then every reflexive relation r on A with the substitution property is a congruence of A.*

PROOF. Let t be a Mal'cev term for V. If $\langle a, b\rangle \in r$ then

$$\langle b, a\rangle = \langle t(a, a, b), t(a, b, b)\rangle \in r.$$

If $\langle a, b\rangle \in r$ and $\langle b, c\rangle \in r$ then

$$\langle a, c\rangle = \langle t(a, b, b), t(b, b, c)\rangle \in r.$$

□

1.7. THEOREM. (Fleischer [55]) *Let V be a congruence permutable variety, $A, B \in V$ and let a subalgebra C of $A \times B$ be a subdirect product. Then there exist an algebra $D \in V$, a homomorphism f of A onto D and a homomorphism g of B onto D such that $C = \{\langle a, b\rangle : f(a) = g(b)\}$.*

PROOF. For $c = \langle a, b\rangle \in C$ put $p(c) = a$ and $q(c) = b$, so that p is a homomorphism of C onto A and q is a homomorphism of C onto B. Put $\alpha = \mathbf{ker}(p) \vee \mathbf{ker}(q)$ and $D = C/\alpha$. There exist a unique homomorphism f of A onto D and a unique homomorphism g of B onto D such that $fp(c) = gq(c) = c/\alpha$ for all $c \in C$. For $\langle a, b\rangle \in C$ we have $f(a) = fp(\langle a, b\rangle) = gq(\langle a, b\rangle) = g(b)$. Now let $a \in A$ and $b \in B$ be such that $f(a) = g(b)$. There are elements $c_1, c_2 \in C$ with $p(c_1) = a$ and $q(c_2) = b$. Since $\langle c_1, c_2\rangle \in \alpha = \mathbf{ker}(p) \circ \mathbf{ker}(q)$, there is an element $c \in C$ with $\langle c_1, c\rangle \in \mathbf{ker}(p)$ and $\langle c, c_2\rangle \in \mathbf{ker}(q)$. Then $p(c) = a$, $q(c) = b$ and $c = \langle a, b\rangle$. □

1.8. THEOREM. (Foster and Pixley [64],[64a]) *Let V be a congruence permutable variety, $S_1, \dots, S_n \in V$ be simple algebras and let a subalgebra A of $S_1 \times \dots \times S_n$ be a subdirect product. Then $A \simeq S_{i_1} \times \dots \times S_{i_k}$ for some $1 \le i_1 < \dots < i_k \le n$.*

PROOF. By induction on n. For $n = 1$ we have $A = S_1$. Let $n > 1$. Clearly, A is isomorphic to (and can be considered identical with) a subalgebra of $B \times S_n$ where B is a subdirect product of $S_1, \dots, S_{n-1}$. By 1.7 there exist an algebra $D \in V$, a homomorphism f of B onto D and a homomorphism g of S_n onto D such that $A = \{\langle b, s\rangle : f(b) = g(s)\}$. Since S_n is simple, either g is an isomorphism or D is trivial. In the first case we have $A = \{\langle b, g^{-1}f(b)\rangle\} \simeq B$ and in the second case $A = B \times S_n$. It remains to use the induction assumption. □

For two relations r, s and a positive integer n we define a relation $(r, s)^n$ as follows:

$$(r, s)^1 = r, \quad (r, s)^2 = r \circ r, \quad (r, s)^3 = r \circ s \circ r, \quad (r, s)^4 = r \circ s \circ r \circ s, \quad \dots$$

An algebra A is said to have *n-permutable congruences* if $(r, s)^n = (s, r)^n$ for any two congruences r, s of A. A variety V is said to have n-permutable congruences if every algebra from V has n-permutable congruences. Clearly, 2-permutability means the same as permutability. For $n < m$, n-permutability implies m-permutability.

1.9. THEOREM. *Let $n \ge 2$. A variety V has n-permutable congruences if and only if there are terms $t_0, \dots, t_n$ in $n + 1$ variables $x_1, \dots, x_{n+1}$ such that $t_0 = x_1$, $t_n = x_{n+1}$ and the following equations are satisfied in V:*

(1) $t_{i-1}(x, x, y, y, \dots) \approx t_i(x, x, y, y, \dots)$ *for i even;*

(2) $t_{i-1}(x,y,y,z,z,\dots) \approx t_i(x,y,y,z,z,\dots)$ *for i odd.*

PROOF. It is similar to the proof of 1.4. □

1.10. THEOREM. *An algebra A has 3-permutable congruences if and only if the following is true: if f is a homomorphism of A onto an algebra B, then for any congruence r of A, the relation $f(r) = \{\langle f(a), f(b)\rangle : \langle a,b\rangle \in r\}$ is a congruence of B.*

PROOF. Let A have 3-permutable congruences, let f be a homomorphism of A onto B and let r be a congruence of A. Clearly, $f(r)$ is a congruence if it is a transitive relation. Let $\langle a,b\rangle \in f(r)$ and $\langle b,c\rangle \in f(r)$. There exist pairs $\langle a_1,b_1\rangle \in r$ and $\langle b_2,c_2\rangle \in r$ such that $a = f(a_1)$, $b = f(b_1) = f(b_2)$ and $c = f(c_2)$. We have $\langle a_1,c_2\rangle \in r \circ \mathbf{ker}(f) \circ r = \mathbf{ker}(f) \circ r \circ \mathbf{ker}(f)$, so that there exist elements d,e such that $\langle a_1,d\rangle \in \mathbf{ker}(f)$, $\langle d,e\rangle \in r$ and $\langle e,c_2\rangle \in \mathbf{ker}(f)$. Since $a = f(d)$, $c = f(e)$ and $\langle d,e\rangle \in r$, we get $\langle a,c\rangle \in f(r)$.

In order to prove the converse, let $\langle a,b\rangle \in r \circ s \circ r$ where r,s are two congruences of A. There exist elements c,d such that $\langle a,c\rangle \in r$, $\langle c,d\rangle \in s$ and $\langle d,b\rangle \in r$. Denote by g the canonical homomorphism of A onto A/s. We have $\langle g(a), g(c)\rangle \in g(r)$ and $\langle g(c), g(b)\rangle = \langle g(d), g(b)\rangle \in g(r)$. Since $g(r)$ is transitive, we get $\langle g(a), g(b)\rangle \in g(r)$, so that there exists a pair $\langle a_1,b_1\rangle \in r$ with $g(a) = g(a_1)$ and $g(b) = g(b_1)$. Hence $\langle a,b\rangle \in s \circ r \circ s$. □

2. Distributive congruences

A variety V is said to be *congruence distributive* if the congruence lattice of any algebra in V is distributive.

2.1. THEOREM. (Jónsson [67]) *A variety V is congruence distributive if and only if for some $n \geq 1$ there are terms $t_0(x,y,z),\dots,t_n(x,y,z)$ in three variables x,y,z such that $t_0 = x$, $t_n = z$ and the following equations are satisfied in V:*

(1) $t_i(x,y,x) \approx x$ *for all i;*
(2) $t_{i-1}(x,x,y) \approx t_i(x,x,y)$ *for $i < n$ odd;*
(3) $t_{i-1}(x,y,y) \approx t_i(x,y,y)$ *for $i < n$ even.*

A nontrivial variety is congruence distributive if and only if its free algebra of rank 3 has distributive congruence lattice.

PROOF. Let V be congruence distributive. Denote by F the free V-algebra over $\{x,y,z\}$, by T the algebra of terms over $\{x,y,z\}$ and by h the homomorphism of T onto F acting as the identity on $\{x,y,z\}$. Put $s = \mathbf{Cg}_F(x,y)$, $t = \mathbf{Cg}_F(y,z)$ and $r = \mathbf{Cg}_F(x,z)$. Since $\langle x,z\rangle \in r \cap (s \vee t) \subseteq (r \cap s) \vee (r \cap t)$, there exist elements $d_0,\dots,d_n \in F$ such that $d_0 = x$, $d_n = z$, $\langle d_{i-1}, d_i\rangle \in r \cap s$ for i odd and $\langle d_{i-1}, d_i\rangle \in r \cap t$ for i even. Take $t_0,\dots,t_n$ in such a way that $h(t_i) = d_i$, $t_0 = x$ and $t_n = z$. Applying 1.3, one can prove that the equations are satisfied in V.

Conversely, suppose that there exist terms $t_0,\dots,t_n$ as above. Let r,s,u be three congruences of an algebra $A \in V$. For every $m \geq 1$ put $q_m = (s,u)^m$.

In order to prove $r \cap (s \vee u) \subseteq (r \cap s) \vee (r \cap u)$, it is sufficient to prove $r \cap q_m \subseteq (r \cap s) \vee (r \cap u)$ by induction on m. This is evident for $m = 1$. Let $\langle a, b \rangle \in r \cap q_{m+1}$. There exists an element c such that $\langle a, b \rangle \in r$, $\langle a, c \rangle \in q_m$ and $\langle c, b \rangle \in q$, where q is either s or u. For $i = 0, \dots, n$ put $d_i = t_i(a, c, b)$. Clearly, $\langle a, d_i \rangle \in r$ for all i. For i odd we have

$$\langle d_{i-1}, t_{i-1}(a, a, b) \rangle \in q_m^{-1}, \quad t_{i-1}(a, a, b) = t_i(a, a, b),$$
$$\langle t_i(a, a, b), d_i \rangle \in q_m, \quad \langle t_{i-1}(a, a, b), a \rangle \in r, \quad \langle t_i(a, a, b), a \rangle \in r;$$

hence $\langle d_{i-1}, d_i \rangle \in (r \cap q_m^{-1}) \circ (r \cap q_m) \subseteq (r \cap s) \vee (r \cap u)$, where we have used the induction assumption. For i even we have evidently $\langle d_{i-1}, d_i \rangle \in q$, so that $\langle d_{i-1}, d_i \rangle \in r \cap q \subseteq (r \cap s) \vee (r \cap u)$. □

Terms $t_0, \dots, t_n$, satisfying the equations in 2.1, are called *Jónsson terms* for the given variety.

2.2. Example. For the variety of lattices, one can put $n = 2$, $t_0 = x$, $t_1 = (x \wedge y) \vee (x \wedge z) \vee (y \wedge z)$ and $t_2 = z$. Consequently, the variety of lattices is congruence distributive.

3. Modular congruences

A variety V is said to be *congruence modular* if the congruence lattice of any algebra in V is modular. If a variety is either congruence permutable or congruence distributive, then it is congruence modular.

3.1. Theorem. (Day [69]) *A variety V is congruence modular if and only if for some $n \geq 1$ there are terms $t_0(x, y, z, u), \dots, t_n(x, y, z, u)$ in four variables x, y, z, u such that $t_0 = x$, $t_n = u$ and the following equations are satisfied in V:*

(1) *$t_i(x, y, y, x) \approx x$ for all i;*
(2) *$t_{i-1}(x, x, y, y) \approx t_i(x, x, y, y)$ for i odd;*
(3) *$t_{i-1}(x, y, y, z) \approx t_i(x, y, y, z)$ for i even.*

Proof. Let V be congruence modular. Denote by F the free V-algebra over $\{x, y, z, u\}$, by T the algebra of terms over $\{x, y, z, u\}$ and by h the homomorphism of T onto F acting as the identity on $\{x, y, z, u\}$. Put $p = \mathbf{Cg}_F(y, z)$, $r = \mathbf{Cg}(\{\langle x, y \rangle, \langle z, u \rangle\})$ and $s = \mathbf{Cg}(\{\langle x, u \rangle, \langle y, z \rangle\})$. By modularity we have $(p \vee r) \cap s \subseteq p \vee (r \cap s)$. Since $\langle x, u \rangle \in (p \vee r) \cap s$, we get $\langle x, u \rangle \in p \vee (r \cap s)$ and there exist elements $d_0, \dots, d_n \in F$ such that $d_0 = x$, $d_n = u$, $\langle d_{i-1}, d_i \rangle \in r \cap s$ for i odd and $\langle d_{i-1}, d_i \rangle \in p$ for i even. Take $t_0, \dots, t_n$ in such a way that $h(t_i) = d_i$, $t_0 = x$ and $t_n = u$. By an easy application of 1.3, the equations are satisfied in V.

Conversely, suppose that there exist terms $t_0, \dots, t_n$ as above. Let p, r, s be three congruences of an algebra $A \in V$ such that $p \subseteq s$. In order to prove $(p \vee r) \cap s \subseteq p \vee (r \cap s)$, it is sufficient to prove $s \cap q_m \subseteq p \vee (r \cap s)$ for all m, where $q_m = (r, p)^m$. For $m < 3$ it is easy. Let $m \geq 3$.

First suppose that m is odd. Let $\langle a, b \rangle \in s \cap q_m$. There are elements c, d such that $\langle a, b \rangle \in s$, $\langle a, c \rangle \in q_{m-2}$, $\langle c, d \rangle \in p$ and $\langle d, b \rangle \in r$. Put $d_i = t_i(a, c, d, b)$

for $i = 0, \dots, n$. We have $\langle a, d_i\rangle \in s$ for all i. For i odd we have

$$\langle d_{i-1}, t_{i-1}(a,a,b,b)\rangle \in q_{m-2}^{-1}, \quad t_{i-1}(a,a,b,b) = t_i(a,a,b,b),$$
$$\langle t_i(a,a,b,b), d_i\rangle \in q_{m-2}, \quad \langle t_{i-1}(a,a,b,b), a\rangle \in s, \quad \langle t_i(a,a,b,b), a\rangle \in s;$$

hence $\langle d_{i-1}, d_i\rangle \in (s \cap q_{m-2}^{-1}) \circ (s \cap q_{m-2}) \subseteq p \vee (r \cap s)$ by induction. For i even we have evidently $\langle d_{i-1}, d_i\rangle \in p \subseteq p \vee (r \cap s)$. Since $d_0 = a$ and $d_n = b$, we get $\langle a, b\rangle \in p \vee (r \cap s)$.

Now let m be even. Let $\langle a, b\rangle \in s \cap q_m$. There exists an element c such that $\langle a, b\rangle \in s$, $\langle a, c\rangle \in q_{m-1}$ and $\langle c, b\rangle \in p$. Put $d_i = t_i(a, c, c, b)$. We have $\langle a, d_i\rangle \in s$ for all i. For i odd we have

$$\langle d_{i-1}, t_{i-1}(a,a,b,b)\rangle \in q_{m-1}^{-1}, \quad t_{i-1}(a,a,b,b) = t_i(a,a,b,b),$$
$$\langle t_i(a,a,b,b), d_i\rangle \in q_{m-1}, \quad \langle t_{i-1}(a,a,b,b), a\rangle \in s, \quad \langle t_i(a,a,b,b), a\rangle \in s;$$

hence $\langle d_{i-1}, d_i\rangle \in (s \cap q_{m-1}^{-1}) \circ (s \cap q_{m-1}) \subseteq p \vee (r \cap s)$ by induction. For i even clearly $d_{i-1} = d_i$. We get $\langle a, b\rangle \in p \vee (r \cap s)$. □

Terms $t_0, \dots, t_n$, satisfying the equations in 3.1, are called *Day terms* for the given variety.

The following result belongs to Gumm [83]; we present a more simple proof due to W. Taylor.

3.2. THEOREM. *A variety V is congruence modular if and only if for some $m \geq 1$ there are terms $p(x, y, z)$ and $q_1(x, y, z), \dots, q_m(x, y, z)$ in three variables x, y, z (called* Gumm terms*) such that the following equations are satisfied in V:*

(1) $p(x, z, z) \approx x$
(2) $p(x, x, z) \approx q_1(x, x, z)$
(3) $q_i(x, y, x) \approx x$ *for all* i
(4) $q_m \approx z$
(5) $q_i(x, z, z) \approx q_{i+1}(x, z, z)$ *for* i *odd*
(6) $q_i(x, x, z) \approx q_{i+1}(x, x, z)$ *for* i *even*

A nontrivial variety is congruence modlar if and only if its free algebra of rank 3 *has modular congruence lattice.*

PROOF. From Gumm terms we can produce Day terms as follows:

$$t_0 = t_1 = x$$
$$t_2(x, y, z, u) = p(x, y, z)$$
$$t_3(x, y, z, u) = q_1(x, y, u)$$
$$t_4(x, y, z, u) = q_1(x, z, u)$$
$$t_{4i+1}(x, y, z, u) = q_{2i}(x, z, u)$$
$$t_{4i+2}(x, y, z, u) = q_{2i}(x, y, u)$$
$$t_{4i+3}(x, y, z, u) = q_{2i+1}(x, y, u)$$
$$t_{4i+4}(x, y, z, u) = q_{2i+1})x, z, u)$$

Conversely, let $t_0, \dots, t_n$ be Day terms. One can assume that n is odd (we can add the term u if necessary). Define the following terms:

$$s_i(x,y,z) = \begin{cases} x \text{ for } i = 0 \\ t_i(s_{i-1}, y, z, s_{i-1}) \text{ for } i > 0 \text{ even} \\ t_i(s_{i-1}, z, y, s_{i-1}) \text{ for } i > 0 \text{ odd} \end{cases}$$

$$r_i(x,y,z) = \begin{cases} x \text{ for } i = 0 \\ t_i(t_{i-1}(x,x,x,z), x, y, t_{i-1}(x,x,x,z)) \text{ for } i > 0 \text{ even} \\ t_i(t_{i-1}(x,x,z,z), z, y, t_{i-1}(x,x,z,z)) \text{ for } i > 0 \text{ odd} \end{cases}$$

$$v^j_{i,i}(x,y,z) = \begin{cases} t_j(r_i(x,z,z), r_i, t_i(x,x,y,z), t_i(x,x,z,z)) \text{ for } i \text{ even} \\ t_j(r_i(x,x,z), r_i, t_i(x,x,y,z), t_i(x,x,x,z)) \text{ for } i \text{ odd} \end{cases}$$

$$v^j_{i,k}(x,y,z) = \begin{cases} t_k(v^j_{i,k-1}, x, z, v^j_{i,k-1}) \text{ for } k > i \text{ even} \\ t_k(v^j_{i,k-1}, z, x, v^j_{i,k-1}) \text{ for } k > i \text{ odd} \end{cases}$$

$$w^j_i = v^j_{i,n}$$

Now define Gumm terms: $p = s_n$ and $q_1, \dots, q_m$ are the terms

$$w^0_0, w^1_0, \dots, w^n_0, w^1_1, \dots, w^n_1, \dots, w^1_n, \dots, w^n_n$$

(so that $m = n^2 + n + 1$).

Claim 1. $p(x,x,z) = x$. (Let us write equations with the equality sign.) Indeed, one can check easily by induction on i that $s_i(x,z,z) = x$.

Claim 2. $s_i(x,x,z) = v^0_{0,i}$. By induction on i. For $i = 0$, both sides are x. For i even, $s_i(x,x,z) = t_i(s_{i-1}(x,x,z), x, z, s_{i-1}(x,x,z)) = t_i(v^0_{0,i-1}, x, z, v^0_{0,i-1}) = v^0_{0,i}$. For i odd, the proof is similar.

Claim 3. $p(x,x,z) = w^0_0 = w^0_0(x,x,z)$. By Claim 2 we have $p(x,x,z) = s_n(x,x,z) = v^0_{0,n} = w^0_0$. Also, observe that w^0_0 does not contain y.

Claim 4. $w^j_i(x,y,x) = x$. It is easy to check that $r_i(x,y,x) = t_i(x,x,y,x)$ and then $v^j_{i,k}(x,y,x) = x$ by induction on $k \geq i$.

Claim 5. $w^n_n = z$. This is obvious.

Claim 6. $v^n_{i-1,i} = v^0_{i,i}$ *for* $0 < i \leq n$. For i even we have $v^n_{i-1,i} = t_i(v^n_{i-1,i-1}, x, z, v^n_{i-1,i-1}) = t_i(t_{i-1}(x,x,x,z), x, z, t_{i-1}(x,x,x,z))$ and $v^0_{i,i} = r_i(x,z,z)$ equals the same. For i odd the proof is similar.

Claim 7. $v^n_{i-1,k} = v^n_{i,k}$ *for* $0 < i \leq k \leq n$. Let us prove it by induction on k. For $k = i$, use Claim 6. Let $k > i$. If k is even then $v^n_{i-1,k} = t_k(v^n_{i-1,k-1}, x, z, v^n_{i-1,k-1}) = t_k(v^0_{i,k-1}, x, z, v^0_{i,k-1}) = v^0_{i,k}$. If k is odd then $v^n_{i-1,k} = t_k(v^n_{i-1,k-1}, z, x, v^n_{i-1,k-1}) = t_k(v^0_{i,k-1}, z, x, v^0_{i,k-1}) = v^0_{i,k}$.

Claim 8. $w^n_{i-1} = w^0_i$. This is Claim 7 with $k = n$.

Claim 9. $r_i(x,x,z) = t_i(x,x,x,z)$ *for* i *even.* We have $r_i(x,x,z) = t_i(t_{i-1}(x,x,x,z), x, x, t_{i-1}(x,x,x,z)) = t_{i-1}(x,x,x,z) = t_i(x,x,x,z)$ by the Day equations.

Claim 10. $r_i(x,z,z) = t_i(x,x,z,z)$ *for* i *odd.* We have $r_i(x,z,z) = t_i(t_{i-1}(x,x,z,z), z, z, t_{i-1}(x,x,z,z)) = t_{i-1}(x,x,z,z) = t_i(x,x,z,z)$ by the Day equations.

Claim 11. $v_{i,i}^j(x,x,z) = v_{i,i}^{j+1}(x,x,z)$ *for* $i+j$ *odd.* If i is odd and j is even then $v_{i,i}^j(x,x,z) = t_j(r_i(x,x,z), r_i(x,x,z), t_i(x,x,x,z), t_i(x,x,x,z)) = t_{j+1}(r_i(x,x,z), r_i(x,x,z), t_i(x,x,x,z), t_i(x,x,x,z)) = v_{i,i}^{j+1}(x,x,z)$. If i is even and j is odd, $v_{i,i}^j(x,x,z) = t_j(r_i(x,z,z), r_i(x,x,z), t_i(x,x,x,z), t_i(x,x,z,z)) = t_j(r_i(x,z,z), r_i(x,x,z), r_i(x,x,z), t_{i-1}(x,x,z,z)) = t_{j+1}(r_i(x,z,z), r_i(x,x,z), r_i(x,x,z), t_i(x,x,z,z)) = v_{i,i}^{j+1}(x,x,z)$ (we have used Claim 9 and several times Day's equations).

Claim 12. $v_{i,i}^j(x,z,z) = v_{i,i}^{j+1}(x,z,z)$ *for* $i+j$ *even.* Using Claim 10, the proof is similar to that of Claim 11.

Claim 13. $w_i^j(x,x,z) = w_i^{j+1}(x,x,z)$ *for* $i+j$ *odd.* Let us prove $v_{i,k}^j(x,x,z) = v_{i,k}^{j+1}(x,x,z)$ by induction on $k \geq i$. For $k = i$, this is Claim 11. Let $k > i$. If k is even then $v_{i,k}^j(x,x,z) = t_k(v_{i,k-1}^j(x,x,z), x, z, v_{i,k-1}^j(x,x,z)) = t_k(v_{i,k-1}^{j+1}(x,x,z), x, z, v_{i,k-1}^{j+1}(x,x,z)) = v_{i,k}^{j+1}(x,x,z)$. For k odd, the proof is similar.

Claim 14. $w_i^j(x,z,z) = w_i^{j+1}(x,z,z)$ *for* $i+j$ *even.* The proof is similar to that of Claim 13.

Equation (1) follows from Claim 1, equation (2) from Claim 3, equations (3) from Claim 4, equation (4) from Claim 5 and the equations (5) and (6) from Claims 8, 13 and 14. □

Let a subalgebra B of a product $A = A_1 \times \cdots \times A_n$ be a subdirect product. For any congruences $r_i \in \mathbf{Con}(A)$ $(1 \leq i \leq n)$, the set of the pairs $\langle\langle a_1, \ldots, a_n\rangle, \langle b_1, \ldots, b_n\rangle\rangle$ such that $\langle a_i, b_i\rangle \in r_i$ for all i, is (clearly) a congruence of A and its intersection with $B \times B$ is a congruence of B. Congruences obtained in this way will be called product congruences. The subdirect product B will be called *skew free* if it has no other congruences.

A set S of algebras of the given signature is said to be totally skew free if B is skew free whenever B is a subdirect product of a finite family of algebras from S.

3.3. Lemma. *Let a subalgebra* B *of* $A_1 \times \cdots \times A_n$ *be a subdirect product. Then* B *is skew free if and only if for any congruence* r *of* B, $r = (r \vee s_1) \cap \cdots \cap (r \vee s_n)$, *where* $s_i = (\mathbf{ker}(p_i)) \cap B^2$ *and* p_i *is the* i*-th projection of the product onto* A_i.

Proof. It is easy. □

3.4. Lemma. *Let* L *be a modular lattice and* $a, b \in L$ *be elements such that* $c = (c \vee a) \wedge (c \vee b)$ *for all* $c \geq a \wedge b$. *Then* $c = (c \vee (a \wedge d)) \wedge (c \vee (b \wedge d))$ *for all* $c, d \in L$ *with* $a \wedge b \wedge d \leq c \leq d$.

Proof. We have $c = c \vee (a \wedge b \wedge d) = (c \vee (a \wedge b)) \wedge d = (c \vee (a \wedge b) \vee a) \wedge (c \vee (a \wedge b) \vee b) \wedge d = (c \vee a) \wedge (c \vee b) \wedge d = (c \vee (a \wedge d)) \wedge (c \vee (b \wedge d))$. □

3.5. Lemma. *Let L be a modular lattice and $a_1, \dots, a_n \in L$ be elements such that $c = (c \vee a_i) \wedge (c \vee a_j)$ whenever $i, j \in \{1, \dots, n\}$ and $c \geq a_i \wedge a_j$. Then $c = (c \vee a_1) \wedge \dots \wedge (c \vee a_n)$ whenever $c \geq a_1 \wedge \dots \wedge a_n$.*

Proof. By induction on n. For $n \leq 2$ there is nothing to prove. Let $n \geq 3$. If c is an element such that $(a_1 \wedge a_i) \wedge (a_1 \wedge a_j) \leq c \leq a_1$ then, by 3.4, $c = (c \vee (a_1 \wedge a_i)) \wedge (c \vee (a_1 \wedge a_j)) = (c \vee a_i) \wedge (c \vee a_j)$. Hence in the sublattice $\downarrow a_1$ of L, the $n-1$ elements $a_1 \wedge a_2, \dots, a_1 \wedge a_n$ satisfy the induction hypothesis. By induction we get that $c \geq a_1 \wedge \dots \wedge a_n$ implies

$$\begin{aligned} c \wedge a_1 &= ((c \wedge a_1) \vee (a_1 \wedge a_2)) \wedge \dots \wedge ((c \wedge a_1) \vee (a_1 \wedge a_n)) \\ &= (c \vee (a_1 \wedge a_2)) \wedge a_1 \wedge \dots \wedge (c \vee (a_1 \wedge a_n)) \wedge a_1 \\ &= a_1 \wedge (c \vee (a_1 \wedge a_2)) \wedge \dots \wedge (c \vee (a_1 \wedge a_n)). \end{aligned}$$

We have $c \vee (a_1 \wedge a_i) \geq a_1 \wedge a_i$, so that $c \vee (a_1 \wedge a_i) = ((c \vee (a_1 \wedge a_i)) \vee a_1) \wedge ((c \vee (a_1 \wedge a_i)) \vee a_i) = (c \vee a_1) \wedge (c \vee a_i)$. Thus

$$\begin{aligned} c = c \vee (c \wedge a_1) &= c \vee (a_1 \wedge (c \vee a_1) \wedge (c \vee a_2) \wedge \dots \wedge (c \vee a_1) \wedge (c \vee a_n)) \\ &= c \vee (a_1 \wedge (c \vee a_2) \wedge \dots \wedge (c \vee a_n)) \\ &= (c \vee a_1) \wedge (c \vee a_2) \wedge \dots \wedge (c \vee a_n) \end{aligned}$$

by modularity. □

3.6. Theorem. *Let V be a congruence modular variety and let S be a subset of V such that B is skew free whenever B is a subdirect product of a pair of algebras from S. Then S is totally skew free.*

Proof. Let $A_1, \dots, A_n \in S$ and let a subalgebra B of $A_1 \times \dots \times A_n$ be a subdirect product. For $i = 1, \dots, n$ put $s_i = (\mathbf{ker}(p_i)) \cap B^2$ where p_i is the i-th projection. For $i \neq j$, the algebra $B/(s_i \cap s_j)$ is isomorphic to a subdirect product of the pair A_i, A_j. By the assumption applied to this algebra and using 3.3, for every congruence r of B with $s_i \cap s_j \subseteq r$ we have $r = (r \vee s_i) \cap (r \vee s_j)$. By 3.5 it follows that $r = (r \vee s_1) \cap (r \vee s_n)$ for any congruence r of B. Consequently, by 3.3, B is skew free. □

3.7. Theorem. *Let V be a congruence distributive variety. Then every subset of V is totally skew free.*

Proof. It follows from 3.3. □

4. Chinese remainder theorem

By a congruence block of an algebra A we mean a subset of A which is a block of a congruence of A.

4.1. Theorem. *The following three conditions are equivalent for an algebra A:*

(1) *The intersection of a finite system of pairwise non-disjoint congruence blocks of A has a nonempty intersection;*

(2) *The intersection of any triple of pairwise non-disjoint congruence blocks of A has a nonempty intersection;*
(3) $r \cap (s \circ u) \subseteq (r \cap s) \circ (r \cap u)$ *for any three congruences of A.*

PROOF. Evidently, (1) implies (2). (2) implies (1): Let us prove by induction on $n \geq 1$ that if $S_1, \dots, S_n$ are pairwise non-disjoint congruence blocks of A, then $S_1 \cap \dots \cap S_n$ is nonempty. This is evident for $n = 1$. Let $n \geq 2$. It follows from (2) that the congruence blocks $S_1 \cap S_2, S_3, \dots, S_n$ are pairwise disjoint. By induction, their intersection is nonempty. But their intersection is $S_1 \cap \dots \cap S_n$.

(2) implies (3): Let $\langle a, b \rangle \in r \cap (s \circ u)$. There exists an element c such that $\langle a, b \rangle \in r$, $\langle a, c \rangle \in s$ and $\langle c, b \rangle \in u$. Put $S_1 = a/r$, $S_2 = c/s$ and $S_3 = b/u$. We have $a \in S_1 \cap S_2$, $c \in S_2 \cap S_3$ and $b \in S_1 \cap S_3$. Consequently, there exists an element $d \in S_1 \cap S_2 \cap S_3$. Since $\langle a, d \rangle \in r \cap s$ and $\langle d, b \rangle \in r \cap u$, we get $\langle a, b \rangle \in (r \cap s) \circ (r \cap u)$.

(3) implies (2): Let S_1 be a block of a congruence r, S_2 be a block of s and S_3 be a block of u; let $a \in S_1 \cap S_2$, $b \in S_2 \cap S_3$ and $c \in S_1 \cap S_3$. We have $\langle a, c \rangle \in r \cap (s \circ u) \subseteq (r \cap s) \circ (r \cap u)$, so that there exists an element d such that $\langle a, d \rangle \in r \cap s$ and $\langle d, c \rangle \in r \cap u$. Clearly, $d \in S_1 \cap S_2 \cap S_3$. □

An algebra A is said to satisfy the *Chinese remainder theorem* if it satisfies the three equivalent conditions of 4.1. A variety V is said to satisfy the Chinese remainder theorem if every algebra in V does.

4.2. EXAMPLE. The ring of integers satisfies the Chinese remainder theorem. This number theoretic result was proved in old China.

By a *ternary majority term* for a variety V we mean a term t in three variables x, y, z such that V satisfies the equations

$$t(x, x, y) \approx x, \quad t(x, y, x) \approx x, \quad t(y, x, x) \approx x.$$

4.3. THEOREM. *A variety V satisfies the Chinese remainder theorem if and only if there exists a ternary majority term for V.*

PROOF. Let V satisfy the Chinese remainder theorem. Denote by F the free V-algebra over x, y, z, by T the algebra of terms over x, y, z and by h the homomorphism of T onto F acting as the identity on x, y, z. Put $r = \mathbf{Cg}_F(x, z)$, $s = \mathbf{Cg}_F(x, y)$ and $u = \mathbf{Cg}_F(y, v)$. We have $\langle x, z \rangle \in r \cap (s \circ u) \subseteq (r \cap s) \circ (r \cap u)$, so that there exists an element $d \in F$ with $\langle x, d \rangle \in r \cap s$ and $\langle d, z \rangle \in r \cap u$. Using 1.3 we can see that any term $t \in T$ such that $h(t) = d$ is a ternary majority term for V.

Conversely, let t be a ternary majority term for V. Let $A \in V$ and $\langle a, b \rangle \in r \cap (s \circ u)$, where r, s, u are three congruences of A. There exists an element c such that $\langle a, b \rangle \in r$, $\langle a, c \rangle \in s$ and $\langle c, b \rangle \in u$. Put $d = t(a, b, c)$. We have $\langle a, d \rangle \in r \cap s$ and $\langle d, b \rangle \in r \cap u$, so that $\langle a, b \rangle \in (r \cap s) \circ (r \cap u)$. □

4.4. THEOREM. *An algebra with permutable congruences has distributive congruences if and only if it satisfies the Chinese remainder theorem. A variety* *with a ternary majority term (i,e., a variety satisfying the Chinese remainder theorem) is congruence distributive.*

Proof. It follows from the above results. □

Example 2.2 actually shows that the variety of lattices satisfies the Chinese remainder theorem.

5. Arithmetical varieties

A variety is said to be *arithmetical* if it is both congruence permutable and congruence distributive.

5.1. Theorem. *The following three conditions are equivalent for a variety V:*

(1) *V is arithmetical;*
(2) *V has both a Mal'cev term and a ternary majority term;*
(3) *there exists a term p in three variables x, y, z such that V satisfies*

$$p(x,y,y) \approx x, \quad p(x,x,y) \approx y, \quad p(x,y,x) \approx x.$$

Proof. The equivalence of the first two conditions follows from 4.3 and 4.4.

(2) implies (3): If t is a Mal'cev term and M is a ternary majority term for V, put $p = M(x, t(x,y,z), z)$.

(3) implies (2): p is a Mal'cev term, and $p(x, p(x,y,z), z)$ is a ternary majority term for V. □

5.2. Example. The variety of Boolean algebras is an arithmetical variety.

5.3. Theorem. (Baker and Pixley [75]) *Let V be an arithmetical variety and $A \in V$ be a finite algebra. An n-ary operation f on A (where $n \geq 1$) is a term operation of A if and only if it preserves subalgebras of A^2 (i.e., whenever S is a subalgebra of A^2 and $\langle a_1, b_1\rangle, \dots, \langle a_n, b_n\rangle \in S$ then $\langle f(a_1, \dots, a_n), f(b_1, \dots, b_n)\rangle \in S$).*

Proof. The direct implication is clear. If S is a subalgebra of A then $\{\langle a, a\rangle : a \in S\}$ is a subalgebra of A^2. From this it follows that f also preserves subalgebras of A. We are going to prove by induction on k that for every k-element subset U of A^n there exists an n-ary term operation of A coinciding with f on U. For $k = 1$ it follows from the fact that f preserves subalgebras of A; for $k = 2$ from the fact that f preserves subalgebras of A^2. Let $k \geq 3$. Take three distinct elements u_1, u_2, u_3 of U. By the induction assumption there exist three term operations g_1, g_2, g_3 of A such that $g_i(u) = f(u)$ for all $u \in U \setminus \{u_i\}$ $(i = 1, 2, 3)$. Where M is the ternary majority term for V, the term operation $g(x_1, \dots, x_n) = M(g_1(x_1, \dots, x_n), g_2(x_1, \dots, x_n), g_3(x_1, \dots, x_n))$ coincides with f on U. □

6. Congruence regular varieties

An algebra A is said to be *congruence regular* if any two congruences of A with a common block are equal. A variety is said to be congruence regular if all its algebras are.

6.1. LEMMA. *An algebra A is congruence regular if and only if for every triple a, b, c of elements of A there exists a subset S of A such that $\langle a, b\rangle$ belongs to the congruence generated by $S \times S$ and $\langle c, d\rangle \in \mathbf{Cg}_A(a, b)$ for all $d \in S$.*

PROOF. Let A be congruence regular and $a, b, c \in A$. Put $S = c/r$, where $r = \mathbf{Cg}_A(a, b)$. The congruence generated by $S \times S$ has a common block S with r and hence equals r, so that it contains $\langle a, b\rangle$. Of course, $\langle c, d\rangle \in r$ for all $d \in S$.

In order to prove the converse, let r, s be two congruences of A with a common block C. It is sufficient to prove that $\langle a, b\rangle \in r$ implies $\langle a, b\rangle \in s$. Take an element $c \in C$. There exists a subset S for the triple a, b, c as above. Since $\langle c, d\rangle \in \mathbf{Cg}_A(a, b) \subseteq r$ for all $d \in S$, we have $S \subseteq C$. Hence $\langle a, b\rangle \in \mathbf{Cg}_A(S \times S) \subseteq \mathbf{Cg}_A(C \times C) \subseteq s$. □

6.2. THEOREM. *A variety V is congruence regular if and only if for some $n \geq 1$ there are terms $t_1, \dots, t_n, u_1, \dots, u_n$ in three variables x, y, z and terms $v_1, \dots, v_n$ in four variables x, y, z, u such that the following equations are satisfied in V:*

(1) $t_i(x, x, z) \approx z$ *for all* i;
(2) $u_i(x, x, z) \approx z$ *for all* i;
(3) $v_1(x, y, z, t_1) \approx x$;
(4) $v_{i-1}(x, y, z, u_{i-1}) \approx v_i(x, y, z, t_i)$ *for* $i = 2, \dots, n$;
(5) $v_n(x, y, z, u_n) \approx y$.

PROOF. Let V be congruence regular. Denote by F the free V-algebra over $\{x, y, z\}$, by T the algebra of terms over $\{x, y, z\}$ and by f the homomorphism of T onto F acting as the identity on $\{x, y, z\}$. By 6.1 there exists a subset S of F such that $\langle x, y\rangle \in \mathbf{Cg}_F(S \times S)$ and $\langle z, a\rangle \in \mathbf{Cg}_F(x, y)$ for all $a \in S$. Since $\langle x, y\rangle \in \mathbf{Cg}_F(S \times S)$, there exist unary polynomials $f_1, \dots, f_n$ of F and pairs $\langle a_1, b_1\rangle, \dots, \langle a_n, b_n\rangle \in S \times S$ for some $n \geq 1$ such that

$$x = f_1(a_1), \quad f_1(b_1) = f_2(a_2), \quad \dots, \quad f_{n-1}(b_{n-1}) = f_n(a_n), \quad f_n(b_n) = y.$$

There exist ternary terms $t_1, \dots, t_n, u_1, \dots, u_n$ such that $a_1 = f(t_1), \dots, a_n = f(t_n), b_1 = f(u_1), \dots, b_n = f(u_n)$. Denote by g the endomorphism of F with $g(x) = x$, $g(y) = x$ and $g(z) = z$. For all i we have $\langle z, a_i\rangle \in \mathbf{ker}(g)$, so that

$$f(z) = z = g(z) = g(a_i) = g(f(t_i)) = f(t_i(x, x, z)).$$

From this we get (1), and (2) can be proved similarly. For every $i = 1, \dots, n$ there exist a positive integer k_i, a term s_i in variables $x_1, \dots, x_{k_i}$ and elements $a_{i,2}, \dots, a_{i,k_i} \in F$ such that $f_i(a) = s_i^F(a, a_{i,2}, \dots, a_{i,k_i})$ for all $a \in F$. We have $a_{i,2} = f(w_{i,2}), \dots, a_{i,k_i} = f(w_{i,k_i})$ for some terms $w_{i,2}, \dots, w_{i,k_i} \in T$. Put

$v_i = s_i(u, w_{i,2}, \dots, w_{i,k_i})$, so that v_i is a term in variables x, y, z, u. We have

$$\begin{aligned} f(v_1(x,y,z,t_1)) = f(s_1(t_1, w_{1,2}, \dots, w_{1,k_1})) &= s_1^F(f(t_1), f(w_{1,2}), \dots, f(w_{1,k_1}) \\ &= s_1^F(a_1, \dots, a_{1,2}, \dots, a_{1,k_1}) = f_1(a_1) = x = f(x), \end{aligned}$$

so that V satisfies (3). One can prove (4) and (5) similarly.

In order to prove the converse, let $A \in V$ and $a, b, c \in A$. For $i = 1, \dots, n$ put $a_i = t_i^F(a, b, c)$ and $b_i = u_i^F(a, b, c)$. Put $S = \{a_1, \dots, a_n, b_1, \dots, b_n\}$. If r is a congruence containing $\langle a, b\rangle$, then $\langle c, a_i\rangle = \langle t_i^F(a, a, c), t_i^F(a, b, c)\rangle \in r$ and similarly $\langle c, b_i\rangle \in r$ for all i. If R is a congruence containing $S \times S$, then

$$\begin{aligned} a &= v_1^F(a,b,c,a_1) \mathrel{R} v_1^F(a,b,c,b_1) = v_2^F(a,b,c,a_2) \mathrel{R} v_2^F(a,b,c,b_2) = \dots \\ &= v_{n-1}^F(a,b,c,a_{n-1}) \mathrel{R} v_{n-1}^F(a,b,c,b_{n-1}) \\ &= v_n^F(a,b,c,a_n) \mathrel{R} v_n^F(a,b,c,b_n) = b. \end{aligned}$$

Now we are able to apply 6.1. □

6.3. Example. The variety of quasigroups is congruence regular: put $n = 1$, $t_1 = y(x\backslash z)$, $u_1 = z$, $v_1 = (y(x\backslash z))/(x\backslash u)$.

7. Congruence distributive varieties

7.1. Theorem. (Jónsson [67]) *Let K be a class of algebras such that the variety $\mathbf{HSP}(K)$ is congruence distributive. Then every subdirectly irreducible algebra from $\mathbf{HSP}(K)$ is a homomorphic image of a subalgebra of an ultraproduct of a family of algebras from K.*

Proof. Let B be a subdirectly irreducible algebra from $\mathbf{HSP}(K)$. There exist a family H of algebras from K (denote its domain by I) and a subalgebra A of ΠH such that $B \simeq A/r$ for a congruence r of A. For every subset J of I define a congruence s_J of A in this way: $\langle f, g\rangle \in s_J$ if and only if $f(i) = g(i)$ for all $i \in J$. Put $D = \{J \subseteq I : s_J \subseteq r\}$. We have $I \in D$, and $J \in D$ implies $J' \in D$ for all $J \subseteq J' \subseteq I$. It follows from Zorn's lemma that there is a filter F of subsets of I which is maximal among the filters contained in D. We are going to prove that F is an ultrafilter.

Suppose, on the contrary, that there exists a subset J of I such that neither J nor $I \smallsetminus J$ belongs to F. It follows from the maximality of F that there exist two subsets $L_1, L_2 \in F$ with $J \cap L_1 \notin D$ and $(I \smallsetminus J) \cap L_2 \notin D$. Put $M = L_1 \cap L_2$, so that $M \in F$. Then M is the disjoint union of two subsets X, Y not belonging to D: put $X = J \cap M$ and $Y = (I \smallsetminus J) \cap M$. Since $M \in D$, we have

$$r = r \vee s_M = r \vee (s_X \cap s_Y) = (r \vee s_X) \cap (r \vee s_Y)$$

by the congruence distributivity. But A/r is subdirectly irreducible, and it follows that either $r = r \vee s_X$ or $r = r \vee s_Y$, i.e., either $X \in D$ or $Y \in D$, a contradiction.

So, F is an ultrafilter. The corresponding ultraproduct of H is the factor of ΠH through the congruence R defined as follows: $\langle f, g\rangle \in R$ if and only if

$\{i : f(i) = g(i)\} \in F$. Since $F \subseteq D$, we have $R \cap (A \times A) \subseteq r$, and hence A/r is a homomorphic image of the ultraproduct. □

7.2. THEOREM. *Let K be a finite set of finite algebras such that $\mathbf{HSP}(K)$ is a congruence distributive variety. Then every subdirectly irreducible algebra from $\mathbf{HSP}(K)$ belongs to $\mathbf{HS}(K)$.*

PROOF. It follows from 7.1, as any ultraproduct of any family of algebras from K is isomorphic to an algebra from K. □

7.3. THEOREM. *Let V be a congruence distributive variety and $A, B \in V$ be two algebras such that A is finite, B is subdirectly irreducible, $\mathbf{card}(A) \leq \mathbf{card}(B)$ and $A \not\simeq B$. Then there is an equation that is satisfied in A but not satisfied in B.*

PROOF. It follows from 7.2. □

8. Congruence meet-semidistributive varieties

A lattice is said to be *meet-semidistributive* if it satisfies the quasiequation $x \wedge y = x \wedge z \to x \wedge y = x \wedge (y \vee z)$. A variety V is said to be congruence meet-semidistributive if the congruence lattice of any algebra from V is meet-semidistributive.

Let α, β, γ be three congruences of an algebra A. Define congruences β_n and γ_n for $n \geq 0$ in this way: $\beta_0 = \beta$, $\gamma_0 = \gamma$, $\beta_{n+1} = \beta \vee (\alpha \cap \gamma_n)$, $\gamma_{n+1} = \gamma \vee (\alpha \cap \beta_n)$. Clearly, the congruences β_n constitute a chain and so their union β_∞ is also a congruence. Similarly, the union of the chain of congruences γ_n is a congruence γ_∞. It is easy to see that $\alpha \cap \beta_\infty = \alpha \cap \gamma_\infty$ and if β', γ' are two congruences such that $\beta \subseteq \beta'$, $\gamma \subseteq \gamma'$ and $\alpha \cap \beta' = \alpha \cap \gamma'$ then $\beta_\infty \subseteq \beta'$ and $\gamma_\infty \subseteq \gamma'$.

8.1. THEOREM. (Willard [00]) *The following are equivalent for a nontrivial variety V:*

(1) *V is congruence meet-semidistributive;*
(2) *in the V-free algebra over $\{x, y, z\}$ we have $\langle x, z\rangle \in \beta_n$ for some n, where $\alpha = \mathbf{Cg}(x, z)$, $\beta = \mathbf{Cg}(x, y)$ and $\gamma = \mathbf{Cg}(y, z)$;*
(3) *there exist a finite set E and ternary terms s_e, t_e ($e \in E$) such that the equations $s_e(x, y, x) \approx t_e(x, y, x)$ are satisfied in V for all $e \in E$, and for any algebra $A \in V$ and any elements $a, b \in A$, $a = b$ if and only if $s_e(a, a, b) = t_e(a, a, b) \leftrightarrow s_e(a, b, b) = t_e(a, b, b)$ for all $e \in E$;*
(4) *for any algebra $A \in V$ and any finite sequence $a_0, a_1, \dots, a_n$ of elements of A such that $a_0 \neq a_n$ there exists an $i < n$ such that $\mathbf{Cg}(a_0, a_n) \cap \mathbf{Cg}(a_i, a_{i+1}) \neq \mathbf{id}_A$;*

PROOF. (1) implies (2): $\langle x, z\rangle \in \alpha \cap (\beta \circ \gamma) \subseteq \alpha \cap (\beta \vee \gamma) \subseteq \alpha \cap (\beta_\infty \vee \gamma_\infty) = \alpha \cap \beta_\infty$ and thus $\langle x, z\rangle \in \beta_n$ for some n.

(2) implies (3): Let us define by induction on $k = 0, \dots, n$ a finite set E_k of finite sequences with k members of positive integers, and for every $e \in E_k$ a pair of ternary terms s_e, t_e with $\langle s_e, t_e\rangle \in \alpha \cap \beta_{n-k}$ if k is even while $\langle s_e, t_e\rangle \in \alpha \cap \gamma_{n-k}$ if k is odd. Let E_0 contain a single element, the empty sequence, and put $s_\emptyset = x$

and $t_\emptyset = z$ (so that $\langle s_\emptyset, t_\emptyset\rangle \in \alpha \cap \beta_n$). Now assume that E_k and s_e, t_e ($e \in E_k$) are already defined for some $k < n$. Take any $e \in E_k$. If k is even then $\langle s_e, t_e\rangle \in \beta_{n-k} = \beta \vee (\alpha \cap \gamma_{n-k-1})$, so that there exists a finite sequence of ternary terms $s_{e1}, t_{e1}, \dots, s_{em}, t_{em}$ such that $\langle s_e, s_{e1}\rangle \in \beta$, $\langle s_{ei}, t_{ei}\rangle \in \alpha \cap \gamma_{n-k-1}$ for $1 \le i \le m$, $\langle t_{ei}, s_{e(i+1)}\rangle \in \beta$ for $1 \le i < m$ and $\langle t_{em}, t_e\rangle \in \beta$; add the sequences $e1, \dots, em$ to E_{k+1}. If k is odd, do the same with β replaced by γ and concersely.

Put $E = E_0 \cup \dots \cup E_n$. The set E can be imagined as a rooted tree, with the root $\emptyset$ and leaves the sequences from E that cannot be extended to longer sequences in E. Clearly, we have $\langle s_e, t_e\rangle \in \alpha$ for all $e \in E$, so that the equation $s_e(x, y, x) \approx t_e(x, y, x)$ is satisfied in V. For any leaf $e \in E_k$ we have $\langle s_e, t_e\rangle \in \beta$ if k is even, while $\langle s_e, t_e\rangle \in \gamma$ if k is odd. Observe that if u, v are ternary terms such that $\langle u, v\rangle \in \beta$ then $u(x, x, y) \approx v(x, x, y)$ is satisfied in V, and if $\langle u, v\rangle \in \gamma$ then $u(x, y, y) \approx v(x, y, y)$ is satisfied in V.

Let $A \in V$, $a, b \in A$ and let $s_e(a, a, b) = t_e(a, a, b) \leftrightarrow s_e(a, b, b) = t_e(a, b, b)$ for all $e \in E$. We will prove by induction on $n - k$ that $s_e(a, a, b) = t_e(a, a, b)$ and $s_e(a, b, b) = t_e(a, b, b)$. If e is a leaf then $\langle s_e, t_e\rangle$ belongs to either β or γ, so that we have one of the two equalities and then by the assumption we have both. Let $e \in E_k$ be not a leaf, so that e can be continued to some $e1$, $\dots$, e_m in E_{k+1} obtained in the above described way. If k is even, each two neighbors in the sequence $s_e(a, a, b)$, $s_{e1}(a, a, b)$, $t_{e1}(a, a, b)$, $\dots$, $s_{em}(a, a, b)$, $t_{em}(a, a, b)$, $t_e(a, a, b)$ are equal because in each case we can either use the induction hypothesis or the pair of the terms belongs to β; thus $s_e(a, a, b) = t_e(a, a, b)$ and then $s_e(a, b, b) = t_e(a, b, b)$ follows by the assumption. If k is odd, we can similarly prove $s_e(a, b, b) = t_e(a, b, b)$ and then obtain $s_e(a, a, b) = t_e(a, a, b)$ by the assumption.

(3) implies (4): Put $a = a_0$ and $b = a_n$. Since $a \ne b$, there exists an $e \in E$ such that either $s_e(a, a, b) = t_e(a, a, b)$ but $s_e(a, b, b) \ne t_e(a, b, b)$, or conversely; assume this first case. Clearly, there exists an $i < n$ such that $s_e(a, a_i, b) = t_e(a, a_i, b)$ while $s_e(a, a_{i+1}, b) \ne t_e(a, a_{i+1}, b)$. Put $c = s_e(a, a_{i+1}, b)$ and $d = t_e(a, a_{i+1}, b)$, so that $c \ne d$. Put $u = s_e(a, a_i, b) = t_e(a, a_i, b)$ and $v = s_e(a, a_{i+1}, a) = t_e(a, a_{i+1}, a)$. The polynomial $f_1(x) = s_e(a, x, b)$ maps $\{a_i, a_{i+1}\}$ onto $\{c, u\}$ and the polynomial $f_2(x) = t_e(a, x, b)$ maps $\{a_i, a_{i+1}\}$ onto $\{u, d\}$, so that $\langle c, d\rangle \in \mathbf{Cg}(a_i, a_{i+1})$. The polynomial $g_1(x) = s_e(a, a_{i+1}, x)$ maps $\{a, b\}$ onto $\{c, v\}$ and the polynomial $g_2(x) = t_e(a, a_{i+1}, x)$ maps $\{a, b\}$ onto $\{v, d\}$, so that $\langle c, d\rangle \in \mathbf{Cg}(a, b)$. Thus $\langle c, d\rangle \in \mathbf{Cg}(a_0, b_0) \cap \mathbf{Cg}(a_i, a_{i+1})$.

(4) implies (1): It is sufficient to prove that if $A \in V$ and α, β, γ are congruences of A with $\alpha \cap \beta = \alpha \cap \gamma = \mathbf{id}_A$ then $\alpha \cap (\beta \vee \gamma) = \mathbf{id}_A$. Suppose $\alpha \cap (\beta \vee \gamma) \ne \mathbf{id}_A$. There are elements $a = a_0, a_1, \dots, a_n = b$ with $a \ne b$, $\langle a, b\rangle \in \alpha$ and $\langle a_i, a_{i+1}\rangle \in \beta \cup \gamma$ for $i < n$. By (4) there is an $i < n$ with $\mathbf{Cg}(a, b) \cap \mathbf{Cg}(a_i, a_{i+1}) \ne \mathbf{id}_A$. But then either $\alpha \cap \beta \ne \mathbf{id}_A$ or $\alpha \cap \gamma \ne \mathbf{id}_A$, a contradiction. $\square$

If condition (3) in 8.1 is satisfied then we also say that V is a meet-semidistributive variety with respect to the *Willard terms* s_e, t_e $(e \in E)$.

8.2. EXAMPLE. The variety of semilattices is a meet-semidistributive variety. For the Willard terms we can take $s_1 = xy$, $t_1 = xyz$, $s_2 = xyz$, $t_2 = yz$.

CHAPTER 8

PROPERTIES OF VARIETIES

1. Amalgamation properties

By an *idempotent* of an algebra A we mean an element $a \in A$ such that $\{a\}$ is a subuniverse of A.

1.1. THEOREM. *The following are equivalent for a variety V:*

(1) *For every subset S of V there exists an algebra $A \in V$ such that every algebra from S is isomorphic to a subalgebra of A.*
(2) *For every pair $A, B \in V$ there exists an algebra $C \in V$ such that both A and B can be embedded into C.*
(3) *Every algebra from V can be embedded into an algebra in V with an idempotent.*
(4) *Whenever H is a family of V-algebras over a set I and an algebra $A \in V$, together with a family of homomorphisms $f_i : H_i \to A$ $(i \in I)$, is a coproduct of H in V, then f_i is injective for every $i \in I$.*

PROOF. Clearly, (4) implies (1), (1) implies (2) and (2) implies (3). It remains to prove that (3) implies (4). Let H be a family of V-algebras over a set I and A together with $f_i : H_i \to A$ be a coproduct in V. For every $i \in I$, H_i is a subalgebra of an algebra $C_i \in V$ such that C_i contains an idempotent e_i. Denote by D the product of the family C_i $(i \in I)$. For $i \in I$ and $a \in H_i$ denote by $g_i(a)$ the element $p \in D$ with $p(i) = a$ and $p(j) = e_j$ for $j \in I \setminus \{i\}$. Clearly, $g_i : H_i \to D$ is an embedding. Since A is a coproduct, there exists a homomorphism $q : A \to D$ with $g_i = qf_i$ for all $i \in I$. Since g_i is injective, it follows that f_i is injective. $\square$

A variety is said to be *extensive* if it satisfies the equivalent conditions of Theorem 1.1.

A class V of algebras is said to have the *amalgamation property* if for any algebras $A, B, C \in V$ and any embeddings $f : A \to B$ and $g : A \to C$ there exists an algebra $D \in V$ and two embeddings $p : B \to D$ and $q : C \to D$ such that $pf = qg$. Clearly, this is equivalent to saying that the pushout of f, g in the category V consists of injective homomorphisms. Also, if V is closed under isomorphic algebras then V has the amalgamation property if and only if for every $A, B, C \in V$ such that A is a subalgebra of B, A is a subalgebra of C and $A = B \cap C$ there exist an algebra $D \in K$, an injective homomorphism f of B into D and an injective homomorphism g of C into D such that f, g coincide on A.

It is not difficult to prove that if a variety V has the amalgamation property, then for every algebra $A \in V$ and for every family of injective homomorphisms $f_i : A \to B_i$ $(i \in I)$ of A into algebras $B_i \in V$, the pushout in the category V of this family of embeddings consists of injective homomorphisms.

A class V of algebras is said to have the *strong amalgamation property* if for any three algebras $A, B, C \in V$ such that A is a subalgebra of both B and C, there exists an algebra $D \in V$ such that both B and C are subalgebras of D. Clearly, this is equivalent to saying that if the pair $f : B \to E$, $g : C \to E$ is a pushout of the pair $\mathbf{id}_A : A \to B$, $\mathbf{id}_A : A \to C$, then f, g are injective homomorphisms and $f(B) \cap g(C) = f(A)$. Also, if V is closed under isomorphic algebras then V has the strong amalgamation property if and only if for every $A, B, C \in V$ such that A is a subalgebra of B, A is a subalgebra of C and $A = B \cap C$ there exists an algebra $D \in K$ such that B, C are subalgebras of D.

Of course, the strong amalgamation property implies the amalgamation property. The variety of distributive lattices is an example of a variety with the amalgamation property which does not have the strong amalgamation property.

1.2. THEOREM. *The variety of all algebras of a given signature has the strong amalgamation property.*

PROOF. It is not difficult to give a construction of a pushout of two injective homomorphisms with the same beginning in the category of all pre-algebras. Its reflection in the variety of all algebras is a pushout in this category. We have given a construction of this reflection in the remark following 3.9.2. $\square$

Let V be a variety. It is easy to see that a morphism of the category V is a monomorphism of this category if and only if it is an injective homomorphism. Also, it is easy to see that if $f : A \to B$ is a homomorphism of A onto B, where $A, B \in V$, then f is an epimorphism of the category V. A variety V is said to have *epimorphisms onto* if every epimorphism $f : A \to B$ of the category V has the property $f(A) = B$.

1.3. EXAMPLE. Let A be the semigroup of integers and B be the semigroup of rational numbers (both with respect to the multiplication of rational numbers). We are going to show that the homomorphism $\mathbf{id}_A : A \to B$ is an epimorphism of the category of semigroups. Let S be a semigroup and $f : B \to C$, $g : B \to C$ be two homomorphisms coinciding on A. For any two integers a, b with $a \neq 0$ we have

$$f(\frac{1}{a}) = f(\frac{1}{a})g(a)g(\frac{1}{a}) = f(\frac{1}{a})f(a)g(\frac{1}{a}) = g(\frac{1}{a}),$$

so that

$$f(\frac{b}{a}) = f(b)f(\frac{1}{a}) = g(b)g(\frac{1}{a}) = g(\frac{b}{a}),$$

and we get $f = g$. So, the variety of semigroups, and also the variety of rings, do not have epimorphisms onto.

Let F be an algebra and X be a subset of A. By an F, X-situation we mean a sixtuple I, J, B, C, r, s such that $I \subseteq X$, $J \subseteq X$, $I \cap J$ is nonempty, $I \cup J = X$, B is the subalgebra of F generated by I, C is the subalgebra of F generated by J, r is a congruence of B, s is a congruence of C and r, s coincide on $B \cap C$. By a solution of an F, X-situation I, J, B, C, r, s we mean a congruence t of F such that $t \cap B^2 = r$ and $t \cap C^2 = s$. By a strong solution of I, J, B, C, r, s we mean a solution t such that if $b \in B$, $c \in C$ and $\langle b, c\rangle \in t$ then there exists an element $a \in B \cap C$ with $\langle b, a\rangle \in r$ and $\langle a, c\rangle \in s$.

1.4. THEOREM. *Let K be a nontrivial variety. The following four conditions are equivalent:*

(1) *K has the amalgamation property*
(2) *the class of finitely generated algebras from K has the amalgamation property*
(3) *whenever F is a K-free algebra over a set X then every F, X-situation has a solution*
(4) *whenever F is a K-free algebra over a finite set X then every F, X-situation has a solution*

Also, the following four conditions are equivalent:

(1') *K has the strong amalgamation property*
(2') *the class of finitely generated algebras from K has the strong amalgamation property*
(3') *whenever F is a K-free algebra over a set X then every F, X-situation has a strong solution*
(4') *whenever F is a K-free algebra over a finite set X then every F, X-situation has a strong solution*

PROOF. We prove the equivalence of the first four conditions and indicate only how to modify the proof to obtain the equivalence of the second four conditions. (1) implies (2) clearly.

Let us prove that (2) implies (4). Let I, J, B, C, r, s be an F, X-situation. Put $A = B \cap C$, so that (as it is easy to see) A is the subalgebra of F generated by $I \cap J$. Put $z = r \cap A^2 = s \cap A^2$. Denote by p_r the canonical homomorphism of B onto B/r and by p_s the canonical homomorphism of C onto C/s. Since $z = \mathbf{ker}(p_r \mathbf{id}_A)$, there exists an injective homomorphism f of A/z into B/r such that the restriction of p_r to A equals $f p_z$. Similarly, there exists an injective homomorphism g of A/z into C/s such that the restriction of p_s to A equals $g p_z$. By the amalgamation property for finitely generated algebras from K there exist an algebra $D \in K$, an injective homomorphism f' of B/r into D and an injective homomorphism g' of C/s into D such that $f'f = g'g$. Since F is K-free, there exists a homomorphism h of F into D with $h(x) = f'(p_r(x))$ for all $x \in I$ and $h(x) = g'(p_s(x))$ for all $x \in J$. (For $x \in I \cap J$ we have $f'(p_r(x)) = f'(f(p_z(x))) = g'(g(p_z(x))) = g'(p_s(x))$.) Clearly, the restriction of h to B equals $f'p_r$ and the restriction of h to C equals $g'p_s$. Put $t = \mathbf{ker}(h)$. Since f' and g' are injective, t extends both r ans s. It is easy to verify that if the range of $f'f$ is the intersection of the ranges of f' and g' then the solution t is strong.

Let us prove that (4) implies (3). Let I, J, B, C, r, s be an F, X-situation. For every finite subset Y of X such that $Y \cap I \cap J$ is nonempty denote by F_Y the subalgebra of F generated by Y; put $I_Y = I \cap Y$, $J_Y = J \cap Y$, $B_Y = B \cap F_Y$, $C_Y = C \cap F_Y$, $r_Y = r \cap B_Y^2$ and $s_Y = s \cap C_Y^2$. By (4), the F_Y, Y-situation $I_Y, J_Y, B_Y, C_Y, r_Y, s_Y$ has at least one solution. Denote by t_Y the intersection of all solutions of this F_Y, Y-situation, so that t_Y is a congruence of F_Y extending both r_Y and s_Y. If $Y_1 \subseteq Y_2$ then the restriction of t_{Y_2} to F_{Y_1} is a congruence of F_{Y_1} extending both r_{Y_1} and s_{Y_1}, so that t_{Y_1} is contained in the restriction of t_{Y_2} to F_{Y_1} which is contained in t_{Y_2}. The union t of the up-directed system of all these t_Y is a congruence of F. Let us prove that it extends r. Let $a, b \in B$. There exists a set Y such that $a, b \in B_Y$. If $\langle a, b\rangle \in r$ then $\langle a, b\rangle \in r_Y$, so that $\langle a, b\rangle \in t_Y$ and $\langle a, b\rangle \in t$. If $\langle a, b\rangle \in t$ then there exists a set Z such that $\langle a, b\rangle \in t_Z$; put $M = Z \cup Y$; clearly, $\langle a, b\rangle \in t_M$, so that $\langle a, b\rangle \in r_M$ and thus $\langle a, b\rangle \in r$. Similarly, t extends s. It is easy to check that if the solutions t_Y are strong then t is strong.

It remains to prove that (3) implies (1). Let $A, B, C \in K$ be such that A is a subalgebra of B, A is a subalgebra of C and $A = B \cap C$. Denote by X the union of the sets B, C and let F be a K-free algebra over X; denote by B' the subalgebra of F generated by B and by C' the subalgebra of F generated by C. Since F is K-free, the identity on B can be extended to a homomorphism h of B' into B and the identity on C can be extended to a homomorphism k of C' into C. The F, X-situation $B, C, B', C', \mathbf{ker}(h), \mathbf{ker}(k)$ has a solution r. Denote by p_r the canonical homomorphism of F onto F/r, by f the restriction of p_r to B and by g the restriction of p_r to C. It is easy to see that f is an injective homomorphism of B into F/r, g is an injective homomorphism of C into F/r and that f, g, p_r coincide on A. If, moreover, r is a strong solution then the range of the restriction of f to A is the intersection of the ranges of f and g. □

1.5. Theorem. *Let V be a variety. Consider the following three conditions:*

(1) *V has the strong amalgamation property.*
(2) *Every monomorphism of the category V is an equalizer of a pair of V-morphisms.*
(3) *V has epimorphisms onto.*

We have (1) $\Rightarrow$ (2) $\Rightarrow$ (3).

Proof. (1) implies (2): Let $f : A \to B$ be a monomorphism of V, i.e., an injective homomorphism. Let the pair $g : B \to C$, $h : B \to C$ be a pushout of the pair $f : A \to B$, $f : A \to B$. We have $gf = hf$; it follows from the strong amalgamation property that $f(A) = \{b \in B : g(b) = h(b)$. Consequently, f is an equalizer of g, h.

(2) implies (3): Let $f : A \to B$ be an epimorphism of the category V. Denote by C the subalgebra of B with the underlying set $f(A)$. The monomorphism $\mathbf{id}_C : C \to B$ is an equalizer of a pair of morphisms $g : B \to D$,

$h : B \to D$. We have $g\mathbf{id}_C = h\mathbf{id}_C$, so that $gf = g\mathbf{id}_C f = h\mathbf{id}_C f = hf$. Since f is an epimorphism, we get $g = h$. But then, $C = B$. □

2. Discriminator varieties and primal algebras

The *discriminator function* on a set A is the ternary operation d on A defined by

$$d(x,y,z) = \begin{cases} x \text{ if } x \neq y \\ z \text{ if } x = y. \end{cases}$$

The *switching function* on A is the quaternary operation s on A defined by

$$s(x,y,z,u) = \begin{cases} z \text{ if } x = y \\ u \text{ if } x \neq y. \end{cases}$$

It is easy to check that

$$s(x,y,z,u) = d(d(x,y,z), d(x,y,u), u) \quad \text{and} \quad d(x,y,z) = s(x,y,z,x).$$

A ternary term t is said to be a *discriminator term* for a class K of algebras if for any $A \in K$, t^A is the discriminator function on A. Similarly, a quaternary term is said to be a *switching term* for K if it represents the switching function on any algebra from K. It follows that K has a discriminator term if and only if it has a switching term.

An algebra is said to be *quasiprimal* if it is finite and has a discriminator term. A variety V is said to be a *discriminator variety* if there exists a term t such that V is generated by all its algebras for which t is a discriminator term. We also say that V is a discriminator variety with respect to t.

2.1. Theorem. *Let $V = \mathbf{HSP}(K)$ where K is a class of algebras such that there exists a term $t(x,y,z)$ serving as a discriminator term for all algebras in K. Then V is an arithmetical variety; a nontrivial algebra $A \in V$ is subdirectly irreducible if and only if it is simple if and only if t serves as a discriminator term for A if and only if $A \in \mathbf{ISP}_U(K)$.*

Proof. A discriminator term satisfies the equations 7.5.1, so the variety generated by K is arithmetical. If an algebra A has a discriminator term then the switching function s is a term operation of A and for any elements $x,y,z,u \in A$ with $x \neq y$ we have $\langle z,u\rangle = \langle s(x,x,z,u), s(x,y,z,u)\rangle \in \mathbf{Cg}(x,y)$, so that A is simple. Clearly, the class of algebras for which t is a discriminator term is closed under ultraproducts and subalgebras. So, all the algebras in $\mathbf{SP}_U(K)$ are simple. By 7.7.1, all subdirectly irreducible algebras from V belong to $\mathbf{HSP}_U(K)$; a homomorphic image of a simple algebra A is either trivial or isomorphic to A. □

2.2. Theorem. (Pixley [71]) *A finite algebra A is quasiprimal if and only if it generates an arithmetical variety and every subalgebra of A is either simple or trivial.*

PROOF. The direct implication follows from 2.1. In order to prove the converse, by 7.5.3 it is sufficient to show that the discriminator function d on A preserves subalgebras of A^2. Let C be a subalgebra of A^2. Denote by A_1 and A_2 the image of C under the first and the second projection, respectively, so that A_1 and A_2 are subalgebras of A and $C \subseteq A_1 \times A_2$ is a subdirect product. By 7.1.7 there exist an algebra $D \in \mathbf{HSP}(A)$, a homomorphism f of A_1 onto D and a homomorphism g of A_2 onto D such that $C = \{\langle x, y\rangle : f(x) = g(y)\}$. Since subalgebras of A are either simple or trivial, either f, g are isomorphisms or D is trivial. In the first case we have $C = \{\langle x, h(x)\rangle : x \in A_1\}$ where h is the isomorphism $g^{-1}f$ of A_1 onto A_2, and in the second case $C = A_1 \times A_2$. Let $\langle a, a'\rangle, \langle b, b'\rangle, \langle c, c'\rangle$ be three elements of C. If $C = \{\langle x, h(x)\rangle : x \in A_1\}$ then $a = b$ if and only if $a' = b'$, so that $d(\langle a, a'\rangle, \langle b, b'\rangle, \langle c, c'\rangle) = \langle d(a, b, c), d(a', b', c')\rangle$ is either $\langle a, a'\rangle \in C$ (if $a \neq b$) or $\langle c, c'\rangle \in C$ (if $a = b$). If $C = A_1 \times A_2$ then $d(\langle a, a'\rangle, \langle b, b'\rangle, \langle c, c'\rangle) \in C$ is obvious. □

2.3. THEOREM. (McKenzie [75]) *Let V be a variety and t be a ternary term.*

(1) *V is a discriminator variety with respect to t if and only if it satisfies the following equations:*

(a) $t(x, y, y) = x$, $t(x, y, x) = x$, $t(x, x, y) = y$,
(b) $t(x, t(x, y, z), y) = y$,
(c) $t(x, y, F(z_1, \dots, z_n)) = t(x, y, F(t(x, y, z_1), \dots, t(x, y, z_n)))$ *for any n-ary F in the signature.*

(2) *If V is a discriminator variety with respect to t then for any algebra $A \in V$ and any elements a, b, c, d of A, $\langle c, d\rangle \in \mathbf{Cg}(a, b)$ if and only if $t(a, b, c) = t(a, b, d)$.*

PROOF. It is easy to check that if V is a discriminator variety with respect to t then the equations (a), (b) and (c) are satisfied in all algebras of the class generating V as a variety and thus in all algebras of V.

Let V be a variety satisfying the equations (a), (b) and c. For an algebra $A \in V$ and elements $a, b \in A$ define a binary relation $\gamma(a, b)$ on A by $\langle x, y\rangle \in \gamma(a, b)$ if and only if $t(a, b, x) = t(a, b, y)$. Then $\gamma(a, b)$ is clearly an equivalence and it follows from (c) that it is a congruence of A. By (a) we have $\langle a, b\rangle \in \gamma(a, b)$, so that $\mathbf{Cg}(a, b) \subseteq \gamma(a, b)$. Let $\langle c, d\rangle \in \gamma(a, b)$. Then $\langle t(a, b, c), c\rangle = \langle t(a, b, c), t(a, a, c)\rangle \in \mathbf{Cg}(a, b)$, $\langle t(a, b, d), d\rangle = \langle t(a, b, d), t(a, a, d)\rangle \in \mathbf{Cg}(a, b)$ by (a) and $t(a, b, c) = t(a, b, d)$, so that $\langle c, d\rangle \in \mathbf{Cg}(a, b)$. Thus $\gamma(a, b) = \mathbf{Cg}(a, b)$ and we have proved (2).

It remains to prove that if V satisfies (a), (b) and (c) then t is a discriminator function on any subdirectly irreducible algebra A in V (since the class of subdirectly irreducible algebras generates V). For this, by (a), it is sufficient to prove that if $c, d \in A$ and $c \neq d$ then $t(c, d, x) = c$ for all $x \in A$. There exist two distinct elements $a, b \in A$ such that $\mathbf{Cg}(a, b)$ is the monolith of A. If $t(a, b, x) \neq a$ for some element $x \in A$ then $\langle a, b\rangle \in \mathbf{Cg}(a, t(a, b, x))$ which means $t(a, t(a, b, x), a) = t(a, t(a, b, x), b)$; but $t(a, t(a, b, x), a) = a$ and $t(a, t(a, b, x), b) = b$ by (a) and (c), a contradiction. Consequently, $t(a, b, x) = a$

for all $x \in a$. Thus $\mathbf{Cg}(a,b) = A^2$ and then also $\mathbf{Cg}(c,d) = A^2$ whenever $c \neq d$. We get $t(c,d,x) = c$ for all $x \in A$. □

An algebra is said to be *primal* if it is finite and every n-ary operation on A, for any $n \geq 1$, is a term operation of A.

2.4. Theorem. (Foster and Pixley [64],[64a]) *A finite algebra A is primal if and only if it is quasiprimal, has no proper subalgebras and has no automorphisms except identity.*

Proof. The direct implication is clear. In order to prove the converse, follow the proof of 2.2 and observe that $A_1 = A_2 = A$ and $h = \mathbf{id}_A$. □

2.5. Example. (1) The two-element Boolean algebra is primal.

(2) For every prime number p, the finite field of integers modulo p is primal. For the discriminator term we can take $d(x,y,z) = x(x-y)^{p-1} + z(1-(x-y)^{p-1})$.

(3) For every $n \geq 2$, the n-element *Post algebra* is the algebra of the signature of Boolean algebras, with the underlying set $\{0,1,\dots,n-1\}$ and operations defined in this way: it is a lattice with respect to the ordering $0 < n-1 < n-2 < \dots < 2 < 1$; $0' = 1$, $1' = 2$, $\dots$, $(n-1)' = 0$. This is a primal algebra. For the discriminator term we can take $d(x,y,z) = (g(x,y) \wedge x) \vee (g(g(x,y),1) \wedge z)$ where $g(x,y) = (\bigwedge_{1 \leq i < n}(\bigwedge_{1 \leq j \leq n}(x^j \vee y^j))^i)'$ and x^i means i applications of $'$ to x. (Observe that $\bigwedge_{1 \leq j \leq n}(x^j \vee y^j) = 0$ if and only if $x = y$; $(\bigwedge_{1 \leq i < n} x^i)'$ is 0 for $x = 0$ and 1 for $x \neq 0$; and $g(x,y)$ is 0 for $x = y$ and 1 for $x \neq y$).

2.6. Theorem. (Foster [53a]) *Let A be a primal algebra. Then the variety generated by A is just the class of algebras isomorphic to a Boolean power of A.*

Proof. By 2.1 it is sufficient to prove that every subdirect power of A is isomorphic to a Boolean power of A. Let a nontrivial subalgebra C of A^I be a subdirect power of A. Clearly, for any $a \in A$ the constant mapping $c_a \in A^I$ with value a belongs to C. The switching function s on A is a term operation of A. For $f,g,p,r \in C$ we have $\mathbf{e}(f = g) \cup \mathbf{e}(p = r) = \mathbf{e}(s(f,g,p,r) = p)$, $\mathbf{e}(f = g) \cap \mathbf{e}(p = r) = \mathbf{e}(s(f,g,p,f) = s(f,g,r,g))$ and $I \smallsetminus \mathbf{e}(f = g) = \mathbf{e}(s(f,g,c_a,c_b) = c_b)$ where a,b are two distinct elements of A. Consequently, the set B of all subsets $\mathbf{e}(f = g)$ of I with $f,g \in C$ is a subalgebra of the Boolean algebra of all subsets of I. Put $X = B^*$. For $f \in C$ and $U \in X$, I is the disjoint union of its finitely many subsets $f^{-1}(a)$ with $a \in A$, so there exists precisely one $a \in A$ with $f^{-1}(a) \in U$; denote this a by $H(f,U)$. Define a mapping h of C into A^X by $h(f)(U) = H(f,U)$. Clearly, all constants of A^X are in $h(C)$. For $f,g \in C$ we have $\mathbf{e}(h(f) = h(g)) = \{U \in X : H(f,U) = H(g,U)\} = \{U \in X : f^{-1}(a) \in U \text{ and } g^{-1}(a) \in U \text{ for some } a \in A\} = \{U \in X : \mathbf{e}(f = g) \in U\}$, which is a clopen set. Thus the sets $\mathbf{e}(f = g)$ with $f,g \in C$ are precisely all the clopen subsets of X. For $f,g \in C$ and a clopen subset $N = \{U \in X : \mathbf{e}(p = r)\}$ of X put $q = (f \restriction \mathbf{e}(p = r)) \cup (g \restriction (I \smallsetminus \mathbf{e}(p = r)))$. Since $q = s(p,r,f,g)$, we have $q \in C$; since $\mathbf{e}(h(q) = h(f)) = \{U \in X : \mathbf{e}(q =$

$f) \in U\} \supseteq \{U \in X : \mathbf{e}(p = r) \in U\} = N$ and $\mathbf{e}(h(q) = h(g)) = \{U \in X : \mathbf{e}(q = g) \in U\} \supseteq \{U \in X : I \smallsetminus \mathbf{e}(p = r) \in U\} = X \smallsetminus N$, we have $h(q) = (h(f) \restriction N) \cup (h(g) \restriction (X \smallsetminus N))$. This proves $h(C) = A[B]^*$. Clearly, h is a bijection and it is not difficult to check that h is a homomorphism. □

Let A be an algebra and B_1, B_2 be two Boolean algebras. For any homomorphism $h : B_1 \to B_2$ we get a homomorphism $\bar{h} : A[B_1]^* \to A[B_2]^*$ if we put (for any $f \in A[B_1]^*$) $\bar{h}(f) = fh^*$ where h^* is the continuous mapping of B_2^* into B_1^* corresponding to h (see 4.4.3). It can be proved that if A is a nontrivial primal algebra then every homomorphism of $A[B_1]^*$ into $A[B_2]^*$ can be obtained from a homomorphism of B_1 into B_2 in this way, and $A[B_1]^* \simeq A[B_2]^*$ if and only if $B_1 \simeq B_2$.

2.7. Theorem. *Let $V = \mathbf{HSP}(K)$ be a discriminator variety and $t(x, y, z)$ be a discriminator term for all algebras in K. Then every algebra from V is isomorphic to a Boolean product of algebras that either belong to K or are trivial.*

Proof. Let A be a nontrivial algebra from V. By 2.1, A is isomorphic to a subdirect product of a family of simple algebras $S_i \in \mathbf{SP}_U(K)$ $(i \in I)$; we can assume that A is equal to the subdirect product. Denote by s the switching term for algebras in K (and thus for all algebras in $\mathbf{SP}_U(K)$). For $x, y, z, u \in A$ we have $\mathbf{Cg}(x, y) = \{\langle z, u\rangle : \mathbf{e}(x = y) \subseteq \mathbf{e}(z = u)\}$, because the right side is easily seen to be a congruence containing $\langle x, y\rangle$ and if $\mathbf{e}(x = y) \subseteq \mathbf{e}(z = u)$ then $\langle z, u\rangle = \langle s(x, x, z, u), s(x, y, z, u)\rangle$. We have $\mathbf{Cg}(x, y) \vee \mathbf{Cg}(z, u) = \mathbf{Cg}(t(x, y, z), t(y, x, u))$, since (as it is easy to check) $\mathbf{e}(t(x, y, z) = t(y, x, u)) = \mathbf{e}(x = y) \cap \mathbf{e}(z = u)$. Also, $\mathbf{Cg}(x, y) \cap \mathbf{Cg}(z, u) = \mathbf{Cg}(s(x, y, z, u), z)$ since $\mathbf{e}(s(x, y, z, u) = z) = \mathbf{e}(x = y) \cup \mathbf{e}(z = u)$. Hence the set L of finitely generated congruences of A (equal to the set of principal congruences of A) is a sublattice of $\mathbf{Con}(A)$. The lattice is distributive and the congruences permute, since V is an arithmetical variety. In order to prove that it is relatively complemented, it is sufficient to show that $\mathbf{Cg}(s(z, u, x, y), y)$ is a complement of $\mathbf{Cg}(z, u)$ in the interval $[\mathbf{id}_A, \mathbf{Cg}(x, y) \vee \mathbf{Cg}(z, u)]$. We have $\mathbf{Cg}(z, u) \cap \mathbf{Cg}(s(z, u, x, y), y) = \mathbf{Cg}(s(z, u, s(z, u, x, y), y), s(z, u, x, y)) = \mathbf{id}_A$, since (as it is easy to check) $s(z, u, s(z, u, x, y), y) = s(z, u, x, y)$. We have $\mathbf{Cg}(z, u) \vee \mathbf{Cg}(s(z, u, x, y), y) = \mathbf{Cg}(t(z, u, s(z, u, x, y)), t(u, z, y))$ since $\mathbf{e}(t(z, u, s(z, u, x, y)) = t(u, z, y)) = \mathbf{e}(x = y) \cap \mathbf{e}(z = u)$. Now we can use 4.5.1. □

Let A be an algebra of signature σ. Denote by $\sigma + A$ the extension of σ by constants C_a, one constant for each element a of A. We denote by A_{+A} the algebra of signature $\sigma + A$ such that A is a reduct of A_{+A} and each constant C_a is interpreted by the element a in A_{+A}.

An algebra A is called *functionally complete* if A_{+A} is a primal algebra, i.e., if every operation on the set A of positive arity is a polynomial of A.

2.8. THEOREM. (Werner [74]) *Let A be a nontrivial algebra generating a congruence permutable variety. Then A is functionally complete if and only if* $\mathbf{Con}(A^2)$ *is the four-element Boolean lattice.*

PROOF. Put $B = A_{+A}$ and denote by $\mathbf{2}$ the two-element lattice. Let A be functionally complete, so that B is primal. Since B generates a congruence distributive variety, it follows from 7.3.7 that congruences of B^2 are precisely the product congruences. Since B is simple, $\mathbf{Con}(B^2) \simeq \mathbf{2}^2$. Clearly, $\mathbf{Con}(A^2) = \mathbf{Con}(B^2)$.

Conversely, let $\mathbf{Con}(A^2) \simeq \mathbf{2}^2$. Then also $\mathbf{Con}(B^2) \simeq \mathbf{2}^2$. The algebra B is simple, because otherwise B^2 would have product congruences other than the obvious four ones. Clearly, B has no proper subalgebras and no non-identical automorphisms. It follows from 7.1.7 that the only subalgebras of B^2 that are subdirect products are A^2 and D, where D is the subalgebra with underlying set $\mathbf{id}_A$. Since $\mathbf{Con}(B^2) \simeq \mathbf{2}^2$, the only congruences of B^2 are the four product congruences. Since $D \simeq B$, D has only two congruences and these are again the product congruences. By 7.3.6 it follows that $\{B\}$ is totally skew free, so that $\mathbf{Con}(B^n) \simeq \mathbf{2}^n$ for all n.

Denote by F the free algebra in $\mathbf{HSP}(B)$ over $\{x, y, z\}$. By 6.4.2 we have $F \in \mathbf{ISP}(B)$ and F is isomorphic to a subalgebra of B^k for some positive integer k. Since B has no proper subalgebras, this subalgebra is a subdirect power of B. By 7.1.8 it follows that $F \simeq B^n$ for some n. Consequently, $\mathbf{Con}(F) \simeq \mathbf{2}^n$ and hence the lattice $\mathbf{Con}(F)$ is distributive. By 7.2.1 the variety $\mathbf{HSP}(B)$ is congruence distributive; since it is also congruence permutable, it is arithmetical. By 2.2 and 2.4 it follows that B is primal, i.e., A is functionally complete. $\square$

3. Dual discriminator varieties

The *dual discriminator* function on a set A is the ternary operation D on A defined by

$$D(x, y, z) = \begin{cases} x \text{ if } x = y \\ z \text{ if } x \neq y. \end{cases}$$

A ternary term t is said to be a *dual discriminator term* for a class K of algebras if for any $A \in K$, t^A is the dual discriminator function on A. A variety V is said to be a *dual discriminator variety* if there exists a term t such that V is generated by all its algebras for which t is a dual discriminator term. We also say that V is a dual discriminator variety with respect to t. The following results belong to Fried and Pixley [79].

3.1. THEOREM. *Let d be the discriminator and D be the dual discriminator function on a set A. Then $D(x, y, z) = d(x, d(x, y, z), z)$ for all $x, y, z \in A$. Consequently, every discriminator variety is a dual discriminator variety.*

PROOF. It is obvious. $\square$

3.2. Example. The term $(x\wedge y)\vee(y\wedge z)\vee(x\wedge z)$ is a dual discriminator term for the two-element lattice. Consequently, the variety of distributive lattices is a dual discriminator variety. It is not a discriminator variety.

3.3. Theorem. *Let $V = \mathbf{HSP}(K)$ where K is a class of algebras such that there exists a term $t(x,y,z)$ serving as a dual discriminator term for all algebras in K. Then V is a congruence distributive variety; a nontrivial algebra $A \in V$ is subdirectly irreducible if and only if it is simple if and only if t serves as a dual discriminator term for A if and only if $A \in \mathbf{ISP}_U(K)$. A dual discriminator variety is a discriminator variety if and only if it is congruence permutable.*

Proof. It is easy to check that the dual discriminator term is a ternary majority term, so that the variety V is congruence distributive (and satisfies the Chinese remainder theorem) by 7.4.4. If an algebra A has a dual discriminator term t then for any elements $a,b,c \in A$ with $a \neq b$ we have $\langle a,c\rangle = \langle t(a,a,c), t(a,b,c)\rangle \in \mathbf{Cg}(a,b)$, so that A is simple. Clearly, the class of algebras for which t is a dual discriminator term is closed under ultraproducts and subalgebras. So, all the algebras in $\mathbf{ISP}_U(K)$ are simple. By 7.7.1, all subdirectly irreducible algebras from V belong to $\mathbf{HSP}_U(K)$; a homomorphic image of a simple algebra A is either trivial or isomorphic to A. $\square$

3.4. Theorem. *Let A be a finite algebra with $|A| \geq 3$. Then A is functionally complete if and only if the dual discriminator function on A is a polynomial of A.*

Proof. The direct implication is obvious. Let the dual discriminator function on A be a polynomial of A. There exists a ternary term t of the signature $\sigma + A$ (σ being the signature of A) such that the corresponding term operation on the algebra $B = A_{+A}$ is the dual discriminator function.

Denote by D the subalgebra of B^2 with the underlying set $\mathbf{id}_A$. Let S be an arbitrary subalgebra of A. Clearly, $D \subseteq S$. Let $S \neq D$, so that $\langle a,b\rangle \in S$ for two elements a,b with $a \neq b$. For all elements $a \in A$ we have $\langle a,c\rangle = \langle t(a,a,c), t(a,b,c)\rangle = t(\langle a,a\rangle, \langle a,b\rangle, \langle c,c\rangle) \in S$. Consequently, for all elements $c,d \in A$ with $c \neq a$ we have $\langle d,c\rangle = \langle t(c,a,d), t(c,c,d)\rangle = t(\langle c,c\rangle, \langle a,c\rangle, \langle d,d\rangle) \in S$. Since $|A| \geq 3$, it follows that all elements of A^2 belong to S. Thus B^2 has only two subalgebras, D and B^2. Clearly, both of them are closed under any operation on B, so that any operation on B is a polynomial of B according to 7.5.3. This means that the algebra B is primal and hence A is functionally complete. $\square$

3.5. Example. The assumption $|A| \geq 3$ in 3.4 is essential: according to 3.2, the two-element lattice has the dual discriminator term but it is not functionally complete (since every polynomial of a lattice is order preserving).

3.6. Theorem. *Let V be a variety and t be a ternary term.*

(1) V is a dual discriminator variety with respect to t if and only if it satisfies the following equations:

(a) $t(x,y,y) = y$, $t(x,y,x) = x$, $t(x,x,y) = x$,
(b) $t(x,y,t(x,y,z)) = t(x,y,z$,
(c) $t(z,t(x,y,z),t(x,y,u)) = t(x,y,z)$,
(d) $t(x,y,F(z_1,\dots,z_n)) = t(x,y,F(t(x,y,z_1),\dots,t(x,y,z_n)))$ *for any n-ary F in the signature.*

(2) *If V is a dual discriminator variety with respect to t then for any algebra $A \in V$ and any elements a,b,c,d of A the following are true:*

(i) $\langle c,d\rangle \in \mathbf{Cg}(a,b)$ *if and only if* $t(c,d,x) = t(c,d,t(a,b,x))$ *for all* $x \in A$;
(ii) $\mathbf{Cg}(a,b) \cap \mathbf{Cg}(c,d) = \mathbf{Cg}(t(a,b,c),t(a,b,d))$,
(iii) $\mathbf{Cg}(a,b)$ *has a complement r in the lattice* $\mathbf{Con}(A)$; *we have* $\langle x,y\rangle \in r$ *if and only if* $t(a,b,x) = t(a,b,y)$.

Proof. It is easy to check that if V is a dual discriminator variety with respect to t then the equations (a), (b), (c) and (d) are satisfied in all algebras of the class generating V as a variety and thus in all algebras of V.

Let V be a variety satisfying the equations (a), (b), (c) and d. By (a), V is congruence distributive. For an algebra $A \in V$ and elements $a,b \in A$ define a binary relation $\gamma(a,b)$ on A by $\langle x,y\rangle \in \gamma(a,b)$ if and only if $t(a,b,x) = t(a,b,y)$. (In the same way as in the proof of 2.3; but now $\gamma(a,b)$ is not $\mathbf{Cg}(a,b)$.) Clearly, $\gamma(a,b)$ is an equivalence and it follows from (d) that it is a congruence of A. By (b) we have $\langle t(a,b,c),c\rangle \in \gamma(a,b)$ for all $a,b,c \in A$.

Claim. $\gamma(z,t(x,y,z) \neq \mathbf{id}_A$ *whenever* $x \neq y$. By (c) we have $t(z,t(x,y,z), t(x,y,u)) = t(z,t(x,y,z),t(x,y,z))$ and thus $\langle t(x,y,z),t(x,y,u)\rangle \in \gamma(z,t(x, y,z))$ for all $x,y,z,u \in A$. If $\gamma(z,t(x,y,z)) = \mathbf{id}_A$ then $t(x,y,z) = t(x,y,u)$ for all u, so that $x = t(x,y,x) = t(x,y,y) = y$.

Let us prove (ii). Denote by R the set of the congruences r of A such that A/r is subdirectly irreducible and $\mathbf{Cg}(t(a,b,c),t(a,b,d)) \subseteq r$. Let $r \in R$. Since t is a dual discriminator function on A/r and $t(a/r,b/r,c/r) = t(a/r,b/r,d/r)$, we have either $a/r = b/r$ or $c/r = d/r$. Thus for any $r \in R$, either $\langle a,b\rangle \in r$ or $\langle c,d\rangle \in r$. By 3.5.4, $\mathbf{Cg}(t(a,b,c),t(a,b,d))$ is the intersection of the congruences $r \in R$. But every $r \in R$ is above either $\mathbf{Cg}(a,b)$ or $\mathbf{Cg}(c,d)$. Thus $\mathbf{Cg}(a,b) \cap \mathbf{Cg}(c,d) = \mathbf{Cg}(t(a,b,c),t(a,b,d))$.

Let us prove (iii). For all $c,d \in A$ we have $\langle c,t(a,b,c)\rangle \in \gamma(a,b)$, $\langle t(a,b,c), t(a,a,c)\rangle \in \mathbf{Cg}(a,b)$, $t(a,a,c) = a = t(a,a,d)$, $\langle t(a,a,d),t(a,b,d)\rangle \in \mathbf{Cg}(a,b)$ and $\langle t(a,b,d),d\rangle \in \gamma(a,b)$, so that $\langle c,d\rangle \in \mathbf{Cg}(a,b) \vee \gamma(a,b)$ and thus $\mathbf{Cg}(a,b) \vee \gamma(a,b) = A^2$. If $\langle c,d\rangle \in \mathbf{Cg}(a,b) \cap \gamma(a,b)$ then $\langle c,d\rangle \in \mathbf{Cg}(a,b) \cap \mathbf{Cg}(c,d) = \mathbf{Cg}(t(a,b,c),t(a,b,d))$ by (ii); but $t(a,b,c) = t(a,b,d)$ and so $c = d$. We get $\mathbf{Cg}(a,b) \cap \gamma(a,b) = \mathbf{id}_A$.

Let us prove (i). By (iii) we have $\langle c,d\rangle \in \mathbf{Cg}(a,b)$ if and only if $\mathbf{Cg}(c,d) \subseteq \mathbf{Cg}(a,b)$ if and only if $\gamma(a,b) \subseteq \gamma(c,d)$ (since the congruence lattice is distributive). If $\langle c,d\rangle \in \mathbf{Cg}(a,b)$ then $\gamma(a,b) \subseteq \gamma(c,d)$; we have $t(a,b,x) = t(a,b,t(a,b,x))$ by (b) and so $t(c,d,x) = t(c,d,t(a,b,x))$ for all $x \in A$. Conversely, let $t(c,d,x) = t(c,d,t(a,b,x))$ for all $x \in A$. Then for all $x,y \in A$

$t(a,b,x) = t(a,b,y)$ implies $t(c,d,x) = t(c,d,y)$, i.e., $\langle x,y\rangle \in \gamma(a,b)$ implies $\langle x,y\rangle \in \gamma(c,d)$, so that $\gamma(a,b) \subseteq \gamma(c,d)$ and hence $\langle c,d\rangle \in \mathbf{Cg}(a,b)$.

It remains to prove that if V satisfies (a), (b), (c) and (d) then t is a dual discriminator function on any subdirectly irreducible algebra A in V. For this, by (a), it is sufficient to prove that if $c,d \in A$ and $c \neq d$ then $t(c,d,x) = c$ for all $x \in A$. By the Claim it is sufficient to prove that if $c \neq d$ then $\gamma(c,d) = \mathbf{id}_A$. Suppose $\gamma(c,d) \neq \mathbf{id}_A$. There exist two distinct elements $a,b \in A$ such that $\mathbf{Cg}(a,b)$ is the monolith of A. We have $\langle a,b\rangle \in \gamma(c,d)$, so that $t(c,d,a) = t(c,d,b)$. Put $e = t(c,d,a) = t(c,d,b)$. By the Claim we have $\gamma(a,e) = \gamma(a,t(c,d,a)) \neq \mathbf{id}_A$ and $\gamma(b,e) = \gamma(b,t(c,d,b)) \neq \mathbf{id}_A$. Hence $\langle a,b\rangle \in \gamma(a,e) \cap \gamma(b,e)$, so that $a = t(a,e,a) = t(a,e,b)$ and $b = t(b,e,b) = t(b,e,a)$. Since $a \neq b$, either $e \neq a$ or $e \neq b$. If $e \neq a$ then taking $x = a$, $y = e$, $z = b$ in the Claim yields $\gamma(b,a) = \gamma(b,t(a,e,b)) \neq \mathbf{id}_A$; hence $\langle a,b\rangle \in \gamma(b,a)$ which implies $a = t(b,a,a) = t(b,a,b) = b$, a contradiction. If $e \neq b$ then taking $x = b$, $y = e$, $z = a$ in the Claim gives $\gamma(a,b) = \gamma(a,t(b,e,a)) \neq \mathbf{id}_A$ which implies $\langle a,b\rangle \in \gamma(a,b)$ and thus $a = t(a,b,a) = t(a,b,b) = b$, a contradiction again. Thus $\gamma(c,d) = \mathbf{id}_A$. □

4. Bounded varieties

An equational theory E is said to be *bounded* if there is a finite set of terms S such that every term is E-equivalent to a term similar to a term in S. (Recall that two terms u,v are similar if $v = h(u)$ for an automorphism h of the algebra of terms.) A variety is said to be bounded if the corresponding equational theory is bounded.

4.1. Theorem. *The set of bounded varieties of signature σ is an ideal in the lattice of all varieties of σ-algebras. The following are true for any bounded variety V:*

(1) *V has only finitely many subvarieties*
(2) *V is finitely generated*
(3) *If the signature is finite then V is finitely based*

Proof. Clearly, a subvariety of a bounded variety is itself bounded. Let V be the join of two bounded varieties V_1 and V_2. There are two finite sets S_1 and S_2 of terms such that every term is V_1-equivalent with a term similar to a term from S_1 and also V_2-equivalent with a term similar to a term from V_2. For each pair $\langle u,v\rangle \in S_1 \times S_2$ select, if possible, a term t that is equivalent modulo V_1 with a term similar to a term from S_1 and equivalent modulo V_2 with a term similar to a term from S_2. Denote by S the set of all terms t selected in this way. Then S is a finite set witnessing the boundedness of V.

Let V be a bounded variety and S be a finite set of terms such that every term is V-equivalent with a term similar to a term from S.

(1) Denote by $x_1,\dots,x_k$ all the variables occurring in some term from S; take pairwise distinct variables $y_1,\dots,y_k$ not belonging to $\{x_1,\dots,x_k\}$ and denote by S' the (finite) set of terms that are similar to a term from S and contain

no other variables than those belonging to $\{x_1, \dots, x_n, y_1, \dots, y_n\}$. Clearly, every equation is equivalent modulo V to an equation from $S' \times S'$. Consequently, every subvariety of V is based (modulo V) on a subset of $S' \times S'$.

(2) It is easy to see that V is locally finite. For every proper subvariety W of V there exists a finitely generated, and thus finite, algebra in $V \smallsetminus W$; select one and denote it by A_W. Clearly, V is generated by the direct product of the finite algebras A_W with W running over all proper subvarieties of V; by (1), this is a direct product of finitely many finite algebras.

(3) There exist positive integers c and d such that every operation symbol of σ is of arity at most c and every term is V-equivalent with a term of length at most d. Take cd pairwise distinct variables $x_1, \dots, x_{cd}$. Denote by E the set of equations satisfied in V, both sides of which are of length at most cd and contain no other variables than $x_1, \dots, x_{cd}$. Then E is finite and we claim that V is based on E. We only need to prove that for every term s of length greater than cd there exists a shorter term t such that $\langle s, t \rangle$ is a consequence of E. Evidently, s has a subterm u such that its length k satisfies $d < k \leq cd$. Then there is a term v of length at most d such that $\langle u, v \rangle$ is satisfied in V. Now $\langle u, v \rangle$ is a consequence of E and then also $\langle s, t \rangle$ is a consequence of E, where t is obtained from s by replacing one occurrence of u by v; the term t is shorter than s. $\square$

The notion of an address (a finite sequence of elementary addresses) and the related notation introduced in Chapter 6 shoud be recalled. By a *direction* we mean an infinite sequence (indexed by nonnegative integers) of elementary addresses. Addresses and directions can be concatenated: for an address $e = a_0 \dots a_{k-1}$ and a direction $d = b_0 b_1 \dots$, ed is the direction $a_0 \dots a_{k-1} b_0 b_1 \dots$.

For an address e and a natural number n, we define an address e^n by induction as follows: e^0 is the empty address; $e^{n+1} = ee^n$. For a nonempty address e, the unique direction that extends e^n for every natural number n is denoted by e^ω. A direction h is said to be *eventually periodic* if $h = ef^\omega$ for some addresses e and f; it is said to be *periodic* if $h = f^\omega$.

Let t be a term and h be a direction. We say that h is *traversible* in the direction h if $t[e]$ is a variable for some initial segment e of h. This e, if it exists, is unique and will be denoted by $\tau_t[h]$; we also denote $t[e]$ as $t[h]$.

By a *coherent triple* we mean a triple $\langle J, m, d \rangle$ where J is a finite set of directions, m is a mapping of J into the set of nonnegative integers and d is a mapping of J into the set of positive integers, such that the following three conditions are satisfied:

(1) Whenever $eh \in J$ then $h \in J$
(2) If $h \in J$ then $h = ef^\omega$ for some e, f such that e is of length $m(h)$ and f is of length $d(h)$
(3) If $h = ah' \in J$ where a is an elementary address then $m(h) \leq m(h')+1$ and $d(h')$ is a multiple of $d(h)$

For every coherent triple $\langle J, m, d\rangle$ we denote by $\Theta(J, m, d)$ the set of equations defined by $\langle u, v\rangle \in \Theta(J, m, d)$ if and only if the following two conditions are satisfied:

(1) For every $h \in J$, u is traversible in the direction h if and only if v is traversible in the direction h
(2) If $h \in J$ and u, v are traversible in the direction h then $u[h] = v[h]$ and either $u = v$ or else $\lambda(\tau_u(h)) \equiv \lambda(\tau_v(h)) \bmod d(h)$ and both $m(h) \leq \lambda(\tau_u(h))$ and $m(h) \leq \lambda(\tau_v(h))$

Observe that if the signature contains only one operation symbol which is of positive arity, then the first condition is always satisfied because then every term is traversible in every direction.

4.2. THEOREM. *$\Theta(J, m, d)$ is an equational theory for every coherent triple $\langle J, m, d\rangle$.*

PROOF. Put $\Theta = \Theta(J, m, d)$. Evidently, Θ is an equivalence on the set of terms.

In order to prove that Θ is a congruence, let F be an n-ary operation symbol and $\langle u_1, v_1\rangle, \dots, \langle u_n, v_n\rangle \in \Theta$; put $u = F(u_1, \dots, u_n)$ and $v = F(v_1, \dots, v_n)$. Let $h \in J$. We have $h = \langle G, i\rangle h'$ for an elementary address $\langle G, i\rangle$ and a direction $h' \in J$. If $G \neq F$ then neither u nor v is traversible in the direction h. Let $G = F$. Since $\langle u_i, v_i\rangle \in \Theta$, u_i is traversible in the direction h' if and only if v_i is traversible in the direction h', and in the positive case $u_i[h'] = v_i[h']$ and the rest of (2) is satisfied. Since $\tau_u(h) = \langle F, i\rangle \tau_{u_i}(h')$ and $\tau_v(h) = \langle F, i\rangle \tau_{v_i}(h')$, it is easy to check that $\langle u, v\rangle \in \Theta$.

It remains to prove that Θ is fully invariant. Let $\langle u, v\rangle \in \Theta$ and let f be an endomorphism of the algebra of terms; we need to show that $\langle f(u), f(v)\rangle \in \Theta$. Let $h \in J$. If u and v are not traversible in the direction h then also $f(u)$ and $f(v)$ are not traversible. Let u and v be traversible. Then $u[h] = v[h] = x$, where x is a variable. Now $f(s)$ is traversible in the direction h if and only if $f(x)$ is traversible in the direction k, where $h = \tau_u(h)k$. Since $\langle J, m, d\rangle$ is a coherent triple, it follows that $k = k'$, where $h = \tau_v(h)k'$. Thus $f(u)$ is traversible in the direction h if and only if $f(v)$ is, and in the positive case $f(u)[h] = f(v)[h]$. The rest is easy to check. □

The variety corresponding to the equational theory $\Theta(J, m, d)$, for a coherent triple $\langle J, m, d\rangle$, will be denoted by $\Xi(J, m, d)$. Such varieties are called *zigzag varieties*.

A coherent triple $\langle J, m, d\rangle$ is said to be *tight* if $m(h) = m(h')$ for any two directions $h, h' \in J$ with a common initial segment of length $m(h)$. Let us define an ordering on the set of tight coherent triples by $\langle J, m, d\rangle \leq \langle J', m', d'\rangle$ if and only if $J \subseteq J'$ and whenever $h \in J$ then $m(h) \leq m'(h)$ and $d'(h)$ is a multiple of $d(h)$. It is not difficult to prove that the set of tight coherent triples is a meet-complete lattice (a lattice that would be complete if the largest element were added) with respect to this ordering.

4.3. THEOREM. *Ξ is an isomorphism of the lattice of tight coherent triples onto the lattice of zigzag varieties of the given signature.* (*Zigzag varieties of the given signature form a lattice with respect to inclusion, although it is not a sublattice of the lattice of all varieties.*) *In particular, every zigzag varietiy can be uniquely expressed as $\Xi(J, m, d)$ for a tight coherent triple $\langle J, m, d\rangle$.*

PROOF. It is easy. □

4.4. THEOREM. *Every zigzag variety is bounded.*

PROOF. Let $\langle J, m, d\rangle$ be a tight coherent triple. Denote by U the set of all operation symbols that occur in elementary addresses on the directions belonging to J, so that U is finite. Denote by k the maximum of the numbers $m(h)$ ($h \in J$) and by p the least common multiple of the numbers $d(h)$ ($h \in J$). Let J' be the set of the directions h' containing no other operation symbols than those in U, and such that $h' = ef^{\omega}$ where $\lambda(e) = k$ and $\lambda(f) = p$. Put $m'(h') = k$ and $d'(h') = p$ for all $h \in J'$. Then $\langle J', m', d'\rangle$ is a tight coherent triple and $\langle J, m, d\rangle \leq \langle J', m', d'\rangle$, so that $\Xi(J, m, d) \subseteq \Xi(J', m', d')$. It is not difficult to see that every term is equivalent modulo $\Theta(J', m', d')$ with a term t such that every address that is an occurrence of a subterm in t is of length less than $k + 2p$ and t contains no operation symbols other than those in U (supplemented by an arbitrary fixed symbol not in U, if there are such symbols). There are only finitely many such terms up to similarity. Consequently, $\Xi(J', m', d')$ is bounded; and then also $\Xi(J, m, d)$ is bounded. □

CHAPTER 9

COMMUTATOR THEORY AND ABELIAN ALGEBRAS

1. Commutator in general algebras

Let α, β, δ be three congruences of an algebra A. We say that α *centralizes* β modulo δ, and write $\mathbf{C}(\alpha, \beta; \delta)$, if

$$t(a, c_1, \dots, c_n)\ \delta\ t(a, d_1, \dots, d_n) \longleftrightarrow t(b, c_1, \dots, c_n)\ \delta\ t(b, d_1, \dots, d_n)$$

for any $n \geq 0$, any $(n+1)$-ary term operation t of A and any $\langle a, b\rangle \in \alpha$ and $\langle c_1, d_1\rangle, \dots, \langle c_n, d_n\rangle \in \beta$.

Clearly, this condition is equivalent to

$$\begin{aligned}&p(a_1, \dots, a_m, c_1, \dots, c_n)\ \delta\ p(a_1, \dots, a_m, d_1, \dots, d_n) \longrightarrow \\ &p(b_1, \dots, b_m, c_1, \dots, c_n)\ \delta\ p(b_1, \dots, b_m, d_1, \dots, d_n)\end{aligned}$$

for any $n, m \geq 0$, any $(n+m)$-ary polynomial p of A and any $\langle a_i, b_i\rangle \in \alpha$, $\langle c_i, d_i\rangle \in \beta$.

1.1. Theorem. *The following are true for congruences $\alpha, \beta, \gamma, \delta, \alpha_i, \beta_i, \delta_i$ of any algebra A:*

(1) *If $\mathbf{C}(\alpha, \beta; \delta)$ then $\mathbf{C}(\alpha_0, \beta_0; \delta)$ for any $\alpha_0 \subseteq \alpha$ and $\beta_0 \subseteq \beta$.*
(2) *If $\mathbf{C}(\alpha_i, \beta; \delta)$ for all $i \in I$, then $\mathbf{C}(\bigvee_{i\in I} \alpha_i, \beta; \delta)$.*
(3) *If $\mathbf{C}(\alpha, \beta; \delta_i)$ for all $i \in I$, then $\mathbf{C}(\alpha, \beta; \bigcap_{i\in I} \delta_i)$.*
(4) *If $\gamma \subseteq \alpha \cap \beta \cap \delta$, then $\mathbf{C}(\alpha, \beta; \delta)$ in A if and only if $\mathbf{C}(\alpha/\gamma, \beta/\gamma; \delta/\gamma)$ in A/γ.*

Proof. It is easy. □

It follows from 1.1(3) that for any two congruences α, β of an algebra A there exists a least congruence δ of A with the property $\mathbf{C}(\alpha, \beta; \delta)$. This least congruence δ is called the *commutator* of α, β; it is denoted by $[\alpha, \beta]$.

1.2. Theorem. *The following are true for congruences α, β of any algebra A:*

(1) *$[\alpha, \beta] \subseteq \alpha \cap \beta$.*
(2) *If $\alpha_1 \subseteq \alpha_2$ and $\beta_1 \subseteq \beta_2$, then $[\alpha_1, \beta_1] \subseteq [\alpha_2, \beta_2]$.*

Proof. It is easy. □

1.3. Example. Let G be a group and α, β be two congruences of G; let H, K be the corresponding normal subgroups. Then $[\alpha, \beta]$ is the congruence of G corresponding to the commutator $[H,K]$ (the subgroup generated by the elements $h^{-1}k^{-1}hk$ with $h \in H$ and $k \in K$).

Let R be a ring and α, β be two congruences of R; let I, J be the corresponding ideals. Then $[\alpha, \beta]$ is the congruence of R corresponding to the ideal generated by $IJ + JI$.

Let A be an algebra. The *center* of A is the binary relation R on A defined as follows. An ordered pair $\langle a, b\rangle$ belongs to the center of A if and only if for every $n \geq 0$, every $(n+1)$-ary term operation t of A and any elements $c_1, \dots, c_n, d_1, \dots, d_n \in A$,

$$t(a, c_1, \dots, c_n) = t(a, d_1, \dots, d_n) \longleftrightarrow t(b, c_1, \dots, c_n) = t(b, d_1, \dots, d_n).$$

It is not difficult to prove that the center of any algebra A is a congruence of A.

An algebra A is said to be *Abelian* if its center is the all-relation A^2. Equivalently, an algebra A is Abelian if and only if $[A^2, A^2] = \mathbf{id}_A$.

A congruence α of an algebra A is said to be Abelian if $[\alpha, \alpha] = \mathbf{id}_A$, i.e., if $\mathbf{C}(\alpha, \alpha; \mathbf{id}_A)$. So, an algebra A is Abelian if and only if A^2 is an Abelian congruence of A.

Let α, β be two congruences of an algebra A. We say that β is *Abelian over* α if $\alpha \subseteq \beta$ and $\mathbf{C}(\beta, \beta; \alpha)$ (i.e., β/α is an Abelian congruence of A/α).

1.4. THEOREM. *An algebra A is Abelian if and only if $\mathbf{id}_A$ is a block of a congruence of $A \times A$.*

PROOF. Clearly, $\mathbf{id}_A$ is a block of a congruence of $A \times A$ if and only if it is a block of the congruence of $A \times A$ generated by $\mathbf{id}_A$. By 6.4.5, this is equivalent to saying that for every unary polynomial

$$f(x) = t^{A \times A}(x, \langle c_1, d_1\rangle, \dots, \langle c_n, d_n\rangle)$$

of $A \times A$ (where t is a term in variables $x, x_1, \dots, x_n$ for some $n \geq 0$), $f(\langle a, a\rangle) \in \mathbf{id}_A$ for some $a \in A$ implies $f(\langle b, b\rangle) \in \mathbf{id}_A$ for all $b \in A$. If we reformulate this using t^A instead of $t^{A \times A}$, we obtain the implication in the definition of an Abelian algebra. $\square$

By an *Abelian variety* we mean a variety, all the algebras of which are Abelian.

Let α, β be two congruences of an algebra A. We say that β is *strongly Abelian* over α if α if $\alpha \subseteq \beta$ and

$$p(a, c_1, \dots, c_n) \stackrel{\alpha}{\equiv} p(b, d_1, \dots, d_n) \rightarrow p(a, e_1, \dots, e_n) \stackrel{\alpha}{\equiv} p(b, e_1, \dots, e_n)$$

whenever p is an $(n+1)$-ary polynomial p of A, $a \stackrel{\beta}{\equiv} b$ and $c_i \stackrel{\beta}{\equiv} d_i \stackrel{\beta}{\equiv} e_i$ for $i = 1, \dots, n$.

We say that β is a *strongly Abelian congruence* of A if β is strongly Abelian over $\mathbf{id}_A$. An algebra A is said to be strongly Abelian if $A \times A$ is a strongly Abelian congruence.

1.5. PROPOSITION. *Let $\alpha \subseteq \beta$ be two congruences of an algebra A.*

(1) *If β is strongly Abelian over α then β is Abelian over α*
(2) *If γ is a congruence of A and $\gamma \subseteq \alpha$ then β is (strongly) Abelian over α if and only if β/γ is (strongly) Abelian over α/γ*

PROOF. It is easy. □

2. Commutator theory in congruence modular varieties

Throughout this section let V be a congruence modular variety and let $d_0, \dots, d_N$ be Day terms for V.

For an algebra $A \in V$ and two congruences $\alpha, \beta \in \mathbf{Con}(A)$ we denote by $\mathbf{M}(\alpha, \beta)$ the set of the 2×2-matrices

$$\begin{pmatrix} t(a_1, \dots, a_m, c_1, \dots, c_n) & t(a_1^1, \dots, a_m, d_1, \dots, d_n) \\ t(b_1, \dots, b_m, c_1, \dots, c_n) & t(b_1^1, \dots, b_m, d_1, \dots, d_n) \end{pmatrix}$$

where $n, m \geq 0$, t is an $(n+m)$-ary term operation of A, $\langle a_i, b_i \rangle \in \alpha$ for $i = 1, \dots, m$ and $\langle c_j, d_j \rangle \in \beta$ for $j = 1, \dots, n$. So, α centralizes β modulo δ if and only if for every $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbf{M}(\alpha, \beta)$, $\langle a, b \rangle \in \delta$ implies $\langle c, d \rangle \in \delta$.

We denote by $\mathbf{X}(\alpha, \beta)$ the set of the ordered pairs $\langle d_i(a, b, d, c), d_i(a, a, c, c) \rangle$ where $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbf{M}(\alpha, \beta)$ and $i \leq N$.

2.1. LEMMA. *Let $A \in V$, $\gamma \in \mathbf{Con}(A)$ and $a, b, c, d \in A$ be elements such that $\langle b, d \rangle \in \gamma$. Then $\langle a, c \rangle \in \gamma$ if and only if $\langle d_i(a, a, c, c), d_i(a, b, d, c) \rangle \in \gamma$ for all $i \leq N$.*

PROOF. If $\langle a, c \rangle \in \gamma$, then $d_i(a, a, c, c) \;\gamma\; d_i(a, a, a, a) = a$ and $d_i(a, b, d, c) \;\gamma\; d_i(a, b, b, a) = a$. Conversely, let $u_i = d_i(a, b, d, c)$, $v_i = d_i(a, a, c, c)$ and assume that $\langle u_i, v_i \rangle \in \gamma$ for all i. Since $u_0 = a$ and $u_N = c$, it is enough to prove $\langle u_{i-1}, u_i \rangle \in \gamma$ for $i = 1, \dots, N$. For i odd we have $u_{i-1} \;\gamma\; v_{i-1} = v_i \;\gamma\; u_i$. For i even we have $u_{i-1} \;\gamma\; d_{i-1}(a, b, b, c) = d_i(a, b, b, c) \;\gamma\; u_i$. □

2.2. LEMMA. *Let $A \in V$. The following conditions are equivalent for $\alpha, \beta, \delta \in \mathbf{Con}(A)$:*

(1) $\mathbf{X}(\alpha, \beta) \subseteq \delta$;
(2) $\mathbf{X}(\beta, \alpha) \subseteq \delta$;
(3) $\mathbf{C}(\alpha, \beta; \delta)$;
(4) $\mathbf{C}(\beta, \alpha; \delta)$;
(5) $[\alpha, \beta] \subseteq \delta$.

PROOF. It is enough to prove (3) $\Rightarrow$ (1) $\Rightarrow$ (4), since then we obtain (4) $\Rightarrow$ (2) $\Rightarrow$ (3) by interchanging α and β, and the equivalence with (5) follows easily.

(3) $\Rightarrow$ (1): Let $\mathbf{C}(\alpha,\beta;\delta)$. Let $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbf{M}(\alpha,\beta)$ be given by t as above. Let $k \le N$ and put $u = d_k(a,a,c,c)$ and $v = d_k(a,b,d,c)$. We have

$$u = d_k(t(a_i,c_j),t(a_i c_j),t(b_i,c_j),t(b_i,c_j)),$$
$$v = d_k(t(a_i,c_j),t(a_i,d_j),t(b_i,d_j),t(b_i,c_j)).$$

If we replace the second occurrences of a_i by b_i and the second occurrences of b_i with a_i, we obtain equal elements; denote the element by w. Now $\begin{pmatrix} u & v \\ w & w \end{pmatrix} \in \mathbf{M}(\alpha,\beta)$, so $\langle u,v\rangle \in \delta$.

(1) $\Rightarrow$ (4): Let $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbf{M}(\beta,\alpha)$ and $\langle a,b\rangle \in \delta$. We need to prove $\langle d,c\rangle \in \delta$; since $\langle b,a\rangle \in \delta$, according to 2.1 this is equivalent to

$$\langle d_i(a,a,b,b), d_i(a,c,d,b)\rangle \in \delta.$$

But these pairs belong to $\mathbf{X}(\alpha,\beta) \subseteq \delta$, since $\begin{pmatrix} a & c \\ b & d \end{pmatrix} \in \mathbf{M}(\alpha,\beta)$. □

2.3. Theorem. *Let α,β,β_i be congruences of an algebra A in a congruence modular variety. Then:*

(1) $[\alpha,\beta] = [\beta,\alpha] = \mathbf{Cg}(\mathbf{X}(\alpha,\beta)) \subseteq \alpha \cap \beta$.
(2) $[\alpha, \bigvee_{i\in I}\beta_i] = \bigvee_{i\in I}[\alpha,\beta_i]$.

Proof. (1) follows from 2.2. Put $\beta = \bigvee_{i\in I}\beta_i$ and $\delta = \bigvee_{i\in I}[\alpha,\beta_i]$. We have $\delta \subseteq [\alpha,\beta]$ by monotonicity (1.2(2). It remains to prove that $\mathbf{C}(\beta,\alpha;\delta)$. We have $\mathbf{C}(\beta_i,\alpha;\delta)$. Let $\begin{pmatrix} u & r \\ v & s \end{pmatrix} \in \mathbf{M}(\beta,\alpha)$, i.e., $\begin{pmatrix} u & v \\ r & s \end{pmatrix} \in \mathbf{M}(\alpha,\beta)$, and let $\langle u,r\rangle \in \delta$. Clearly, there exist finite sequences $x_0,\dots,x_k,z_0,\dots,z_k$ such that $\begin{pmatrix} x_0 & x_k \\ z_0 & z_k \end{pmatrix} = \begin{pmatrix} u & v \\ r & s \end{pmatrix}$ and $\begin{pmatrix} x_{j-1} & x_j \\ z_{j-1} & z_j \end{pmatrix}$ for $j = 1,\dots,k$ with the same term operations as for $\begin{pmatrix} u & v \\ r & s \end{pmatrix}$. Thus inductively $\langle x_j,z_j\rangle \in \delta$. Hence $\langle v,s\rangle \in \delta$. □

It follows that for an algebra A in a congruence modular variety and for any two congruences α,β of A there exists a largest congruence γ with $[\beta,\gamma] \subseteq \alpha$. This largest congruence γ is denoted by $\alpha : \beta$.

2.4. Theorem. *Let A, B be two algebras in a congruence modular variety and f be a homomorphism of A onto B, with kernel r.*

(1) *If α,β are two congruences of A with $\alpha,\beta \supseteq r$, then $[f(\alpha),f(\beta)] = f([\alpha,\beta]\vee r)$.*
(2) *For any $\alpha,\beta \in \mathbf{Con}(A)$ we have $[\alpha,\beta]\vee r = f^{-1}[f(\alpha\vee r), f(\beta\vee r)]$.*

Proof. (1) f maps the generating relation $\mathbf{X}(\alpha,\beta)\vee r$ of $[\alpha,\beta]\vee r$ onto a generating relation of $[f(\alpha),f(\beta)]$.

(2) This follows from (1) and from $[\alpha,\beta]\vee r = [\alpha\vee r,\beta\vee r]\vee r$, which is a consequence of 2.3(2). □

2.5. Theorem. *Let A be an algebra in a congruence modular variety and B be a subalgebra of A; let $\alpha,\beta\in\mathbf{Con}(A)$. Then $[\alpha|_B,\beta|_B]\subseteq[\alpha,\beta]|_B$.*

Proof. Since $\mathbf{C}(\alpha,\beta;[\alpha,\beta])$, we have $\mathbf{C}(\alpha|_B,\beta|_B;[\alpha,\beta]|_B)$. $\square$

2.6. Theorem. *Let $A=\Pi_{i\in I}A_i$ where A_i are algebras in a congruence modular variety. For $\alpha_i\in\mathbf{Con}(A_i)$ $(i\in I)$ denote by $\Pi_{i\in I}\alpha_i$ the congruence α of A defined by $\langle f,g\rangle\in\alpha$ if and only if $\langle f(i),g(i)\rangle\in\alpha_i$ for all $i\in I$. Define a congruence λ of A by $\langle f,g\rangle\in\lambda$ if and only if $f(i)=g(i)$ for all but finitely many $i\in I$. Then $(\alpha_i)_{i\in I}\mapsto\lambda\cap\Pi_{i\in I}\alpha_i$ is an embedding of the lattice $\Pi_{i\in I}\mathbf{Con}(A_i)$ into the lattice $\mathbf{Con}(A)$. For two families of congruences $\alpha_i,\beta_i\in\mathbf{Con}(A_i)$ we have*

$$[\Pi_{i\in I}\alpha_i,\Pi_{i\in I}\beta_i]\subseteq\Pi_{i\in I}[\alpha_i,\beta_i],$$
$$[\lambda\cap\Pi_{i\in I}\alpha_i,\lambda\cap\Pi_{i\in I}\beta_i]=\lambda\cap\Pi_{i\in I}[\alpha_i,\beta_i].$$

Proof. It is easy to see that the mapping is a lattice embedding. Put $\alpha=[\lambda\cap\Pi\alpha_i,\lambda\cap\Pi\beta_i]$, $\beta=\lambda\cap\Pi[\alpha_i,\beta_i]$. Let $p_i:A\to A_i$ be the projections. For $i\in I$ put $\gamma_i=\bigcap_{j\neq i}Ker(p_j)$. For $\gamma\in\mathbf{Con}(A_i)$ put $\gamma^*=p_i^{-1}(\gamma)\in\mathbf{Con}(A)$. Clearly, $\lambda\cap\Pi\alpha_i=\lambda\cap\bigcap\alpha_i^*$. We have $\alpha=[\lambda\cap\bigcap\alpha_i^*,\lambda\cap\bigcap\beta_i^*]$ and $\beta=\lambda\cap\bigcap[\alpha_i,\beta_i]^*=\lambda\cap\bigcap([p_i^{-1}\alpha_i,p_i^{-1}\beta_i]\vee\mathbf{ker}(p_i))=\lambda\cap\bigcap([\alpha_i^*,\beta_i^*]\vee\mathbf{ker}(p_i))$, because $[\alpha_i,\beta_i]^*=p_i^{-1}[\alpha_i,\beta_i]=[p_i^{-1}\alpha_i,p_i^{-1}\beta_i]\vee\mathbf{ker}(p_i)$ by Theorem 2.4. By monotonicity, $\alpha\subseteq\beta$. Similarly, $[\Pi\alpha_i,\Pi\beta_i]\subseteq\Pi[\alpha_i,\beta_i]$.

It remains to prove $\beta\subseteq\alpha$. Easily, $\lambda\cap\bigcap_i\alpha_i^*=\bigvee_i(\alpha_i^*\cap\gamma_i)$. Hence $\beta=\lambda\cap\bigcap[\alpha_i,\beta_i]^*=\bigvee_i([\alpha_i,\beta_i]^*\cap\gamma_i)=\bigvee(([\alpha_i^*,\beta_i^*]\vee\mathbf{ker}(p_i))\cap\gamma_i)$. By modularity, $\alpha_i^*=(\alpha_i^*\cap\gamma_i)\vee\mathbf{ker}(p_i)$. Hence $[\alpha_i^*,\beta_i^*]\vee\mathbf{ker}(p_i)=[\alpha_i^*\cap\gamma_i)\vee\mathbf{ker}(p_i),(\beta_i^*\cap\gamma_i)\vee\mathbf{ker}(p_i)]=[\alpha_i^*\cap\gamma_i,\beta_i^*\cap\gamma_i]\vee\mathbf{ker}(p_i)$. Hence $([\alpha_i^*,\beta_i^*]\vee\mathbf{ker}(p_i))\cap\gamma_i=([\alpha_i^*\cap\gamma_i,\beta_i^*\cap\gamma_i]\vee\mathbf{ker}(p_i))\cap\gamma_i=[\alpha_i^*\cap\gamma_i,\beta_i^*\cap\gamma_i]$ (the last by modularity). Since $\gamma_i\subseteq\lambda$, we get $\beta\subseteq\alpha$. $\square$

2.7. Theorem. *The class of Abelian algebras in a congruence modular variety V is a subvariety of V.*

Proof. It follows from the above theorems. $\square$

For a more detailed exposition of commutator theory in congruence modular varieties see Freese, McKenzie [87].

3. Abelian and Hamiltonian varieties

An algebra A is said to be *Hamiltonian* if every subalgebra of A is a block of a congruence of A. A variety V is said to be Hamiltonian if every algebra from V is Hamiltonian.

3.1. Theorem. (Klukovits [75])

(1) *An algebra A is Hamiltonian if and only if for any term $t(x,y_1,\dots,y_n)$ of its signature and any elements $a,b,c_1,\dots,c_n\in A$ there is a ternary term $s(x,y,z)$ such that $s(a,b,t(a,c_1,\dots,c_n))=t(b,c_1,\dots,c_n)$.*

(2) *A variety V is Hamiltonian if and only if for any term $t(x, y_1, \dots, y_n)$ of its signature there exists a ternary term $s(x, y, z)$ such that $s(x, y, t(x, z_1, \dots, z_n)) \approx t(y, z_1, \dots, z_n)$ is satisfied in V.*

PROOF. Let A be Hamiltonian and $t, a, b, c_1, \dots, c_n$ be given. Denote by B the subalgebra of A generated by $\{a, b, t(a, c_1, \dots, c_n)\}$. Then B is a block of a congruence r of A. Since $\langle t(a, c_1, \dots, c_n), t(b, c_1, \dots, c_n)\rangle \in r$, we have $t(b, c_1, \dots, c_n) \in B$ and hence $t(b, c_1, \dots, c_n) = s(a, b, t(a, c_1, \dots, c_n))$ for a ternary term s.

Conversely, let A be an algebra such that s exists for any t and any $a, b, c_1, \dots, c_n \in A$. Let B be a subalgebra of A. Denote by r the congruence generated by B^2 and suppose that B is not a block of r, so that $\langle a, b\rangle \in r$ for some $a \in B$ and some $b \in A - B$. There exists a Mal'cev chain from a to b with respect to B^2 and at least one link in that chain must consist of a pair of elements, one from B and the other from $A - B$. Thus there exist elements $b_1, b_2 \in B$ and a unary polynomial f of A such that $f(b_1) \in B$ and $f(b_2) \in A - B$. We have $f(x) = t(x, c_1, \dots, c_n)$ for some term $t(x, y_1, \dots, y_n)$ and some elements $c_1, \dots, c_n \in A$. Where s is the ternary term the existence of which is guaranteed by our assumption, we have $f(b_2) = t(b_2, c_1, \dots, c_n) = s(b_1, b_2, t(b_1, c_1, \dots, c_n)) \in B$, a contradiction.

(2) follows easily from (1) if we consider the free algebra in V over $n + 2$ generators. □

3.2. THEOREM. *Every Hamiltonian variety is Abelian.*

PROOF. It follows from 1.4. □

3.3. EXAMPLE. The eight-element group of quaternions is Hamiltonian. Thus not every Hamiltonian algebra is Abelian. From 1.4 it follows only that if A^2 is Hamiltonian then A is Abelian.

We are going to prove that a locally finite Abelian variety is Hamiltonian. First we need to introduce some terminology.

Two n-ary polynomials p and q of an algebra A are said to be *twins* if there is a term $t(x_1, \dots, x_n, y_1, \dots, y_m)$ for some m and elements c_i, d_i of A such that $p(a_1, \dots, a_n) = t(a_1, \dots, a_n, c_1, \dots, c_m)$ and $q(a_1, \dots, a_n) = t(a_1, \dots, a_n, d_1, \dots, d_m)$ for all $a_1, \dots, a_n \in A$. If, moreover, β is a congruence of A and $\langle c_i, d_i\rangle \in \beta$ for all i then p, q are said to be r-twins.

Clearly, if p, q are twin polynomials of an Abelian algebra A, then $\mathbf{ker}(p) = \mathbf{ker}(q)$.

For an algebra A and a subset S of A, the subsets $p(S^n) = \{p(s_1, \dots, s_n) : s_i \in S \text{ for all } i\}$ for n-ary polynomials p of A ($n \geq 1$ arbitrary) are called *neighborhoods* of S. If p, q are two n-ary twin polynomials of A then $p(S^n)$, $q(S^n)$ are called *twin neighborhoods* of S.

3.4. LEMMA. *If S is a finite subset of an Abelian algebra A and p, q are n-ary twin polynomials of A then $|p(S^n)| = |q(S^n)|$.*

PROOF. We have $|p(S^n)| = |S^n/(\mathbf{ker}(p) \cap S^n)| = |S^n/(\mathbf{ker}(q) \cap S^n)| = |q(S^n)|$. $\square$

3.5. LEMMA. *Let A be an Abelian algebra, S be a finite subset of A, p be an n-ary polynomial of A such that $T = p(S^n)$ is a maximal neighborhood of S and T is finite, and let q be a β-twin of p where $\beta = \mathbf{Cg}_A(S^2)$. Then $q(S^n) = T$.*

PROOF. We have $p(x_1, \dots, x_n) = t(x_1, \dots, x_n, c_1, \dots, c_m)$ and $q(x_1, \dots, x_n) = t(x_1, \dots, x_n, d_1, \dots, d_m)$ for an $(n+m)$-ary term t and some pairs $\langle c_i, d_i \rangle \in \beta$. Put $p_i(x_1, \dots, x_n) = t(x_1, \dots, x_n, d_1, \dots, d_i, c_{i+1}, \dots, c_m)$ for $i = 0, \dots, m$, so that $p_0 = p$ and $p_m = q$. It is sufficient to prove that $p_{i-1}(S^n) = T$ implies $p_i(S^n) = T$. We have $p_{i-1}(x_1, \dots, x_n) = f(x_1, \dots, x_n, c)$ and $q(x_1, \dots, x_n) = f(x_1, \dots, x_n, d)$ for an $(n+1)$-ary polynomial f and a pair $\langle c, d \rangle \in \beta$. There exists a Mal'cev chain $a_0, \dots, a_k$ from c to d, where $\langle a_{i-1}, a_i \rangle = \langle g_i(r_i), g_i(s_i) \rangle$ for some unary polynomial g_i and elements $r_i, s_i \in S$. Put $h_i(x_1, \dots, x_n, y) = f(x_1, \dots, x_n, g_i(y))$. Thus $h_1(x_1, \dots, x_n, r_1) = f(x_1, \dots, x_n, c) = p(x_1, \dots, x_n)$ and hence $h_1(S^n, r_1) = p(S^n) = T$. (By $h_i(S^n, r)$ we mean the set of the elements $h_i(s_1, \dots, s_n, r)$ with $s_1, \dots, s_n$ running over all n-tuples of elements of S.) We will be done if we prove $h_i(S^n, s) = T$ for all i and all $s \in S$.

Let us prove that if $h_i(S^n, r) = T$ for some $r \in S$ then $h_i(S^n, s) = T$ for all $s \in S$. We have $T \subseteq h_i(S^{n+1})$, so that by the maximality of T, $h_i(S^{n+1}) = T$. Let $s \in S$. Then $h_i(S^n, s) \subseteq T$; but $h_i(x_1, \dots, x_n, r)$ and $h_i(x_1, \dots, x_n, s)$ are twins, so that $|h_i(S^n, s)| = |h_i(S^n, r)| = |T|$ according to 3.4; we get $h_i(S^n, s) = T$.

In particular, $h_1(S^n, s) = T$ for all $s \in S$. Let us continue by induction on i. Let $h_{i-1}(S^n, s) = T$ for all s. Then $h_{i-1}(S^n, s_{i-1}) = T$; but $h_i(S^n, r_i) = h_{i-1}(S^n, s_i) = T$ and by the above claim we get $h_i(S^n, s) = T$ for all $s \in S$. $\square$

3.6. LEMMA. *Let A be an algebra generating an Abelian variety, S be a finite subset of A, $\beta = \mathbf{Cg}_A(S^2)$ and T be a maximal neighborhood of S; let T be finite. If T' is a twin neighborhood of S lying in the same block of β as T, then $T' = T$.*

PROOF. We have $T = p(S^n)$ and $T' = q(S^n)$ for some n-ary twin polynomials p, q of A; $p(x_1, \dots, x_n) = t(x_1, \dots, x_n, c_1, \dots, c_m)$ and $q(x_1, \dots, x_n) = t(x_1, \dots, x_n, d_1, \dots, d_m)$ for a term t and elements c_i, d_i. Suppose $T \neq T'$. Then $p(s_1, \dots, s_n) \neq q(s_1, \dots, s_n)$ for some elements $s_i \in S$. Put $0_p = p(s_1, \dots, s_n)$ and $0_q = q(s_1, \dots, s_n)$. We have $\langle 0_p, 0_q \rangle \in \beta$. Put $k = |T|$ and let $u_1, \dots, u_k$ be all elements of T. For each $i = 1, \dots, k$ there is some $\alpha_i = (\alpha_{i,1}, \dots, \alpha_{i,n})$ with $p(\alpha_i) = u_i$. Put $v_j = (\alpha_{1,j}, \dots, \alpha_{k,j})$ for $j = 1, \dots, n$. Thus v_j is the j-th column in the matrix with rows $\alpha_1, \dots, \alpha_k$. For every $a \in A$ denote by $\hat{a}$ the sequence $(a, a, \dots, a)$ of length k, so that $\hat{a} \in A^k$. Denote by C the subalgebra of A^k generated by $\{v_1, \dots, v_n\} \cup \{\hat{a} : a \in A\}$. For $e = (e_1, \dots, e_m) \in A^m$ put $V_e = t(v_1, \dots, v_n, \hat{e}_1, \dots, \hat{e}_m)$ (computed in the algebra C), so that $V_{c_1, \dots, c_m} = (u_1, \dots, u_k)$ and $V_{d_1, \dots, d_m}$ is some k-tuple of elements of T'. Put $\delta = \mathbf{Cg}(\hat{0}_p, \hat{0}_q)$.

Let us prove that if $w = (w_1, \ldots, w_k) \in C$ satisfies $\langle w, V_{c_1,\ldots,c_m} \rangle \in \delta$ then $\{w_1, \ldots, w_k\} = T$, i.e., $w_1, \ldots, w_k$ is a permutation of $u_1, \ldots, u_k$. There is a Mal'cev chain from w to $V_{c_1,\ldots,c_m}$. Thus to prove the claim, it is sufficient to prove that for any unary polynomial f of C, if $f(\hat{0}_p)$ is a permutation of $u_1, \ldots, u_k$ then $f(\hat{0}_q)$ is. There are a term $g(x, y_1, \ldots, y_{m'})$ and elements $\gamma_1, \ldots, \gamma_{m'}$ of C with $f(x) = g(x, \gamma_1, \ldots, \gamma_{m'})$. Each γ_i belongs to C, so we may assume that $f(x) = g(x, v_1, \ldots, v_n, \hat{a}_1, \ldots, \hat{a}_M)$ for some M and $a_1, \ldots, a_M \in A$. Put $h(x, y_1, \ldots, y_n) = g(x, y_1, \ldots, y_n, a_1, \ldots, a_M)$, so that h is a polynomial of A. We have $f(\hat{0}_p) = h(\hat{0}_p, v_1, \ldots, v_n)$ and $f(\hat{0}_q) = h(\hat{0}_q, v_1, \ldots, v_n)$. The polynomials $h(0_p, y_1, \ldots, y_n)$ and $h(0_q, y_1, \ldots, y_n)$ are β-twins, since $\langle 0_p, 0_q \rangle \in \beta$. We have $T = \{h(0_p, \alpha_{1,1}, \ldots, \alpha_{1,n}), \ldots, h(0_p, \alpha_{k,1}, \ldots, \alpha_{k,n})\} \subseteq h(0_p, S^n)$. By the maximality of T we get $h(0_p, S^n) = T$ and thus, by 3.5, $h(0_q, S^n) = T$. Hence $\{h(0_q, \alpha_{1,1}, \ldots, \alpha_{1,n}), \ldots, h(0_q, \alpha_{k,1}, \ldots, \alpha_{k,n})\} \subseteq T$. These elements must be pairwise distinct since if $h(0_q, \alpha_{i,1}, \ldots \alpha_{i,n}) = h(0_q, \alpha_{j,1}, \ldots, \alpha_{j,n})$ then the Abelian property of A implies $h(0_p, \alpha_{i,1}, \ldots \alpha_{i,n}) = h(0_p, \alpha_{j,1}, \ldots, \alpha_{j,n})$ and hence $i = j$. Thus $f(\hat{0}_q)$ is a permutation of $u_1, \ldots, u_k$.

We have $\langle t(\hat{s}_1, \ldots, \hat{s}_n, \hat{c}_1, \ldots, \hat{c}_m), t(\hat{s}_1, \ldots, \hat{s}_n, \hat{d}_1, \ldots, \hat{d}_m) \rangle = \langle 0_p, 0_q \rangle \in \delta$ and so, since C/δ is Abelian, $\langle t(v_1, \ldots, v_n, \hat{c}_1, \ldots, \hat{c}_m), t(v_1, \ldots, v_n, \hat{d}_1, \ldots, \hat{d}_m) \rangle \in \delta$. The first member of this pair is the k-tuple $u_1, \ldots, u_k$ and the second is a k-tuple of some elements of T'; by the above claim we get $T' = T$. □

3.7. LEMMA. *Let A be an algebra generating an Abelian variety and B be a finite subalgebra of A. Then B is a block of some congruence of A.*

PROOF. Put $N = |B|$ and $\beta = \mathbf{Cg}_A(B^2)$. Let us first prove that every neighborhood of B has at most N elements. Let $T = p(B^n)$ for an n-ary polynomial p of A, where $p(x_1, \ldots, x_n) = t(x_1, \ldots, x_n, a_1, \ldots, a_m)$ for a term t and elements $a_i \in A$. Take any m-tuple $b_1, \ldots, b_m$ of elements of B. Then $q(x_1, \ldots, x_n) = t(x_1, \ldots, x_n, b_1, \ldots, b_m)$ is twin with p and $|p(B^n)| = q(B^n)|$ according to 3.4. Since B is a subalgebra and q is a polynomial of B, we have $q(B^n) \subseteq B$ and thus $|q(B^n)| \leq N$.

Suppose that B is not a block of β. Then there exist elements $a \in A - B$ and $b \in B$ with $\langle a, b \rangle \in \beta$. There exists a Mal'cev chain from a to b with respect to β and at least one link in that chain must yield a pair $b_1, b_2 \in B$ with $f(b_1) \notin B$ and $f(b_2) \in B$ for a unary polynomial f of A. Now $f(B)$ is a neighborhood of B; since by the above claim there is a bound on the sizes of neighborhoods of B, $f(B)$ is contained in a maximal neighborhood T of B. Both $T \cap B$ and $T \cap (A - B)$ are nonempty. Form a twin polynomial g of f by taking the constants in B, and set $T' = g(B)$. Since g is a unary polynomial of B, we have $T' \subseteq B$ and since $T \cap B$ is nonempty, it follows that T, T' are contained in the same block of β. But then $T' = T$ by 3.6. We get a contradiction with $T \cap (A - B) \neq \emptyset$. □

3.8. THEOREM. (Kiss and Valeriote [93]) *Every locally finite Abelian variety is Hamiltonian.*

PROOF. It follows easily from 3.8 and 3.1. □

CHAPTER 10

FINITELY BASED VARIETIES

1. A sufficient condition for a finite base

1.1. THEOREM. (Birkhoff [35]) *Let the signature σ be finite. Let E be an equational theory which has a base consisting of equations in n variables $x_1, \dots, x_n$. If the free algebra over $x_1, \dots, x_n$ in the variety determined by E is finite, then E is finitely based.*

PROOF. Let T_n be the subalgebra of the term algebra generated by $x_1, \dots, x_n$, let B be a free algebra over $x_1, \dots, x_n$ in the variety V corresponding to E, and let h be the homomorphism of T_n onto B extending the identity on $x_1, \dots, x_n$. For every element $b \in B$ let us take one element $b^* \in T_n$ with $h(b^*) = b$, in such a way that if $b \in \{x_1, \dots, x_n\}$, then $b^* = b$. Denote by Q the set of the equations $\langle F(b_1^*, \dots, b_k^*), b^* \rangle$ where F is any operation symbol of σ, k is the arity of F, $b_1, \dots, b_k$ is any k-tuple of elements of B, and $b = F_B(b_1, \dots, b_k)$. Since both σ and B are finite, the set Q is finite. It is easy to see that $Q \subseteq E$.

Let A be any model of Q. In order to prove that A is a model of E, it is sufficient to take any equation $\langle u, v \rangle \in E$ such that $u \in T_n$ and $v \in T_n$, and to prove that $f(u) = f(v)$ for any homomorphism f of T_n into A. Define a mapping g of B into A by $g(b) = f(b^*)$. Then g is a homomorphism of B into A, since if $F_B(b_1, \dots, b_k) = b$, then $F_A(g(b_1), \dots, g(b_k)) = F_A(f(b_1^*), \dots, f(b_k^*)) = f(F(b_1^*, \dots, b_k^*)) = f(b^*) = g(b)$. For any i we have $gh(x_i) = g(x_i) = f(x_i^*) = f(x_i)$, so $gh = f$. Consequently, $f(u) = gh(u) = gh(v) = f(v)$.

Since any model of Q is a model of E, the set Q is a finite base for E. $\square$

2. Definable principal congruences

We say that a variety V has *definable principal congruences* if there exists a formula $\varphi(x, y, z, u)$ (with no other free variables than x, y, z, u) such that for any algebra $A \in V$ and any quadruple a, b, c, d of elements of A, $\langle a, b \rangle \in \mathbf{Cg}(c, d)$ if and only if $\varphi(a, b, c, d)$ in A. (By this we mean that ϕ is satisfied in A under the interpretation sending x to a, y to b, z to c and u to d.)

By a *principal congruence formula* we mean a formula $\psi(x, y, z, u)$ obtained in the following way. Take a finite sequence $t_i(v_1, \dots, v_k)$ ($n \geq 0$, $0 \leq i \leq n$) of terms in some variables $v_1, \dots, v_k$ different from x, y, z, u; denote by χ the conjunction of the equations

$x = t_0(w_0, v_2, \dots, v_k)$,
$t_{i-1}(w'_{i-1}, v_2, \dots, v_k) = t_i(w_i, v_2, \dots, v_k)$ for $i = 1, \dots, n$,

$t_n(w'_n, v_2, \dots, v_k) = y$

where $\{w_i, w'_i\} = \{z, u\}$ for all i; and let ψ be the formula $(\exists v_2) \dots (\exists v_k)\psi$.

2.1. Lemma. *Let A be an algebra and a, b, c, d be elements of A. Then $\langle a, b\rangle \in \mathbf{Cg}(c, d)$ if and only if $\psi(a, b, c, d)$ for at least one principal congruence formula $\psi(x, y, z, u)$.*

Proof. It follows from 6.4.5. □

2.2. Lemma. *A variety has definable principal congruences if and only if there is a finite set S of principal congruence formulas in four free variables x, y, z, u such that for any algebra $A \in V$ and any quadruple a, b, c, d of elements of A, $\langle a, b\rangle \in \mathbf{Cg}(c, d)$ if and only if $\psi(a, b, c, d)$ in A for at least one $\psi \in S$.*

Proof. Clearly, it is sufficient to prove the direct implication. Let a variety V of signature σ have definable principal congruences with respect to a formula $\varphi(x, y, z, u)$. Denote by σ' the signature obtained from σ by extending it with four new constants C_a, C_b, C_c, C_d. Denote by S_1 the set of all principal congruence formulas of signature σ in the variables x, y, z, u and by S_2 the set of σ'-sentences $\neg\psi(C_a, C_d, C_b, C_d)$ with $\psi(x, y, z, u) \in S_1$. The theory $E \cup \{\varphi(C_a, C_b, C_c, C_d)\} \cup S_2$ is inconsistent, so that by 5.6.1 there exists a finite subset S of S_2 such that $E \cup \{\varphi(C_a, C_b, C_c, C_d)\} \cup S$ is inconsistent. But that means that for any algebra $A \in V$ and any quadruple a, b, c, d of elements of A, $\langle a, b\rangle \in \mathbf{Cg}(c, d)$ if and only if $\psi(a, b, c, d)$ in A for at least one $\psi \in S$. □

2.3. Theorem. (McKenzie [78]) *Let V be a locally finite and residually very finite variety of finite signature; let V have definable principal congruences. Then V is finitely based.*

Proof. Let V have definable principal congruences with respect to a formula $\varphi(x, y, z, u)$ which can be chosen, according to 2.2, as the disjunction of finitely many principal congruence formulas. Let Ψ_1 be a sentence that is satisfied in an (arbitrary) algebra A of the given signature if and only if for any $c, d \in A$, $\mathbf{Cg}_A(c, d)$ is just the set of all $\langle a, b\rangle \in A^2$ for which $\varphi(a, b, c, d)$ is satisfied in A. For example, Ψ_1 can be the universal closure of the conjunction of the following formulas:

$\varphi(z, u, z, u)$,

$\varphi(x, x, z, u)$,

$\varphi(x, y, z, u) \to \varphi(y, x, z, u)$,

$(\varphi(x, y, z, u) \& \varphi(y, w, z, u)) \to \varphi(x, w, z, u)$,

$(\varphi(x_1, y_1, z, u) \& \dots \& \varphi(x_n, y_n, z, u)) \to \varphi(F(x_1, \dots, x_n), F(y_1, \dots, y_n, z, u))$

for any operation symbol F in σ of arity $n > 0$.

Thus a σ-algebra A satisfies Ψ_1 if and only if for all $a, b, c, d \in A$, $\langle a, b\rangle \in \mathbf{Cg}_A$ if and only if $\varphi(a, b, c, d)$ in A. Let Ψ_2 be a sentence expressing the fact that an algebra is a subdirectly irreducible algebra in V. (There are, up to isomorphism, only finitely many such algebras in V and all of them are finite, so the existence of such a Ψ_2 should be clear.) Let Ψ_3 be a sentence expressing the fact that an algebra satisfies Ψ_1 and that if it is subdirectly irreducible, then it is a subdirectly irreducible algebra in V. For example, we could take the sentence

$$\Psi_1 \& (\exists x)(\exists y)(x \neq y \& (\forall z)(\forall u)(z \neq u \rightarrow \varphi(x,y,z,u))) \rightarrow \Psi_2.$$

Then Ψ_3 is a consequence of E, where E is the equational theory of V. By 5.6.2 there exists a finite subset E_0 of E such that Ψ_3 is a consequence of E_0. Thus all subdirectly irreducible algebras in the variety based on E_0 satisfy Ψ_2, so that they belong all to V. Since a variety is uniquely determined by its subdirectly irreducible members, it follows that E_0 is a finite base for V. □

3. Jónsson's finite basis theorem

Recall that a class of algebras is elementary if it is axiomatizable by a single sentence (which is the same like to be axiomatizable by finitely many sentences). If K is elementary then both K and its complement are closed under ultraproducts.

3.1. THEOREM. (Jónsson [95]) *Let V be a variety. If there exist an elementary class K and an axiomatizable class L such that $V \subseteq K$, every subdirectly irreducible algebra from K belongs to L and $V \cap L$ is elementary, then V is finitely based.*

PROOF. Suppose that V is not finitely based. Since V is contained in an elementary class, the signature is finite. Denote by I the set of positive integers. For every $i \in I$ take an algebra $A_i \in K - V$ satisfying all the equations $\langle u, v \rangle$ such that both u and v are terms of length at most n. Let U be an ultrafilter over I containing all complements of finite subsets of I. The ultraproduct A of the family A_i $(i \in I)$ over U belongs to V. Each algebra A_i has a homomorphic image B_i such that B_i is subdirectly irreducible and $B_i \notin V$. The ultraproduct B of the family B_i $(i \in I)$ over U is a homomorphic image of A and thus $B \in V$. We have $B \in K$ and since K is elementary, $\{i \in I : B_i \in K\} \in U$. For $i \in U$ we have $B_i \in K$ and so, since B_i is subdirectly irreducible, $B_i \in L$. Thus $B \in V \cap L$. But $V \cap L$ is elementary and B is an ultraproduct of algebras not belonging to $V \cap L$; we get a contradiction. □

An algebra A is said to be *finitely subdirectly irreducible* if for any $a, b, c, d \in A$ with $a \neq b$ and $c \neq d$, $\mathbf{Cg}(a,b) \cap \mathbf{Cg}(c,d) \neq \mathbf{id}_A$.

3.2. LEMMA. *Let V be a residually very finite variety. Then every finitely subdirectly irreducible algebra in V is subdirectly irreducible.*

PROOF. There exists a positive integer n such that every subdirectly irreducible algebra in V has cardinality less than n. Suppose that there exists a finitely subdirectly irreducible algebra $A \in V$ which is not subdirectly irreducible. Clearly, A is infinite. Take n pairwise different elements $a_1, \dots, a_n \in A$. Since A is finitely subdirectly irreducible, the intersection r of the principal congruences $\mathbf{Cg}(a_i, a_j)$ $(1 \leq i < j \leq n)$ is a nontrivial congruence. There exist elements a, b with $\langle a, b \rangle \in r$ and $a \neq b$. There exists a maximal congruence s of A with the property $\langle a, b \rangle \notin s$. The factor A/s is a subdirectly irreducible algebra belonging to V. It has at least n pairwise different elements a_i/s $(i = 1, \dots, n)$, a contradiction. □

3.3. THEOREM. *Let V be a residually very finite variety of a finite signature. Let $V \subseteq H$ where H is an elementary class for which there exists a formula $M(x, y, z, u)$ with four free variables such that whenever $A \in H$ and $a, b, c, d \in A$ then $M(a, b, c, d)$ in A if and only if $\mathbf{Cg}(a, b) \cap \mathbf{Cg}(c, d) \neq \mathbf{id}_A$. Then V is finitely based.*

PROOF. By 3.2, every finitely subdirectly irreducible algebra A in V is subdirectly irreducible. Denote by L the class of subdirectly irreducible algebras in V and denote by K the class of the algebras $A \in H$ such that A is not finitely subdirectly irreducible, unless A is a subdirectly irreducible algebra from V. Thus $V \subseteq K$. It follows from the assumptions that both K and L are elementary. Consequently, V is finitely based by 3.1. □

4. Meet-semidistributive varieties

The aim of this section is to prove the finite basis theorem for congruence meet-semidistributive varieties.

For every set A denote by $A^{(2)}$ the set of all precisely two-element subsets of A.

We will later make use of *Ramsey's theorem*, which will now be explained. Define positive integers $R(i, j)$ by induction for any integers $i, j \geq 2$ as follows: $f(i, 2) = f(2, i) = i$ for all $i \geq 2$; $f(i, j) = f(i-1, j) + f(i, j-1)$ for $i, j \geq 3$. It is not difficult to prove that for any set A of cardinality $R(i, j)$ and any subset S of $A^{(2)}$ one of the following two cases takes place: either there exists a subset B of A with $|B| = i$ and $B^{(2)} \cap S = \emptyset$, or else there exists a subset C of A such that $|C| = j$ and $C^{(2)} \subseteq S$. In particular, for any set A of cardinality $R(i, i)$ and any subset S of $A^{(2)}$ there exists a subset B of A with $|B| = i$ such that either $B^{(2)} \subseteq S$ or $B^{(2)} \cap S = \emptyset$.

In the following let V be a congruence meet-semidistributive variety of finite signature with Willard terms s_e, t_e $(e \in E)$. For an algebra $A \in V$ denote by U the set of those unary polynomials of A that can be expressed either as $F(c_1, \dots, c_{i-1}, x, c_{i+1}, \dots, c_n)$ for an n-ary operation symbol F in the signature, some $i \in \{1, \dots, n\}$ and elements $c_j \in A$, or as one of $s_e(x, c, d)$, $s_e(c, x, d)$, $s_e(c, d, x)$, $t_e(x, c, d)$, $t_e(c, x, d)$, $t_e(c, d, x)$ for some $e \in E$ and $c, d \in A$. For $k \geq 0$ denote by U_k the set of the unary polynomials of A that can be expressed as a composition of at most k polynomials from U. (Thus $U_0 = \{\mathbf{id}_A\}$.) For two elements $\{a, b\}$ and $\{c, d\}$ of $A^{(2)}$ write

(1) $\{a, b\} \to_k \{c, d\}$ if $\{f(a), f(b)\} = \{c, d\}$ for some $f \in U_k$,
(2) $\{a, b\} \Rightarrow_{k,n} \{c, d\}$ if there exists a sequence $c_0, \dots, c_n$ from c to d such that for every $i < n$ either $c_i = c_{i+1}$ or $\{a, b\} \to_k \{c_i, c_{i+1}\}$,
(3) $\{a, b\} \Rightarrow_k \{c, d\}$ if $\{a, b\} \Rightarrow_{k,n} \{c, d\}$ for some n.

Thus $\langle c, d\rangle \in \mathbf{Cg}(a, b)$ if and only if $\{a, b\} \Rightarrow_k \{c, d\}$ for some k.

Observe that $\{a,b\} \Rightarrow_{k,n} \{c,d\} \Rightarrow_{l,m} \{r,s\}$ implies $\{a,b\} \Rightarrow_{k+l,nm} \{r,s\}$. Also, if $\{a,b\} \rightarrow_{k+l} \{c,d\}$ then $\{a,b\} \rightarrow_k \{r,s\} \rightarrow_l \{c,d\}$ for some $\{r,s\}$.

For a mapping f of A into A, $f\{a,b\}$ will stand for $\{f(a),f(b)\}$. By a sequence from a to b we will mean a finite sequence $S=(a_0,\dots,a_n)$ of elements of A such that $a_0=a$ and $a_n=b$; let fS stand for $(f(a_0),\dots,f(a_n))$. By a link in S we will mean any pair $\{a_i,a_{i+1}\}$ with $i<n$ and $a_i\neq a_{i+1}$. Put $S^{\leftarrow}=(a_n,\dots,a_1,a_0)$. If S is a sequence from a to b and T is a sequence from b to c, denote by ST the sequence from a to c obtained by concatenation.

4.1. Lemma. *Let $A\in V$, $\{a,b\}\in A^{(2)}$ and let S be a sequence of elements of A from a to b. Then there exist a $\{c,d\}\in A^{(2)}$ and a link $\{x,y\}$ in S such that $\{x,y\}\Rightarrow_{1,2}\{c,d\}$ and $\{a,b\}\Rightarrow_{1,2}\{c,d\}$.*

Proof. Four unary polynomials from U_2 witnessing these facts were constructed in the proof of implication (3) ⇒ (4) of Theorem 7.8.1. □

4.2. Lemma. *Let $A\in V$, $\{a,b\}\in A^{(2)}$ and let $S_1,\dots,S_n$ be sequences of elements of A from a to b. Then there exist a $\{c,d\}\in A^{(2)}$ and, for each $i=1,\dots,n$, a link $\{x_i,y_i\}$ in S_i such that $\{a,b\}\Rightarrow_{n,2^n}\{c,d\}$ and $\{x_i,y_i\}\Rightarrow_{n,2^n}\{c,d\}$ for all i.*

Proof. By induction on n. For $n=1$ the claim is 4.1. Let $n>1$. By the induction hypothesis applied to $S_1,\dots,S_{n-1}$ there exist a $\{u,v\}$ and, for each $i<n$, a link $\{x_i,y_i\}$ in S_i such that $\{a,b\}\Rightarrow_{n-1,2^{n-1}}\{u,v\}$ and $\{x_i,y_i\}\Rightarrow_{n-1,2^{n-1}}\{u,v\}$ for all $i<n$. There exists a sequence $u_0,\dots,u_m$ from u to v for some $m\le 2^{n-1}$ such that $\{a,b\}\rightarrow_{n-1}\{u_j,u_{j+1}\}$ for all $j<m$. (The use of the braces also means that $u_j\neq u_{j+1}$.) We have $\{u_j,u_{j+1}\}=f_j\{a,b\}$ for some $f_j\in U_{n-1}$. For $j<m$ put $T_j=f_jS_n$ if $\langle f_j(a),f_j(b)\rangle=\langle u_j,u_{j+1}\rangle$, while $T_j=f_jS_n^{\leftarrow}$ if $\langle f_j(a),f_j(b)\rangle=\langle u_{j+1},u_j\rangle$. Denote by T the sequence $T_0\dots T_{m-1}$, so that T is a sequence from u to v. By 4.1 there exist a $\{c,d\}$ and a link $\{x,y\}$ in T such that $\{x,y\}\Rightarrow_{1,2}\{c,d\}$ and $\{u,v\}\Rightarrow_{1,2}\{c,d\}$. This $\{x,y\}$ is a link in T_j for some $j<m$, so that there exists a link in S_n, which we denote by $\{x_n,y_n\}$, such that $\{x,y\}=f_j\{x_n,y_n\}$. We have

$\{x_i,y_i\}\Rightarrow_{n-1,2^{n-1}}\{u,v\}\Rightarrow_{1,2}\{c,d\}$ and so $\{x_i,y_i\}\Rightarrow_{n,2^n}\{c,d\}$ for $i<n$,
$\{x_n,y_n\}\rightarrow_{n-1}\{x,y\}\Rightarrow_{1,2}\{c,d\}$ and so $\{x_n,y_n\}\Rightarrow_{n,2^n}\{c,d\}$,
$\{a,b\}\Rightarrow_{n-1,2^{n-1}}\{u,v\}\Rightarrow_{1,2}\{c,d\}$ and so $\{a,b\}\Rightarrow_{n,2^n}\{c,d\}$. □

4.3. Lemma. *Let $A\in V$, $\{a_1,b_1\},\dots,\{a_n,b_n\},\{u,v\}\in A^{(2)}$ and let $\{a_i,b_i\}\Rightarrow_k\{u,v\}$ for $i=1,\dots,n$. Then there exist $\{c,d\}\in A^{(2)}$ and $\{x_i,y_i\}\in A^{(2)}$ for $1\le i\le n$ such that $\{u,v\}\Rightarrow_{n,2^n}\{c,d\}$ and $\{a_i,b_i\}\rightarrow_k\{x_i,y_i\}\Rightarrow_{n,2^n}\{c,d\}$ for all i. In particular, $\{a_i,b_i\}\Rightarrow_{k+n,2^n}\{c,d\}$ for all i.*

Proof. For every $i=1,\dots,n$ there exists a sequence S_i from u to v such that $\{a_i,b_i\}\rightarrow_k\{x,y\}$ for any link $\{x,y\}$ in S_i. Apply 4.2 to $\{u,v\}$ and $S_1,\dots,S_n$. □

For an integer $m\ge 2$ put $M=R(m+1,m+1)$, $D=3+(M+1)M$, $L=(m+1)m/2$ and $C=(3+DM)L$. (It would be more appropriate to write $M=M(m)$, $D=D(m)$, $L=L(m)$, $C=C(m)$.)

4.4. Lemma. *Let $A \in V$ and let $m \geq 2$. Then one of the following two cases takes place:*

(1) *there exist $\{a,b\} \in A^{(2)}$ and a subset S of A with $|S| = m+1$ such that $\{x,y\} \Rightarrow_{DM+L,2^L} \{a,b\}$ for all $\{x,y\} \in S^{(2)}$;*

(2) *for all $\{a,b\},\{c,d\} \in A^{(2)}$ with $\mathbf{Cg}(a,b) \cap \mathbf{Cg}(c,d) \neq \mathbf{id}_A$ there exists a $\{u,v\} \in A^{(2)}$ such that $\{a,b\} \Rightarrow_{DM+2,4} \{u,v\}$ and $\{c,d\} \Rightarrow_{DM+2,4}$ $\{u,v\}$.*

Proof. Let us say that two elements $\{a,b\},\{c,d\}$ of $A^{(2)}$ are n-bounded if there exist $\{r,s\},\{r',s'\},\{u,v\} \in A^{(2)}$ such that $\{a,b\} \to_n \{r,s\} \Rightarrow_{2,4} \{u,v\}$ and $\{c,d\} \to_n \{r',s'\} \Rightarrow_{2,4} \{u,v\}$. If $\{a,b\},\{c,d\}$ are DM-bounded whenever $\mathbf{Cg}(a,b) \cap \mathbf{Cg}(cd) \neq \mathbf{id}_A$, then case (2) takes place. Let there exist $\{a,b\},\{c,d\} \in A^{(2)}$ with $\mathbf{Cg}(a,b) \cap \mathbf{Cg}(c,d) \neq \mathbf{id}_A$ such that $\{a,b\},\{c,d\}$ are not DM-bounded. It follows from 4.3 that they are n-bounded for some n. We have $n > DM$. Put $t = n - DM$, so that there exist elements a^z_j, b^z_j for $z \in \{1,2\}$ and $j \in \{0,\dots,M\}$ with $\{a,b\} \to_t \{a^1_0, b^1_0\}$, $\{c,d\} \to_t \{a^2_0, b^2_0\}$ and

$$\{a^z_0, b^z_0\} \to_D \{a^z_1, b^z_1\} \to_D \cdots \to_D \{a^z_M, b^z_M\} \Rightarrow_{2,4} \{u,v\}$$

for $z \in \{1,2\}$. Then

$$\{a^z_0, b^z_0\} \to_{Dj} \{a^z_j, b^z_j\} \to_{D(M-j)} \{a^z_M, b^z_M\}$$

for all z and j. For each z,j choose $f^z_j \in U_{D(M-j)}$ witnessing $\{a^z_j, b^z_j\} \to_{D(M-j)}$ $\{a^z_M, b^z_M\}$. We can assume that $f^z_j(a^z_j) = a^z_M$ and $f^z_j(b^z_j) = b^z_M$ (if this is not the case, switch the two elements). Since $\{a^z_M, b^z_M\} \Rightarrow_{2,4} \{u,v\}$, there are elements $u = u^z_0$, u^z_1, u^z_2, u^z_3, $u^z_4 = v$ and polynomials g^z_0, g^z_1, g^z_2, g^z_3 from U_2 such that $\{u^z_k, u^z_{k+1}\} = g^z_k\{a^z_M, b^z_M\}$ for $k < 4$.

For $0 \leq i < j \leq M$, $z \in \{1,2\}$ and $0 \leq k < 4$ denote by $R^z_{i,j}$ the sequence $f^z_j(a^z_j, a^z_i, b^z_i, b^z_j)$ from a^z_M to b^z_M and by $S^z_{i,j,k}$ the sequence from u^z_k to u^z_{k+1} obtained by applying g^z_k to either $R^z_{i,j}$ or its inverse. Put $S_{i,j} = S^z_{i,j,0} S^z_{i,j,1} S^z_{i,j,2} S^z_{i,j,3}$. Thus $S^z_{i,j}$ is a sequence from u to v and for every link $\{x,y\}$ of $S_{i,j}$ one of the following three cases takes place:

(i) $\{a^z_i, a^z_j\} \to_{D(M-j)+2} \{x,y\}$ (and $a^z_i \neq a^z_j$),
(ii) $\{b^z_i, b^z_j\} \to_{D(M-j)+2} \{x,y\}$ (and $b^z_i \neq b^z_j$),
(iii) $\{a^z_i, b^z_i\} \to_{D(M-j)+2} \{x,y\}$.

We have obtained $(M+1)M$ sequences from u to v. By 4.2 there exist $\{u',v'\} \in A^{(2)}$ and links $\{x^z_{i,j}, y^z_{i,j}\}$ in $S^z_{i,j}$ such that $\{x^z_{i,j}, y^z_{i,j}\} \Rightarrow_{(M+1)M,Q} \{u',v'\}$ where $Q = 2^{(M+1)M}$.

Suppose that for some i,j the link $\{x^z_{i,j}, y^z_{i,j}\}$ satisfies (iii). Then

$$\{a^z_0, b^z_0\} \to_{Di} \{a^z_i, b^z_i\} \to_{D(M-j)+2} \{x^z_{i,j}, y^z_{i,j}\} \Rightarrow_{(M+1)M} \{u',v'\}$$

and hence $\{a^z_0, b^z_0\} \Rightarrow_{DM-1} \{u',v'\}$ because $Di + D(M-j) + 2 + (M+1)M \leq DM - 1$. Thus if there is a link $\{x^1_{i,j}, y^1_{i,j}\}$ satisfying (iii) and also a link $\{x^2_{i',j'}, y^2_{i',j'}\}$ satisfying (iii), $\{a^1_0, b^1_0\},\{a^2_0, b^2_0\}$ would be $(DM-1)$-bounded by 4.3, a contradiction with $n < DM$. So, there exists a $z \in \{1,2\}$ such that no link $\{x^z_{i,j}, y^z_{i,j}\}$ satisfies (iii); let us take this z. Thus for all i, j, $\{x^z_{i,j}, y^z_{i,j}\}$ satisfies either (i) or (ii). By the definition of M, it follows from Ramsey's theorem that there is a subset J of $\{1, \dots, M\}$ with $|J| = m + 1$ such that either $\{x^z_{i,j}, y^z_{i,j}\}$ satisfies (i) for all $i, j \in J$ with $i < j$ or $\{x^z_{i,j}, y^z_{i,j}\}$ satisfies (ii) for all $i, j \in J$ with $i < j$. We can assume without loss of generality that the first case takes place. If $i, j \in J$ and $i < j$ then $\{a^z_i, a^z_j\} \rightarrow_{D(M-j)+2} \{x^z_{i,j}, y^z_{i,j}\} \Rightarrow_{(M+1)M,Q} \{u', v'\}$ and thus $\{a^z_i, a^z_j\} \Rightarrow_{DM} \{u', v'\}$. By 4.3 there exists a $\{c, d\}$ with $\{a_i, a_j\} \Rightarrow_{DM+L,2^L} \{c, d\}$ for all $i, j \in J$ with $i < j$. Thus (1) takes place. $\square$

Since the signature σ is finite, it is clear that for any k, n there exists a formula $\phi_{k,n}(x, y, z, u)$ which defines the relation $\{x, y\} \Rightarrow_{k,n} \{z, u\}$ on any σ-algebra. Consequently, for any $m \geq 2$ there is a sentence Φ_m which is satisfied in a σ-algebra A if and only if there exist $\{a, b\} \in A^{(2)}$ and a subset S of A with $|S| = m + 1$ such that $\{x, y\} \Rightarrow_{C,2^L} \{a, b\}$ for all $\{x, y\} \in S^{(2)}$. Observe that $DM + L \leq C$, so that if an algebra satisfies 4.4(1) then it also satisfies Φ_m.

4.5. Lemma. *Let $A \in V$ be subdirectly irreducible; let $|A| > m \geq 2$. Then A satisfies Φ_m.*

Proof. Suppose that A does not satisfy Φ_m, so that it also does not satisfy 4.4(1) and consequently it satisfies 4.4(2). Since, moreover, A is subdirectly irreducible, for any $\{a, b\}, \{c, d\} \in A^{(2)}$ there exists a $\{u, v\} \in A^{(2)}$ such that $\{a, b\} \Rightarrow_{DM+2,4} \{u, v\}$ and $\{c, d\} \Rightarrow_{DM+2,4} \{u, v\}$. It follows by induction on k that for any subset S of $A^{(2)}$ with $|S| \leq 2^k$ there exists a $\{u, v\} \in A^{(2)}$ with $\{a, b\} \Rightarrow_{DM+2,k} \{u, v\}$ for all $\{a, b\} \in S$. Let $S = B^{(2)}$ where B is a subset of A with $|B| = m + 1$, so that $|S| = L$. We get $\{a, b\} \Rightarrow_{(DM+2)L} \{u, v\}$ for all $a, b \in B$ with $a \neq b$. By 4.3 there exist two different elements u', v' in A such that $\{a, b\} \Rightarrow_{L+(DM+2)L,2^L} \{u', v'\}$ for all $\{a, b\} \in B$ with $a \neq b$. Here $L + (DM + 2)L = C$ and we get Φ_m in A. $\square$

4.6. Theorem. (Willard [00]) *Let V be a congruence meet-semidistributive, residually very finite variety of a finite signature. Then V is finitely based.*

Proof. There exists a positive integer m such that every subdirectly irreducible algebra in V has cardinality less than m. By 7.8.1 there exists a finite collection s_e, t_e of Willard terms for V. Define M, D, L, C as above. Let V^* be the elementary class defined by the formulas 7.8.1(3) for Willard terms, so that $V \subseteq V^*$. V satisfies $\neg\Phi_m$, since any model of Φ_m has a subdirectly irreducible homomorphic image with more than m elements. There is a formula $\mu(x, y, z, u)$ which is satisfied by a quadruple a, b, c, d of elements of any algebra $A \in V^*$ if and only if $a \neq b$, $c \neq b$ and there exist $u, v \in A$ with $u \neq v$, $\{a, b\} \Rightarrow_{DM+2,4} \{u, v\}$ and $\{c, d\} \Rightarrow_{DM+2,4} \{u, v\}$. Let H be the class of all algebras $A \in V^*$ satisfying $\neg\Phi_m$, so that H is elementary and $V \subseteq H$. By 4.4, the relation $\mathbf{Cg}(x, y) \cap \mathbf{Cg}(z, u) \neq \mathbf{id}$ is defined in algebras from H by μ. Now we can apply Theorem 3.2. $\square$

4.7. THEOREM. (Baker [77]) *A finitely generated, congruence distributive variety of finite signature is finitely based.*

PROOF. It follows from 4.6 and 7.7.2. □

Comments

McKenzie [70] proved that every finite lattice is finitely based. This result was generalized by Baker [77] to finite algebras in any congruence distributive variety of finite signature. A shorter and more elegant proof to Baker's theorem was obtained by Jónsson [79]. Working more with the original Baker's techniques, the result was further extended by Willard [00] to congruence meet-semidistributive varieties as presented here in Theorem 4.6.

Another extension of Baker's theorem was given by McKenzie [87]: Every finitely generated, congruence modular and residually small variety of finite signature is finitely based.

Every finite group is finitely based according to Oates and Powell [64]. Bryant [82] gives an example of a finite pointed group which is not finitely based, destroying any hope that the finite basis theorem for finite groups could have a generalization in the setting of universal algebra.

CHAPTER 11

NONFINITELY BASED VARIETIES

1. Inherently nonfinitely based varieties

Given a variety V and a nonnegative integer N, we denote by $V^{(N)}$ the variety based on the equations in no more than N variables that are satisfied in V. As it is easy to see, an algebra belongs to $V^{(N)}$ if and only if every its N-generated subalgebra belongs to V. (By an N-generated algebra we mean an algebra having a set of generators of cardinality at most N.) As it is easy to see, V is the intersection of the chain of varieties $V^{(0)} \supseteq V^{(1)} \supseteq \dots$, and if V is finitely based then $V = V^{(N)}$ for some N.

Recall that an algebra is locally finite if its every finitely generated subalgebra is finite, and a variety is locally finite if its every algebra is locally finite. Clearly, a variety is locally finite if and only if all its finitely generated free algebras are finite. The following easy observation will be useful: The variety generated by an algebra A is locally finite if and only if A is locally finite and for each $N \geq 0$ there is a finite upper bound for the sizes of the N-generated subalgebras of A.

A variety V is said to be *inherently nonfinitely based* if it is locally finite and every locally finite variety containing V is nonfinitely based. An algebra is said to be inherently nonfinitely based if it generates an inherently nonfinitely based variety.

1.1. Theorem. (McNulty [85]) *Let σ be finite. For a locally finite variety V, the following three conditions are equivalent:*

1. *V is inherently nonfinitely based;*
2. *$V^{(N)}$ is not locally finite for any N;*
3. *for infinitely many positive integers N, there is an algebra B_N such that B_N is not locally finite and every N-generated subalgebra of B_N belongs to V.*

Proof. (1) implies (2) by 1.1, and (2) implies (3) clearly. Let (3) be satisfied and suppose that there is a finitely based, locally finite variety $W \supseteq V$. Since the equational theory of W has a base consisting of equations in no more than N variables, we have $W = W^{(i)}$ for all $i \geq N$. By (3), there is an $i \geq N$ such that $V^{(i)}$ is not locally finite, so that $V^{(i)}$ is not contained in W. But $V^{(i)} \subseteq W^{(i)} = W$, a contradiction. $\square$

2. The shift-automorphism method

The aim of this section is to prove the following theorem, belonging to Baker, McNulty and Werner [89]. An element o of an algebra A is called a *zero* of A if $F_A(a_1, \dots, a_n) = o$, whenever $F \in \sigma$ and o appears among $a_1, \dots, a_n$. By a unary polynomial of A we mean a mapping of the form $a \mapsto t(a, c_1, \dots, c_r)$ for a term $t(x, x_1, \dots, x_r)$ in $r+1$ variables and an r-tuple $c_1, \dots, c_r$ of elements of A. (As it is easy to guess, by $t(a, c_1, \dots, c_r)$ we mean the image of t under the homomorphism of the term algebra into A sending x to a and x_i to c_i.)

2.1. Theorem. *Let σ be finite. Let A be an infinite, locally finite algebra with zero o and an automorphism α such that the following three conditions are satisfied:*

(1) *$\{o\}$ is the only orbit of α that is finite;*
(2) *for every $F \in \sigma$, the set $\{\langle a_1, \dots, a_n, a_{n+1}\rangle : F_A(a_1, \dots, a_n) = a_{n+1} \neq o\}$ (where n is the arity of F) is partitioned by α into finitely many orbits;*
(3) *$\alpha(a) = p(a) \neq a$ for an element a and a unary polynomial function p of A.*

Then A is inherently nonfinitely based.

The proof will be divided into several lemmas.

2.2. Lemma. *For the proof of* 2.1*, it is sufficient to assume that the number of orbits of α is finite.*

Proof. If the number is infinite, denote by A' the subset of A consisting of o and the elements that are members of $\langle a_1, \dots, a_n, a_{n+1}\rangle$ for some $F_A(a_1, \dots, a_n) = a_{n+1} \neq o$. Then A' is an underlying set of a subalgebra A' of A, the restriction of α to A' is an automorphism of A' with finitely many orbits and A' together with this restriction of α satisfy all the three conditions of 2.1. It remains to show that if A' generates a locally finite variety, then the same is true for A. But if b_N is an upper bound for the sizes of N-generated subalgebras of A', then it is easy to see that $b_N + N + 1$ is an upper bound for the sizes of N-generated subalgebras of A. □

So, from now on we assume that the number of orbits of α is finite; the number will be denoted by m. In particular, A is countable. It is possible to fix an enumeration $\dots, e_{-2}, e_{-1}, e_0, e_1, e_2, \dots$ of $A \setminus \{o\}$ in such a way that $\alpha(e_i) = e_{i+m}$ for all i.

Two elements of $A \setminus \{o\}$ will be called operationally related if they are both members of some $\langle a_1, \dots, a_n, a_{n+1}\rangle$ with $F_A(a_1, \dots, a_n) = a_{n+1} \neq o$ (for some $F \in \sigma$). Denote by M the maximal possible distance between two operationally related elements of $A \setminus \{o\}$; the distance of e_i and e_j is the number $|i - j|$.

Let N be a positive integer such that $N \geq n$ whenever n is the arity of a symbol from σ. By Theorem 1.1, we will be done with the proof of 2.1 if we construct an algebra B_N such that B_N is not locally finite and every N-generated subalgebra of B_N belongs to the variety V generated by A; we also need to show that V is locally finite.

2.3. LEMMA. *There is a positive integer w such that whenever S is a subalgebra of A generated by a subset Y of $A \smallsetminus \{o\}$, then the following are true: if r is such that $r \leq i$ for all $e_i \in Y$, then $r - w \leq i$ for all $e_i \in S$; and if s is such that $i \leq s$ for all $e_i \in Y$, then $i \leq s + w$ for all $e_i \in S$.*

PROOF. Consider first the case $X = \{e_r, \dots, e_{r+M-1}\}$ for some integer r. The subalgebra generated by X is finite; moreover, up to the automorphisms of A there are only finitely many possibilities for the subalgebra generated by such an X. Hence there is a w satisfying the claim for all such sets X. Now let X be arbitrary and $r \leq i$ for all $e_i \in X$. Clearly, the subalgebra generated by X is contained in the subalgebra generated by the set $\{e_r, e_{r+1}, \dots\}$ and, as it is easy to see, every element e_i of this subalgebra with $i < r$ is contained in the subalgebra generated by $\{e_r, e_{r+1}, \dots, e_{r+M-1}\}$, so that $i \geq r - w$. The proof is similar for the indexes on the right. $\square$

2.4. LEMMA. *For the proof of* 2.1*, it is sufficient to assume that $m > N(M + 2w)$.*

PROOF. The automorphism α can be replaced with α^k for any $k \geq 1$, and taking k sufficiently large, the number m increases beyond any bound. It is only necessary to show that if a was such that $\alpha(a) = p(a) \neq a$ for a unary polynomial p, then $\alpha^k(a) = q(a)$ for a unary polynomial q. As it is easy to see,

$$\alpha^k(a) = (\alpha^{k-1} p \alpha^{-(k-1)}) \dots (\alpha^2 p \alpha^{-2})(\alpha p \alpha^{-1}) p(a).$$

Since the composition of polynomials is a polynomial, we see that it is sufficient to prove that $\gamma p \gamma^{-1}$ is a polynomial for any automorphism γ of A. However, if $p(x) = t(x, c_1, \dots, c_r)$, then $\gamma p \gamma^{-1}(x) = t(x, \gamma(c_1), \dots, \gamma(c_r))$. $\square$

2.5. LEMMA. *Let S be a subalgebra of A generated by the union of N (or less) orbits of α. Then there is a subalgebra S_0 of A such that the sets $\alpha^k(S_0)$ (for various integers k) are pairwise disjoint, S is their union and no element of S_0 is operationally related to any element of $\alpha^k(S_0)$ with $k \neq 0$.*

PROOF. By 2.4, we continue to work under the assumption $m > N(M + 2w)$. Denote by Y the union of the N (or fewer) orbits of α, so that S is generated by Y. By the assumption on m, there are two elements e_i, e_j of Y with $j > i + M + 2w$ such that $e_k \notin Y$ for all $i < k < j$. Denote by Y_0 the set of the elements e_k with $j \leq k \leq i + m$, and by S_0 the subalgebra generated by Y_0. As it is easy to check, this subalgebra S_0 serves the purpose. $\square$

Let us define an algebra B in the following way. Its elements are 0 and the ordered pairs $\langle a, i \rangle$ where $a \in A = \{o\}$ and i is an integer. The operations are defined in such a way that

$$F_B(\langle a_1, i_1 \rangle, \dots, \langle a_n, i_n \rangle) = \langle F_A(a_1, \dots, a_n), \max(i_1, \dots, i_n) \rangle$$

if $F_A(a_1, \dots, a_n) \neq o$; in all other cases, the value of F_B is 0.

We also define an automorphism β of B by $\beta(0) = 0$ and $\beta(\langle a, i\rangle) = \langle \alpha(a), i-1\rangle$. The equivalence on B, corresponding to the partition into orbits of β, will be denoted by $\sim$.

2.6. LEMMA. $B \in V$.

PROOF. The mapping $h : B \to A^Z$ (where Z denotes the set of integers) defined by $h(\langle a, i\rangle)(j) = o$ for $j < i$, $h(\langle a, i\rangle)(j) = a$ for $j \geq i$, and $h(0)(j) = o$ for all j, is an embedding of B into a direct power of A. □

2.7. LEMMA. *The equivalence $\sim$ is almost a congruence of B, in the following sense. If $F \in \sigma$ is n-ary and $\langle a_1, i_1\rangle \sim \langle b_1, j_1\rangle, \dots, \langle a_n, i_n\rangle \sim \langle b_n, j_n\rangle$, then*

$$F_B(\langle a_1, i_1\rangle, \dots, \langle a_n, i_n\rangle) \sim F_B(\langle b_1, j_1\rangle, \dots, \langle b_n, j_n\rangle)$$

whenever both elements

$$F_B(\langle a_1, i_1\rangle, \dots, \langle a_n, i_n\rangle) \text{ and } F_B(\langle b_1, j_1\rangle, \dots, \langle b_n, j_n\rangle)$$

are different from 0.

PROOF. By the choice of N we have $N \geq n$. By the definition of $\sim$, $\langle b_1, j_1\rangle = \beta^{k_1}(\langle a_1, i_1\rangle), \dots, \langle b_n, j_n\rangle = \beta^{k_n}(\langle a_n, i_n\rangle)$ for some integers $k_1, \dots, k_n$. Clearly, we will be done if we show that $k_1 = \dots = k_n$. Denote by S the subalgebra of A generated by the α-orbits of $a_1, \dots, a_n$. Since $n \leq N$, there exists a subalgebra S_0 as in 2.5. Since each two members of $a_1, \dots, a_n$ are operationally related, there exists an r such that $\{a_1, \dots, a_n\} \subseteq \alpha^r(S_0)$. Similarly, there exists an s with $\{b_1, \dots, b_n\} \subseteq \alpha^s(S_0)$. Since each α-orbit intersects S_0 in at most one point, we get $k_1 = s - r, \dots, k_n = s - r$. □

We denote by B_N the factor of B through $\sim$, with operations defined in the following way:

$$F_{B_N}(\langle a_1, i_1\rangle/\sim, \dots, \langle a_n, i_n\rangle/\sim) = F_B(\langle a_1, i_1\rangle, \dots, \langle a_n, i_n\rangle)/\sim$$

whenever $F_B(\langle a_1, i_1\rangle, \dots, \langle a_n, i_n\rangle) \neq 0$; in all other cases, the value of F_{B_N} is $0/\sim$. By 2.6, this definition is correct.

2.8. LEMMA. *The algebra B_N is not locally finite.*

PROOF. There exist an element a, a term t and elements $c_1, \dots, c_r$ with $\alpha(a) = t(a, c_1, \dots, c_r) \neq a$. Let us prove by induction on k that the elements $\langle a, 0\rangle/\sim, \langle c_1, 0\rangle/\sim, \dots, \langle c_r, 0\rangle/\sim$ generate $\langle \alpha^k(a), 0\rangle/\sim$. Since

$$t(\langle a, k\rangle, \langle c_1, 0\rangle, \dots, \langle c_r, 0\rangle) = \langle \alpha(a), k\rangle$$

in B, we have

$$\begin{gathered} t(\langle \alpha^k(a), 0\rangle/\sim, \langle c_1, 0\rangle/\sim, \dots, \langle c_r, 0\rangle/\sim) = \\ t(\langle a, k\rangle/\sim, \langle c_1, 0\rangle/\sim, \dots, \langle c_r, 0\rangle/\sim) = \\ \langle \alpha(a), k\rangle/\sim = \langle \alpha^{k+1}(a), 0\rangle/\sim \end{gathered}$$

in B_N. □

2.9. Lemma. *Every N-generated subalgebra of B_N is isomorphic to a subalgebra of B, and hence belongs to V.*

Proof. Let N (or fewer) elements of B_N other than $0/\sim$ be given, and denote by U the union of all these blocks of $\sim$. The projection of U onto $A \setminus \{o\}$ is the union of at most N orbits of α, so it generates a subalgebra S for which S_0 as in 2.5 exists. Then $R = \{0\} \cup ((S_0 \setminus \{o\}) \times Z)$ is a subalgebra of B. It is easy to see that the restriction of $\sim$ on R is the identity and the mapping $x \mapsto x/\sim$ is an isomorphism of R onto the original subalgebra of B_N. □

2.10. Lemma. *The N-generated subalgebras of A are bounded in size.*

Proof. The generators are contained in N or fewer orbits of α. These orbits generate a subalgebra S for which S_0 as in 2.5 exists. Now S_0 has at most m elements, and the generators are contained in the union of at most N sets of the form $\alpha^k(S_0)$ for some integers k; the union is a subalgebra of cardinality at most Nm. □

The proof of Theorem 2.1 is now finished.

3. Applications

The following theorem summarizes the case where Theorem 2.1 can be applied with only one infinite orbit of α to prove that a finite algebra is inherently nonfinitely based. By a Z-sequence we mean a sequence of elements indexed by arbitrary integers. For a Z-sequence s and an integer k we define a Z-sequence s^k by $s^k_i = s_{i-k}$. Such sequences s^k, for various integers k, are called translates of s. If A is an algebra and $F \in \sigma$, then for any Z-sequences $s_1, \dots, s_n$ of elements of A, $F(s_1, \dots, s_n)$ is the Z-sequence defined componentwise, i.e., the sequence computed from $s_1, \dots, s_n$ in the direct power A^Z.

3.1. Theorem. *Let A be a finite algebra of a finite signature σ, with a zero element o. Suppose that a sequence $s = \dots e_{-1}e_0e_1 \dots$ (indexed by integers) of elements of $A \setminus \{o\}$ can be found with the following properties:*

(1) *any fundamental operation of σ applied to translates of s yields either a translate of s or a sequence containing o;*
(2) *there are only finitely many situations $F(s^{k_1}, \dots, s^{k_n}) = s^k$ where $F \in \sigma$ and $k_i = 0$ for some i;*
(3) *there is at least one situation $F(s^{k_1}, \dots, s^{k_n}) = s^1$ where $F \in \sigma$ and $k_i = 0$ for some i such that F_A actually depends on the i-th argument.*

Then A is inherently nonfinitely based.

Proof. Denote by B the subalgebra of A^Z generated by the translates of s. By (1), every element of B is either a translate of s or a Z-sequence containing o. Denote by R the equivalence on B with all blocks singletons except one, consisting of the Z-sequences containing o. Clearly, R is a congruence of B. Since B/R belongs to the variety generated by A, it is sufficient to prove that B/R is inherently nonfinitely based. The mapping $s^k \mapsto s^{k+1}$ induces an automorphism of B/R with a single orbit other than $\{o\}$, and it is easy to verify the conditions of Theorem 2.1. $\square$

Let G be an unoriented graph, possibly with loops but without multiple edges. The *graph algebra* of G is the groupoid with the underlying set $G \cup \{o\}$ (where o is an element not belonging to G) and multiplication defined by $ab = a$ if a and b are joined by an edge, and $ab = o$ otherwise.

3.2. THEOREM. *The graph algebra of any of the four graphs M, T, L_3, P_4 described in the following is inherently nonfinitely based:*

(1) *M has two vertices a, b and two edges $\{a, b\}, \{b, b\}$;*
(2) *T has three vertices a, b, c and three edges $\{a, b\}, \{b, c\}, \{c, a\}$;*
(3) *L_3 has three vertices a, b, c and five edges $\{a, b\}, \{b, c\}, \{a, a\}, \{b, b\}, \{c, c\}$;*
(4) *P_4 has four vertices a, b, c, d and three edges $\{a, b\}, \{b, c\}, \{c, d\}$.*

PROOF. Theorem 3.1 can be applied with respect to the Z-sequences

$$\dots bbbbababbabbbabbbba \dots,$$
$$\dots abababcababab \dots,$$
$$\dots aaabccc \dots,$$
$$\dots ababcdcd \dots,$$

respectively. $\square$

It has been proved in Baker, McNulty and Werner [87] that the graph algebra of a given finite graph G is finitely based if and only if it does not contain an induced subgraph isomorphic to any of the four graphs listed in 3.2.

The graph algebra of the first of these four graphs is the Murskiĭ's groupoid, the first algebra found to be inherently nonfinitely based in Murskij [65]. Here is its multiplication table:

	o	a	b
o	o	o	o
a	o	o	a
b	o	b	b

3.3. THEOREM. (Ježek [85a]) *Each of the three idempotent groupoids with the following multiplication tables is inherently nonfinitely based.*

	a	b	c
a	a	b	b
b	b	b	c
c	b	c	c

	a	b	c
a	a	c	b
b	c	b	c
c	b	c	c

	a	b	c
a	a	b	b
b	c	b	c
c	b	c	c

PROOF. Let G be any of these three groupoids. It is easy to verify that in each of the three cases, the subgroupoid of $G \times G$ generated by $\langle a, c \rangle$ and $\langle c, a \rangle$

maps homomorphically onto the four-element groupoid A with multiplication table

$$\begin{array}{c|cccc} & o & a & b & c \\ \hline o & o & o & o & o \\ a & o & a & c & o \\ b & o & c & b & o \\ c & o & o & o & c \end{array}$$

So, it is sufficient to prove that A is inherently nonfinitely based. Denote by B the subalgebra of A^Z generated by the translates of $s = \dots cccaaa\dots$ and $t = \dots cccbaaa\dots$. We have $st = \dots ccccaaa\dots$, a translate of s. Except for the translates of s and t, all the other elements of B are Z-sequences containing o. The proof, based on Theorem 2.1, can be finished similarly as in the case of Theorem 3.1; in the present case, the automorphism has two infinite orbits. □

4. The syntactic method

The following result can also be obtained as an application of the shift-automorphism method, but we prefer to present a more syntactical proof, one close to the original proof given in Perkins [84]. By a unit element of a groupoid we mean an element e such that $xe = ex = x$ for all elements x of the groupoid. By an *absorption equation* we mean an equation $\langle u, v\rangle$ such that $u \neq v$ and either u or v is a variable.

4.1. THEOREM. *Let A be a finite groupoid with zero o and unit e. Suppose that A is not commutative, is not associative and does not satisfy any absorption equation. Then A is inherently nonfinitely based.*

PROOF. Let V be a locally finite variety containing A. Denote by E_0 the equational theory of A and by E the equational theory of V. For a finite sequence $x_1, \dots, x_k$ of variables, denote by $x_1 \dots x_k$ the term $((x_1x_2)x_3)\dots x_k$. Denote by H the set of the terms $x_1 \dots x_k$, where $k \geq 1$ and $x_1, \dots, x_k$ are pairwise different variables. It is not difficult to verify that if $\langle u, v\rangle \in E_0$ and $u \in H$, then $u = v$.

Now suppose that E has a finite base B, and let q be a positive integer larger than the cardinality of $\mathbf{S}(u)$, for any $\langle u, v\rangle \in B \cup B^{-1}$. For any $i \geq 1$ denote by t_i the term which is the product of the first i variables in the sequence $x_1, \dots, x_q, x_1, \dots, x_q, x_1, \dots, x_q, \dots$. Since E_0 is the equational theory of a locally finite variety, there exist two indexes $i < j$ such that $\langle t_i, t_j\rangle \in E_0$. There exists a derivation $u_0, \dots, u_k$ of $\langle t_i, t_j\rangle$ based on B. However, one can easily see that whenever u is a term such that $f(u) \subseteq t_i$ for a substitution f, then $u \in H$; consequently, $u_0 = u_1 = \dots = u_k$ and we get a contradiction with $t_i \neq t_j$. □

Comments

It has been proved in Lyndon [51] that every two-element algebra of finite signature is finitely based. A more simple proof is given in Berman [80].

Murskij [75] proves that a random finite groupoid is finitely based in the following sense: if $n(k)$ is the number of groupoids with a fixed underlying set of k elements and $m(k)$ is the number of groupoids with the same underlying set that are finitely based then $\lim_{n\to\infty} \frac{m(k)}{n(k)} = 1$.

CHAPTER 12

ALGORITHMS IN UNIVERSAL ALGEBRA

1. Turing machines

Many problems in mathematics can be reformulated in the following way: Given a finite alphabet A, the question is to find a partial function f, enjoying a particular property, from the set A^* of words over A into A^*. A proof of the existence of f is not satisfactory in many cases: the proof may not give us any way how to actually compute $f(w)$, given a word w in the domain of f. We need a set of rules, on the basis of which it would be possible to construct a machine accepting as an input arbitrary words $w \in A^*$, producing $f(w)$ in a finite number of steps for any word w belonging to the domain of f, and perhaps working forever for the other words w. Such a set of rules is called an *algorithm* (over A).

This informal definition may be sufficient if we investigate algorithms from the positive point of view. But we need a mathematically adequate definition if we want to prove that a particular problem has no algorithmic solution. The notion of a Turing machine will serve the purpose.

Let A be a finite alphabet. Let Q be a finite set disjoint with A and containing two distinguished elements $\mathbf{I}$ and $\mathbf{H}$. Let $\mathbf{L}, \mathbf{R}$ and $\mathbf{O}$ be three symbols not belonging to $A \cup Q$; put $A' = A \cup \{\mathbf{O}\}$. By a *Turing machine* over A, with states in Q, we mean a mapping T of $(Q \setminus \{\mathbf{H}\}) \times A'$ into $Q \times (A' \cup \{\mathbf{L}, \mathbf{R}\})$. The elements of Q are called the *states* of T; $\mathbf{I}$ is the *initial state*, and $\mathbf{H}$ is the *halting state*. The symbol $\mathbf{O}$ is called the *empty symbol*. The quadruples $\langle s, a, s', e\rangle$ such that $T(\langle s, a\rangle) = \langle s', e\rangle$ are called the *instructions* of T.

By a *configuration* for T we mean a word w over $A' \cup Q$ which can be written as $w = usav$ for a state s, a symbol $a \in A'$ and two words u, v over A'; the state s is called the state of the configuration w. By a *halting configuration* we mean a configuration, the state of which is the halting state.

Given a non-halting configuration $w = usav$ for T, we define a new configuration w', called the configuration *next to* w and denoted by $T[w]$, as follows:

(1) If $T(\langle s, a\rangle) = \langle s', b\rangle$ for a symbol $b \in A'$, put $w' = us'bv$.
(2) If $T(\langle s, a\rangle) = \langle s', \mathbf{L}\rangle$ and u is empty, put $w' = s'\mathbf{O}av$.
(3) If $T(\langle s, a\rangle) = \langle s', \mathbf{L}\rangle$ and $u = \bar{u}b$ for some $b \in A'$, put $w' = \bar{u}s'bav$.
(4) If $T(\langle s, a\rangle) = \langle s', \mathbf{R}\rangle$ and v is empty, put $w' = uas'\mathbf{O}$.
(5) If $T(\langle s, a\rangle) = \langle s', \mathbf{R}\rangle$ and v is nonempty, put $w' = uas'v$.

Given an arbitrary configuration w for T, we define a sequence (either finite or infinite) of configurations $T^0[w], T^1[w], \dots$ as follows: $T^0[w] = w$; if $T^i[w]$ is a non-halting configuration, then $T^{i+1}[w] = T[T^i[w]]$; otherwise, if $T^i[w]$ is halting, then it is the last member of the sequence. If this sequence is finite, with the last member v, we say that T *halts from* w (at v).

Given a Turing machine T over A, we can define a partial function f from A^* into A^* (a unary partial operation on A^*) as follows. Let $u \in A^*$. Put $w = \mathbf{IO}u$, so that w is a configuration for T. If T halts from w at a configuration v, let $f(u)$ be the word obtained from v by deleting the symbols not belonging to A; if T does not halt from w, let $f(u)$ be undefined. We say that f is the partial function realized by T. A partial function from A^* to A^* is said to be *computable* if it is realized by some Turing machine.

More generally, an n-ary partial operation f on A^* is said to be computable if there exists a Turing machine T over A such that $f(u_1, \dots, u_n) = u$ if and only if T halts from the configuration $\mathbf{IO}u_1 \dots \mathbf{O}u_n$ at a configuration v and u is obtained from v by deleting the symbols not belonging to A.

It is clear that every computable function can be computed by an algorithm in the intuitive sense. We subscribe to the *Church-Turing Thesis*, according to which the functions that can be computed via any mechanical process, are exactly the functions computable in the above sense. We will use this thesis in two ways: We claim that a function is computable if we have an algorithm computing it in an intuitive sense. Also, we claim that a function is not computable by any means if it is not computable according to the above definition.

A subset S of A^* is said to be *recursive* if there exists an algorithm deciding for every word $w \in A^*$ whether $w \in S$. In other words, S is recursive if its characteristic function is computable; the characteristic function of S assigns a specified nonempty word to every word from S, and assigns the empty word to the words not in S.

A subset of A^* is said to be *recursively enumerable* if it is the range of a computable function. It is easy to see that a subset S of A^* is recursive if and only if both S and $A^* \setminus S$ are recursively enumerable.

So far we considered only algorithms, operating with words over a finite alphabet. But it should be clear that the set A^* can be replaced by any set of objects constructible in the sense that they can be coded by words. For example, we could take the set of all matrices with rational coefficients, the set ot integers, the set of Turing machines over a given finite alphabet with states from a given set of constructible objects, or the set of all finite algebras of a given finite signature, with the underlying sets contained in the set of nonnegative integers.

1.1. THEOREM. *Let A be a nonempty finite alphabet. There exists a computable binary partial operation h on A^* such that for every computable partial function g from A^* into A^* there exists a word $u \in A^*$ with this property: an* *arbitrary word $v \in A^*$ belongs to the domain of g if and only if $\langle u, v\rangle$ belongs to the domain of h, and $g(u) = h(u, v)$ if it does.*

PROOF. Arrange into an infinite sequence $T_0, T_1, \dots$ all Turing machines over A in some standard, 'computable' way. Similarly, arrange into an infinite sequence $w_0, w_1, \dots$ all words over A. The algorithm realizing h can be defined in the following way. Let u, v be two words over A. Find the index i such that $u = w_i$. Try to compute $g(v)$, where g is the partial function realized by T_i. If the computation halts, output $h(u, v) = g(v)$. □

1.2. THEOREM. *Let A be a nonempty finite alphabet. There exists a computable partial function f from A^* into A^* such that the range of f consists of just two words and f cannot be extended to a computable, everywhere defined function.*

PROOF. Take a symbol $a \in A$, and define a mapping p of A^* into A^* as follows: $p(w)$ is the empty word for any nonempty word w, and $p(w) = a$ for w empty. Thus $p(w) \neq w$ for all words w. Let h be a computable binary partial operation with the property stated in 1.1. Define $f(w) = p(h(w, w))$ for all the words w such that $f(w, w)$ is defined; for the other words w let $f(w)$ be undefined. Clearly, f is computable and the range of f consists of just two words. Suppose that f can be extended to a computable, everywhere defined function g. There exists a word u such that $g(v) = h(u, v)$ for all v. In particular, $g(u) = h(u, u)$ and hence $g(u) = f(u) = p(h(u, u)) \neq h(u, u) = g(u)$, a contradiction. □

1.3. THEOREM. *There exists a Turing machine T over a one-element alphabet such that the set of the configurations for T from which T halts is not recursive.*

PROOF. It follows from 1.2. □

2. Word problems

Let V be a variety of a finite signature σ. An algebra $A \in V$ is said to be *finitely presented* (with respect to V) if there exist a positive integer n and a binary relation r in the algebra $\mathbf{T}_n$ of terms over $\{x_1, \dots, x_n\}$ such that A is isomorphic to the factor $\mathbf{T}_n/R$, where R is the congruence of $\mathbf{T}_n$ generated by the the equations in variables $x_1, \dots, x_n$ that belong either to r or to the equational theory of V; we say that the pair, consisting of the number n and the relation r, is a *finite presentation* of A (in V).

We say that a finitely presented algebra in V, given by its finite presentation (n, r), has *solvable word problem* if there is an algorithm deciding for every $\langle u, v\rangle \in \mathbf{T}_n \times \mathbf{T}_n$ whether $\langle u, v\rangle$ belongs to the congruence of $\mathbf{T}_n$ generated by the equations in variables $x_1, \dots, x_n$ that belong either to r or to the equational theory of V. We say that a variety V has solvable word problem if every finitely presented algebra in V has solvable word problem. We say that a variety V has globally solvable word problem if there exists an algorithm, deciding for any positive integer n, any finite relation r in $\mathbf{T}_n$ and any $\langle u, v\rangle \in \mathbf{T}_n \times \mathbf{T}_n$ whether $\langle u, v\rangle$ belongs to the congruence of $\mathbf{T}_n$ generated by the equations in variables $x_1, \dots, x_n$ that belong either to r or to the equational theory of V. Clearly, global solvability of the word problem implies solvability.

2.1. THEOREM. *Let V be a variety of algebras of a finite signature and let $A \in V$. The following three conditions are equivalent:*

(1) *A is finitely presented in V;*
(2) *A is a reflection of a finite partial algebra in V;*
(3) *There exists a finite nonempty subset S_0 of A such that for every subset S containing S_0, A together with $\mathbf{id}_S$ is a reflection of $A \restriction S$ in V.*

PROOF. (1) implies (3): Let (n, r) be a finite presentation of A in V, where $r = \{\langle u_1, v_1\rangle, \dots, \langle u_k, v_k\rangle\}$. Denote by R the congruence of $\mathbf{T}_n$ generated by the equations in $x_1, \dots, x_n$ that belong either to r or to the equational theory of V. We can assume that $A = \mathbf{T}_n/R$. Denote by S_0 the finite subset of A consisting of the elements t/R where $t \in \mathbf{T}_n$ is either an element of $\{x_1, \dots, x_n\}$ or a subterm of at least one of the terms $u_1, \dots, u_k, v_1, \dots, v_k$. One can easily see that for every $S_0 \subseteq S \subseteq A$, $\mathbf{id}_S : A \restriction S \to A$ is a reflection of $A \restriction S$ in V.

It remains to prove that (2) implies (1). Let $f : Q \to A$ be a reflection of a finite partial algebra Q in V. Denote by $a_1, \dots, a_n$ the elements of Q and denote by r the set of the pairs $\langle F(x_{i_1}, \dots, x_{i_m}), x_{i_{m+1}}\rangle$ where F is a symbol of arity m in the signature, $i_1, \dots, i_{m+1}$ are elements of $\{1, \dots, n\}$, and $F_Q(a_{i_1}, \dots, a_{i_m}) = a_{i_{m+1}}$. It is easy to see that (n, r) is a finite presentation of A in V. $\square$

Let V be a variety of a finite signature. We say that the *embedding problem* for V is solvable if there exists an algorithm, deciding for every finite partial algebra of the given signature whether it can be embedded into an algebra from V.

2.2. THEOREM. *A variety V of a finite signature has globally solvable word problem if and only if it has solvable embedding problem.*

PROOF. Assume that V has globally solvable word problem. Let Q be a finite partial algebra. Denote by $a_1, \dots, a_n$ the elements of Q and define a finite relation r on $\mathbf{T}_n$ in the same way as in the proof of (2) $\to$ (1) in 2.1. Denote by R the congruence of $\mathbf{T}_n$ generated by the equations in $x_1, \dots, x_n$ that belong either to r or to the equational theory of V. Put $f(a_i) = x_i/R$ for all i; one can easily see that $f : Q \to \mathbf{T}_n/R$ is a reflection of Q in V, and f is injective if and only if Q can be embedded into an algebra from V. Now (n, r) is a finite presentation for $\mathbf{T}_n/R$, f is injective if and only if $\langle x_i, x_j\rangle \in R$ implies $i = j$, and we are able to decide this question.

Conversely, assume that V has solvable embedding problem. Let n be a positive integer and $r = \{\langle u_1, v_1\rangle, \dots, \langle u_n, v_n\rangle\}$ be a finite relation on $\mathbf{T}_n$. Let $u, v \in \mathbf{T}_n$. We need to find a way how to decide whether $\langle u, v\rangle$ belongs to the congruence R generated by the equations in $x_1, \dots, x_n$ that belong either to r or to the equational theory of V. Denote by Y the set of the terms that either belong to $\{x_1, \dots, x_n\}$ or are subterms of one of the terms $u, v, u_1, \dots, u_n, v_1, \dots, v_n$. Denote by $\bar{r}$ the congruence of the partial algebra $\mathbf{T}_n \restriction Y$ generated by r; denote by Q the factor, and by q the corresponding canonical homomorphism onto Q. We are going to prove that $\langle u, v\rangle$ does not belong to R if and only if there exists a homomorphism g of Q onto a partial algebra Q' such that $g(u/R) \neq g(v/R)$ and Q' can be embedded into an algebra from V. This will give us the desired algorithm.

Let $\langle u, v\rangle \notin R$. Since $R \cap (Y \times Y)$ is a congruence of $\mathbf{T}_n \restriction Y$ and $r \subseteq R \cap (Y \times Y)$, we have $\bar{r} \subseteq R \cap (Y \times Y)$. It follows that there exists a homomorphism $g : Q \to \mathbf{T}_n/R$ such that $gq(t) = t/R$ for all $t \in Y$. We can put $Q' = (\mathbf{T}_n/R) \restriction S$, where S is the range of g; the partial algebra Q' can be embedded into the algebra $\mathbf{T}_n/R \in V$.

Conversely, let g and Q' be as above and let f be an embedding of Q' into an algebra $B \in V$. There exists a unique homomorphism $h : \mathbf{T}_n \to B$ such that $h(x_i) = fgq(x_i)$ for all i. The composition fgq is the restriction of h to Y, since both these homomorphisms of $\mathbf{T}_n \restriction Y$ into B coincide on a generating subset. From this we get $r \subseteq \mathbf{ker}(h)$. Since $\mathbf{T}_n/\mathbf{ker}(h) \in V$, we get $R \subseteq \mathbf{ker}(h)$. Now $h(u) = fgq(u) \neq fgq(v) = h(v)$ and hence $\langle u, v\rangle \notin R$. □

2.3. Example. Clearly, the variety of all algebras of a given finite signature has globally solvable word problem.

A finite partial groupoid A can be embedded into a commutative groupoid if and only if $ab = ba$ for any pair a, b of elements of A such that both ab and ba are defined. We can see that the variety of commutative groupoids has solvable embedding problem. Consequently, it has globally solvable word problem.

Similarly, the variety of idempotent groupoids and also the variety of commutative idempotent groupoids has globally solvable word problem.

3. The finite embedding property

A variety V is said to have the *finite embedding property* if for every algebra $A \in V$ and every finite subset S of A, the partial algebra $A \restriction S$ can be embedded into a finite algebra in V.

Next we give two equivalent formulations for this property. An algebra A is said to be *residually finite* if for every pair a, b of different elements of A there exists a congruence r of A such that $\langle a, b\rangle \notin r$ and A/r is finite.

3.1. Theorem. *The following three conditions are equivalent for a variety V of a finite signature:*

1. *V has the finite embedding property;*
2. *For every finitely presented algebra $A \in V$ and every finite subset S of A, the partial algebra $A \restriction S$ can be embedded into a finite algebra in V;*
3. *Every finitely presented algebra in V is residually finite.*

PROOF. Obviously, (1) implies (2). (2) implies (3): Let A be a finitely presented algebra in V and let a, b be two different elements of A. According to 2.1, there exists a finite subset S of A such that $a, b \in S$ and $\mathbf{id}_S : A \restriction S \to A$ is a reflection of $A \restriction S$ in V. By (2) there exists a finite algebra $B \in V$ and an injective homomorphism $f : A \restriction S \to B$. By the definition of reflection, there exists a homomorphism $g : A \to B$ such that $f \subseteq g$. We can put $r = \mathbf{ker}(g)$.

(3) implies (2): For every pair a, b of different elements of S take a congruence $r_{a,b}$ of A such that $\langle a, b\rangle \notin r_{a,b}$ and $A/r_{a,b}$ is finite. Denote by B the product of all the algebras $A/r_{a,b}$ obtained in this way, so that B is a finite algebra in V. For $h \in A$ let $f(h)$ be the element of B with $f(h)(a, b) = h/r_{a,b}$ for all a, b. Then f is a homomorphism of A into B and the restriction of f to S is an injective homomorphism of $A \restriction S$ into B.

(2) implies (1): Let $A \in V$ and let S be a finite subset of A. There exists a reflection $f : A \restriction S \to B$ of $A \restriction S$ in V. By the definition of reflection there exists a homomorphism $g : B \to A$ such that $\mathbf{id}_S = gf$. Consequently, f is injective. By 2.1, B is finitely presented in V and so, by (2), there exists a finite algebra $C \in V$ and an injective homomorphism $h : B \restriction Y \to C$, where Y is the range of f. The composition hf is an injective homomorphism of $A \restriction S$ into C. □

3.2. THEOREM. *Let V be a finitely based variety of a finite signature with the finite embedding property. Then V has globally solvable word problem.*

PROOF. Let B be a finite base for the equations of V. Let n be a positive integer, r be a finite relation on $\mathbf{T}_n$ and let $u, v \in \mathbf{T}_n$. We need to decide whether $\langle u, v\rangle$ belongs to the congruence R of $\mathbf{T}_n$ generated by the equations in $x_1, \dots, x_n$ that belong either to r or to the equational theory of V.

Denote by U the least relation on the algebra $\mathbf{T}$ of terms with the following three properties:

(1) $r \subseteq U$;
(2) If $\langle t_1, t_2\rangle \in B$ then $\langle f(t_1), f(t_2)\rangle \in U$ for every substitution f;
(3) If $\langle t_1, t_2\rangle \in U$ then $\langle L(t_1), L(t_2)\rangle \in U$ for every lift L.

Clearly, U is a recursive set of equations.

By an admissible sequence we mean (just for the purpose of this proof) a finite nonempty sequence $t_1, \dots, t_m$ of terms such that for any $i = 2, \dots, m$ either $t_{i-1} = t_i$ or $\langle t_{i-1}, t_i\rangle \in U$ or $\langle t_i, t_{i-1}\rangle \in U$. Clearly, the set of admissible sequences is a recursive set of finite sequences of terms.

Denote by s_1 the set of the ordered pairs $\langle t_1, t_2\rangle$ such that t_1 is the first and t_2 is the last member of an admissible sequence, and put $s_2 = s_1 \cap (\mathbf{T}_n \times \mathbf{T}_n)$. We are going to prove that $s_2 = R$. Clearly, s_2 is a congruence of $\mathbf{T}$ and $\mathbf{T}/s_1 \in V$. Hence s_2 is a congruence of $\mathbf{T}$ and $\mathbf{T}/s_2 \in V$. Since $r \subseteq s_2$, we get $R \subseteq s_2$. There exists a homomorphism $f : \mathbf{T} \to \mathbf{T}_n/R$ such that $f(x_i) = x_i/R$ for $i = 1, \dots, n$. Clearly, $R = \mathbf{ker}(f) \cap (\mathbf{T}_n \times \mathbf{T}_n)$. Since $r \subseteq \mathbf{ker}(f)$ and $\mathbf{T}/\mathbf{ker}(f) \in V$, we have $U \subseteq \mathbf{ker}(f)$, so that $s_1 \subseteq \mathbf{ker}(f)$ and then $s_2 \subseteq R$.

We have proved that $\langle u, v\rangle$ belongs to R if and only if u is the first and v is the last member of an admissible sequence.

Similarly as in the proof of the converse implication in 2.2 one can construct a finite partial algebra Q and its congruence $\bar{r}$ with this property: If $P_1, \dots, P_k$ are (up to isomorphism) all the finite partial algebras P for which there exists a homomorphism g of Q onto P with $g(u/\bar{r}) \neq g(v/\bar{r})$, then $\langle u, v\rangle \notin R$ if and only if at least one of the partial algebras P_i can be embedded into an algebra from V. It follows from the finite embedding property that $\langle u, v\rangle \notin R$ if and only if at least one of the partial algebras $P_1, \dots, P_k$ can be embedded into a finite algebra from V.

We have obtained two different characterizations of the relation R. The desired algorithm can be constructed by their combination. Let $p_1, p_2, \dots$ be a standard ordering of all finite sequences of terms and let $A_1, A_2, \dots$ be a standard ordering of all finite algebras of the given signature. For every $i = 1, 2, \dots$ we can answer the following two questions: (i) Is p_i an admissible sequence and are u the first and v the last members of this sequence? (ii) Is $A_i \in V$ (this can be verified, since V is finitely based) and can one of the partial algebras $P_1, \dots, P_k$ be embedded into A_i? After a finite number of steps we must obtain a positive answer. If the positive answer is obtained for p_i, the pair $\langle u, v\rangle$ belongs to R, while a positive answer for A_i means that $\langle u, v\rangle$ does not belong to R. □

3.3. Example. Theorem 3.2 can be used to show that the variety of lattices, the variety of Abelian groups, the variety of quasigroups and the variety of loops have globally solvable word problem.

4. Unsolvability of the word problem for semigroups

Let T be a Turing machine over a finite nonempty alphabet A, with the set of states Q. Denote by P the semigroup of nonempty words over $A \cup \{\mathbf{O}, \mathbf{L}, \mathbf{R}\}$, and define five finite relations $r_1, \dots, r_5$ in P:

r_1 is the set of the pairs $\langle qa, pb\rangle$ where $T(\langle q, a\rangle) = \langle p, b\rangle$ and $b \in A \cup \{\mathbf{O}\}$;
r_2 is the set of the pairs $\langle bqa, pba\rangle$ where $T(\langle q, a\rangle) = \langle p, \mathbf{L}\rangle$ and $b \in A \cup \{\mathbf{O}\}$;
r_3 is the set of the pairs $\langle qab, apb\rangle$ where $T(\langle q, a\rangle) = \langle p, \mathbf{R}\rangle$ and $b \in A \cup \{\mathbf{O}\}$;
$r_4 = \{\langle \mathbf{L}, \mathbf{LO}\rangle\}$;
$r_5 = \{\langle \mathbf{OR}, \mathbf{R}\rangle\}$.

Denote by r the union of these five relations and by R the congruence of P generated by r.

4.1. Lemma. *If w and w' are two configurations such that $T[w] = w'$, then $\langle \mathbf{L}w\mathbf{R}, \mathbf{L}w'\mathbf{R}\rangle \in R$.*

Proof. It is easy. □

By an inessential extension of a word $w \in P$ we mean any word $\mathbf{O}^k w \mathbf{O}^m$, where k, m are nonnegative integers.

4.2. Lemma. *Let w be a halting and w' be a non-halting configuration for T. Then $\langle \mathbf{L}w\mathbf{R}, \mathbf{L}w'\mathbf{R}\rangle \in R$ if and only if there exists a finite sequence $w_1, \dots, w_n$ of configurations such that w_1 is an inessential extension of w, w_n is an inessential extension of w' and $w_i = T[w_{i-1}]$ for all $i = 2, \dots, n$.*

PROOF. The converse follows from 4.1. It remains to prove the direct implication. For $i = 1, \dots, 5$ denote by $\bar{r}_i$ the set of the ordered pairs $\langle s, t\rangle \in P \times P$ such that $\langle s, t\rangle = \langle us_0v, ut_0v\rangle$ for some $\langle s_0, t_0\rangle \in r_i$ and some words u, v. Put $\bar{r} = \bar{r}_1 \cup \bar{r}_2 \cup \bar{r}_3 \cup \bar{r}_4 \cup \bar{r}_5$. By an admissible sequence we mean (just for the purpose of this proof) a finite sequence $u_1, \dots, u_m$ of words from P such that $u_1 = \mathbf{L}w\mathbf{R}$, $u_m = \mathbf{L}w'\mathbf{R}$ and for any i=2, $\dots, m$, either $\langle u_{i-1}, u_i\rangle \in \bar{r}$ or $\langle u_i, u_{i-1}\rangle \in \bar{r}$. It is easy to see that there exists at least one admissible sequence. Also, it is easy to see that if $u_1, \dots, u_m$ is an admissible sequence then $u_i = \mathbf{L}\bar{u}_i\mathbf{R}$ for a word $\bar{u}_i$ not containing $\mathbf{L}, \mathbf{R}$ and containing precisely one occurrence of a state of T. However, $\bar{u}_i$ is not necessarily a configuration, as the state can be the last symbol in $\bar{u}_i$.

Let $u_1, \dots, u_k$ be a minimal admissible sequence. Suppose that for some i we have $\langle u_{i-1}, u_i\rangle \notin \bar{r}_4$ and $\langle u_i, u_{i+1}\rangle \in \bar{r}_4$. The pair $\langle u_i, u_{i-1}\rangle$ does not belong to $\bar{r}_4$ (then we would have $u_{i-1} = u_{i+1}$). One can easily see that the sequence $u_1, \dots, u_{i-1}, \mathbf{LO}\bar{u}_{i-1}\mathbf{R}, u_{i+1}, \dots, u_k$ is also admissible, of the same length as $u_1, \dots, u_k$. So, we can assume that in $u_1, \dots, u_k$, all applications of $\langle \mathbf{L}, \mathbf{LO}\rangle$ come at the beginning; similarly we can assume that the applications of $\langle \mathbf{LO}, \mathbf{L}\rangle$ come at the end, all applications of $\langle \mathbf{R}, \mathbf{OR}\rangle$ come at the beginning and all applications of $\langle \mathbf{OR}, \mathbf{R}\rangle$ come at the end. Let $u_c, \dots, u_d$ be the middle, the more interesting part of $u_1, \dots, u_k$: for $i = c+1, \dots, d$ we have either $\langle u_{i-1}, u_i\rangle \in \bar{r}_1 \cup \bar{r}_2 \cup \bar{r}_3$ or $\langle u_i, u_{i-1}\rangle \in \bar{r}_1 \cup \bar{r}_2 \cup \bar{r}_3$.

Suppose that there exists an index i with $\langle u_i, u_{i-1}\rangle \in \bar{r}_1 \cup \bar{r}_2 \cup \bar{r}_3$; let i be the maximal such index. Clearly, the halting state does not occur in u_i. But it occurs in u_k, and hence in u_d. We get $i < d$ and $\langle u_i, u_{i+1}\rangle \in \bar{r}_1 \cup \bar{r}_2 \cup \bar{r}_3$. But then $u_{i-1} = u_{i+1}$, a contradiction with the minimality of k.

Put $n = d - c + 1$ and define $w_1, \dots, w_n$ by $u_c = \mathbf{L}w_1\mathbf{R}$, $u_{c+1} = \mathbf{L}w_2\mathbf{R}$, $\dots$, $u_d = \mathbf{L}w_n\mathbf{R}$. Clearly, w_1 is an inessential extension of w, w_n is an inessential extension of w' and it should be now evident that $w_i = T[w_{i-1}]$ for $i = 2, \dots, n$. □

4.3. LEMMA. *There exists a Turing machine T such that the finitely presented semigroup P/R has unsolvable word problem.*

PROOF. This follows from 1.3 and 4.2. □

4.4. THEOREM. *There exists a two-generated finitely presented semigroup with unsolvable word problem.*

PROOF. By 4.3 there exist a finite nonempty alphabet $B = \{a_1, \dots, a_n\}$ and a finite relation r on the semigroup P of nonempty words over B such that the congruence R of P generated by r is not a recursive set. Denote by U the semigroup of nonempty words over a two-element alphabet $\{a, b\}$. For every $i = 1, \dots, n$ put $w_i = aba^{i+1}b^{i+1}$, and define an injective mapping $u \mapsto \bar{u}$ of P into U as follows: if $u = a_{i_1} \dots a_{i_m}$, then $\bar{u} = w_{i_1} \dots w_{i_m}$. Put $s = \{\langle \bar{u}, \bar{v}\rangle : \langle u, v\rangle \in r\}$, so that s is a finite relation on U. Denote by S the congruence of U generated by s.

Denote by S_1 the set of the pairs $\langle p\bar{u}q, p\bar{v}q\rangle$ where $\langle u,v\rangle \in r$ and p,q are words over $\{a,b\}$.

Suppose $\langle \bar{u}, w\rangle \in S_1$ for a word $u = a_{i_1}\dots a_{i_m} \in P$ and a word $w \in U$. We have $\bar{u} = p\bar{u}_0 q$ and $w = p\bar{v}_0 q$ for a pair $\langle u_0, v_0\rangle \in r$ and some words p,q over $\{a,b\}$. One can easily see that there exist indexes j,k $(0 \le j \le k \le m)$ such that $p = w_{i_1}\dots w_{i_j}$, $\bar{u}_0 = w_{i_{j+1}}\dots w_{i_k}$ and $q = w_{i_{k+1}}\dots w_{i_m}$. We get $w = \bar{v}$, where $a_{i_1}\dots a_{i_j} v_0 a_{i_{k+1}}\dots a_{i_m}$, and $\langle u,v\rangle \in R$.

Quite similarly, if $\langle w, \bar{v}\rangle \in S_1$ for a word $w \in U$ and a word $v \in P$, then there is a word $u \in P$ such that $w = \bar{u}$ and $\langle u,v\rangle \in R$.

Define a relation S_2 on U as follows: $\langle u,v\rangle \in S_2$ if and only if there exists a finite sequence $u_1,\dots,u_m$ such that $u = u_1$, $v = u_m$ and for all $i = 2,\dots,m$, either $\langle u_{i-1}, u_i\rangle \in S_1$ or $\langle u_i, u_{i-1}\rangle \in S_1$. It follows that $\langle \bar{u}, \bar{v}\rangle \in S_2$ implies $\langle u,v\rangle \in R$. Evidently, S_2 is a congruence of U and hence $S \subseteq S_2$.

Let $u,v \in P$. We have proved that $\langle \bar{u}, \bar{v}\rangle \in S$ implies $\langle u,v\rangle \in R$. Conversely, it is easy to see that $\langle u,v\rangle \in R$ implies $\langle \bar{u}, \bar{v}\rangle \in S$. So, if we could decide whether $\langle \bar{u}, \bar{v}\rangle \in S$, then we would be also able to decide whether $\langle u,v\rangle \in R$. The semigroup U/S has undecidable word problem. □

5. An undecidable equational theory

5.1. THEOREM. *There exists an undecidable, finitely based equational theory of the signature consisting of two unary symbols.*

PROOF. Denote the two unary symbols by F and G. By 4.4 there exists a finite relation r on the semigroup P of nonempty words over $\{F,G\}$ such that the congruence R of P generated by r is not a recursive set. Let E be the equational theory based on the equations $\langle ux, vx\rangle$ where $\langle u,v\rangle \in r$ and x is a variable. One can easily see that an equation $\langle t_1, t_2\rangle$ belongs to E if and only if either $t_1 = t_2$ or $\langle t_1, t_2\rangle = \langle ux, vx\rangle$ for a variable x and a pair $\langle u,v\rangle \in R$. Consequently, E is not a recursive set. □

Comments

The results on word problems and the finite embedding property are due to T. Evans.

We did not include the most important and sophisticated result in this direction, that of McKenzie [96], [96a] and [96b]: There is no algorithm deciding for any finite algebra of finite signature whether it is finitely based. This has been accomplished by assigning a finite algebra $A(T)$ of a finite signature to any Turing machine T in such a way that $A(T)$ is finitely based if and only if T halts. At the same time, McKenzie proves that there is no algorithm deciding for any finite algebra A of finite signature whether the variety generated by A is residually finite.

CHAPTER 13

TERM REWRITE SYSTEMS

A broadly discussed question of equational logic is to find ways to decide which equations are consequences of a given finite set of equations, that is, to establish decidability of a given finitely based equational theory. This question is undecidable in general, so attention has been focused on special cases, as general as possible, for which there is hope of finding an algorithm. Evans [51] and Knuth and Bendix [70] introduced the technique of term rewriting, which has been further developed in a large number of papers; see Dershowitz and Jouannaud [90] for an overview and for an extensive bibliography. We are going to explain in this chapter the basics of term rewriting and present an alternative but closely related technique, that of perfect bases, introduced in Ježek and McNulty [95b].

By a *normal form function* for an equational theory E we mean a mapping ν of the set of terms into itself, satisfying the following three conditions:

(nf1) $u \approx v \in E$ if and only if $\nu(u) = \nu(v)$;
(nf2) $t \approx \nu(t) \in E$ for all terms t;
(nf3) $\nu(\nu(t)) = \nu(t)$ for all terms t.

An equational theory E is *decidable* if and only if it has a computable normal form function.

1. Unification

By a *unifier* of a finite collection $t_1, \dots, t_k$ of terms we mean a substitution f with $f(t_1) = \dots = f(t_k)$. By a *minimal unifier* of $t_1, \dots, t_k$ we mean a unifier f such that for any other unifier g of $t_1, \dots, t_k$ there exists a substitution h with $g = hf$. Easily, the term $f(t_1) = \dots = f(t_k)$ is uniquely determined by $t_1, \dots, t_k$ up to similarity; it will also be called the unifier of $t_1, \dots, t_k$ (from the context it will be always clear if the unifier is a term, or a substitution).

1.1. Theorem. *If a finite collection of terms has a unifier, then it has a minimal one. There is an algorithm accepting any finite k-tuple $t_1, \dots, t_k$ of terms as an input and outputting either a minimal unifier f of $t_1, \dots, t_k$ or else the information that the k-tuple has no unifier. If f is output and $u = f(t_1) = \dots = f(t_k)$, where $k \geq 1$, then $\mathbf{S}(u) \subseteq \mathbf{S}(t_1) \cup \dots \cup \mathbf{S}(t_k)$ and $f(x) = x$ for any $x \in X \smallsetminus (\mathbf{S}(t_1) \cup \dots \cup \mathbf{S}(t_k))$.*

Proof. Clearly, it suffices to prove the theorem for $k = 2$.

Denote by S the set of the terms that are subterms of either t_1 or t_2, and by $\equiv$ the smallest equivalence on S such that $t_1 \equiv t_2$ and

$$F(u_1, \dots, u_n) \equiv F(v_1, \dots, v_n) \Rightarrow u_i \equiv v_i \text{ for all } i.$$

Define a binary relation r on S by $\langle u, v\rangle \in r$ if and only if there are terms u', v' with $u \equiv u'$ and $v \equiv v'$, such that u' is a proper subterm of v'. Denote by R the transitive closure of r.

If there exists a substitution f with $f(t_1) = f(t_2)$, then clearly $f(u) = f(v)$ whenever $u \equiv v$ and, consequently, the following two conditions are satisfied:

(1) if $F(u_1, \dots, u_n) \equiv G(v_1, \dots, v_m)$, then $F = G$;
(2) there is no term u with $\langle u, u\rangle \in R$.

Conversely, we will show that if these two conditions are satisfied, then the pair t_1, t_2 has a minimal unifier and we are going to construct it.

Let (1) and (2) be satisfied. Then R is an irreflexive, antisymmetric and transitive relation, i.e., a strict order on S. Define a sequence $M_0, M_1, \dots$ of pairwise disjoint subsets of S recursively in this way: M_i is the set of the elements of $S \smallsetminus (M_0 \cup \dots \cup M_{i-1})$ that are minimal with respect to R. For every term $u \in S$ there is precisely one index i with $u \in M_i$; this i will be called (locally in this proof) the rank of u. Clearly, $u \equiv v$ implies that the terms u and v have the same rank.

For every term $u \in S$, we will define a term $f(u)$ by induction on the rank of u.

Let u be of rank 0, so that u is either an element of X or a constant. If there is a constant c with $u \equiv c$, then c is also of rank 0, and c is unique; put $f(u) = c$ in that case. If there is no such c, then $u \equiv v$ implies that v is an element of X, and of rank 0; with respect to a fixed well ordering of X, take the first element x of X with $u \equiv x$ and put $f(u) = x$.

Now let u be of rank $i + 1$, and suppose that $f(v)$ has been defined for all terms v of ranks at most i in such a way that $f(v) = f(v')$ whenever $v \equiv v'$. Since u is not of rank 0, even if it is an element of X, we have $u \equiv F(v_1, \dots, v_n)$ for a symbol F of arity $n \geq 0$ and some terms $v_1, \dots, v_n$ of ranks $\leq i$. The symbol F is uniquely determined by u, and the terms v_j are determined uniquely up to $\equiv$. So, it makes sense to define $f(u) = F(f(v_1), \dots, f(v_n))$ and it is clear that if $u \equiv u'$, then $f(u) = f(u')$.

In particular, we have $f(t_1) = f(t_2)$. Put, moreover, $f(x) = x$ for any $x \in X \smallsetminus (\mathbf{S}(t_1) \cup \mathbf{S}(t_2))$. It is not difficult to see that f can be uniquely extended to a substitution, and this extension is a minimal unifier of the pair t_1, t_2. □

By a *unifying k-tuple* for a finite collection $t_1, \dots, t_k$ of terms we mean a k-tuple $f_1, \dots, f_k$ of substitutions with $f_1(t_1) = \dots = f_k(t_k)$. By a *minimal unifying k-tuple* for $t_1, \dots, t_k$ we mean a unifying k-tuple $f_1, \dots, f_k$ such that for any other unifying k-tuple $g_1, \dots, g_k$ for $t_1, \dots, t_k$ there exists a substitution h with $g_1 = hf_1, \dots, g_k = hf_k$. Easily, the term $f_1(t_1) = \dots = f_k(t_k)$ is uniquely determined by $t_1, \dots, t_k$ up to similarity; it will be called the *multi-unifier* of $t_1, \dots, t_k$.

1.2. THEOREM. *Let X be infinite. If a finite collection of terms has a unifying k-tuple, then it has a minimal one. There is an algorithm accepting any finite k-tuple $t_1, \dots, t_k$ of terms as an input and outputting either a minimal unifying k-tuple $f_1, \dots, f_k$ of $t_1, \dots, t_k$ or else the information that the terms have no unifying k-tuple.*

PROOF. Clearly, any of the terms $t_1, \dots, t_k$ can be replaced with a similar term without affecting the result. So, since X is infinite, we can assume that the sets $\mathbf{S}(t_1), \dots, \mathbf{S}(t_k)$ are pairwise disjoint. Under this assumption, the existence of a (minimal) unifying k-tuple is equivalent to that of the existence of a (minimal) unifier, so Theorem 1.1 can be applied. $\square$

1.3. EXAMPLE. If the set X is finite, then the minimal unifying pair for a given pair of terms need not exist even if there are some unifying pairs. Consider, for example, the pair $(xy)z$, $x(yz)$ of terms over the set $X = \{x, y, z\}$. This pair has a has a unifying pair, but no minimal one over X.

2. Convergent graphs

Let $\langle G, \to \rangle$ be a (directed) graph. (I.e., G is a nonempty set and $\to$ is a relation on G.) A finite nonempty sequence $a_0, \dots, a_k$ of vertices (i.e., elements of G) is called a *directed path* if $a_i \to a_{i+1}$ for all $i \in \{0, \dots, k-1\}$; it is called an *undirected path* if, for all i, either $a_i \to a_{i+1}$ or $a_{i+1} \to a_i$. In both cases we say that the path starts at a_0 and terminates at a_k, or that the path is from a_0 to a_k. Two vertices a, b are called *connected* if there exists an undirected path starting at a and terminating at b.

A vertex $a \in G$ is called *terminal* if there is no vertex b with $a \to b$. The graph $\langle G, \to \rangle$ is called *finitely terminating* if there is no infinite sequence $a_0, a_1, \dots$ of vertices with $a_i \to a_{i+1}$ for all $i \geq 0$. Clearly, a finitely terminating graph contains no cycles (and, in particular, no loops). If the graph is finitely terminating, then for every vertex a there exists a directed path starting at a and terminating at a terminal vertex.

The graph $\langle G, \to \rangle$ is called *confluent* if for any triple of vertices a, b, c such that there are two directed paths, one from a to b and the other from a to c, there exists a fourth vertex d, a directed path from b to d, and a directed path from c to d.

The graph $\langle G, \to \rangle$ is called *convergent* if it is both finitely terminating and confluent.

2.1. THEOREM. *Let $\langle G, \to \rangle$ be a convergent directed graph. Then every vertex of G is connected to precisely one terminal vertex of G.*

PROOF. Since the graph is finitely terminating, every vertex is connected to at least one terminal vertex (via a directed path). In order to prove the uniqueness, it remains to show that two different terminal vertices cannot be connected. Suppose there exists an undirected path $a_0, \dots, a_k$ such that $a_0 \neq a_k$ and both a_0 and a_k are terminal, and choose one for which the set $\{i \in \{1, \dots, k-1\} : a_i \to a_{i-1} \text{ and } a_i \to a_{i+1}\}$ has the least possible cardinality. The elements of this set of indices will be called peaks. Since both a_0 and a_k are terminal, we have $a_1 \to a_0$ and $a_{k-1} \to a_k$; from this one can easily see that there is at least one peak. Denote by i the first peak, so that $a_i \to a_{i-1} \to \dots \to a_0$. Let j be the largest index with $a_i \to a_{i+1} \to \dots \to a_j$, so that either $j = k$ or $a_{j+1} \to a_j$. By the confluency, there exist a vertex d, a directed path from a_0 to d and a directed path $c_0, \dots, c_n$ from a_j to d. Since a_0 is terminal, we have $d = a_0$. But then it is easy to see that $c_n, \dots, c_0, a_{j+1}, \dots, a_k$ is an undirected path from a_0 to a_k which has a smaller number of peaks than $a_0, \dots, a_k$, a contradiction. $\square$

A directed graph $\langle G, \to \rangle$ is called *locally confluent* if for any triple of vertices a, b, c with $a \to b$ and $a \to c$, there exists a fourth vertex d, a directed path from b to d, and a directed path from c to d.

2.2. THEOREM. *Let $\langle G, \to \rangle$ be a finitely terminating and locally confluent directed graph. Then $\langle G, \to \rangle$ is confluent.*

PROOF. Let a, b, c be three vertices such that there are a directed path from a to b and a directed path from a to c, and suppose that there is no vertex d with a directed path from b to d and a directed path from c to d. Since the graph is finitely terminating, there exist a directed path from b to a terminal vertex b' and a directed path from c to a terminal vertex c'; we have $b' \neq c'$. Denote by P the set of the vertices e for which there exist two different terminal vertices f and g, a directed path from e to f and a directed path from e to g. We have seen that the set P is nonempty. Since the graph is finitely terminating, there exists a vertex $e \in P$ such that whenever $e \to e'$, then $e' \notin P$. For such a vertex e, let $e = f_0, \dots, f_k$ and $e = g_0, \dots, g_l$ be two directed paths with $f_k \neq g_l$ and both f_k and g_l terminal. By the local confluency, there exist a vertex h, a directed path from f_1 to h and a directed path from g_1 to h. Since the graph is finitely terminating, there is a directed path from h to a terminal vertex h'. Now $f_1 \notin P$ and there are two directed paths from f_1 to two terminal vertices, from which we get $f_k = h'$. Quite similarly, $g_1 \notin P$ implies $g_l = h'$. Hence $f_k = g_l$, a contradiction. $\square$

3. Term rewrite systems

Let B be a set of equations. In order to stress their asymmetric character, the equations from B are called *rewrite rules*, and the set B is called a *term rewrite system*.

Let us define a directed graph $\langle G, \to \rangle$ in the following way: G is the set of all terms; $u \to v$ if and only if the equation $\langle u, v \rangle$ is an immediate consequence of an equation from B. This directed graph will be called associated with B. If $u \to v$, we also say that u can be rewritten (with respect to B) in one step to v. If there exists a directed path from u to v, we say that u can be rewritten to v (with respect to B). Of course, two terms u, v are connected if and only if there exists a derivation of $\langle u, v\rangle$ from B, i.e., if $\langle u, v\rangle \in \mathbf{Eq}(B)$. A term u is called terminal (with respect to B) if it is a terminal vertex in the associated graph.

The set of equations B is called finitely terminating, or (locally) confluent if the associated graph is finitely terminating, or (locally) confluent, respectively. Finitely terminating confluent term rewrite systems are called convergent.

By a *critical pair* for B we mean a pair of terms that can be obtained in the following way. Let $\langle u_1, v_1\rangle \in B$, $\langle u_2, v_2\rangle \in B$, let a be an occurrence of a subterm $w \notin X$ in u_1 having a common substitution instance with u_2, and let f, g be a minimal unifying pair for w, u_2. Then a is (clearly) also an occurrence of $f(w) = g(u_2)$ in $f(u_1)$, and $\langle f(v_1), f(u_1)[a : f(w) \to g(v_2)]\rangle$ is a critical pair for B.

Clearly, if $\langle u, v\rangle$ is a critical pair for B, then $\langle u, v\rangle \in \mathbf{Eq}(B)$. In the above notation, we have $f(u_1) \to f(v_1)$ and $f(u_1) \to f(u_1)[a : f(w) \to g(v_2)]$.

The set B is said to have *confluent critical pairs* if for each pair $\langle u, v\rangle$ critical for B, there exists a term t such that both u and v can be rewritten to t with respect to B.

3.1. Theorem. *A set of equations is locally confluent if and only if it has confluent critical pairs.*

Proof. The direct implication is clear. Let B have confluent critical pairs and let p, q, r be three terms with $p \to q$ and $p \to r$, so that $\langle p, q\rangle$ is an immediate consequence of an equation $\langle u_1, v_1\rangle \in B$ and $\langle p, r\rangle$ is an immediate consequence of an equation $\langle u_2, v_2\rangle \in B$. There exists an address a_1 in p such that $p[a_1] = f(u_1)$ for a substitution f and $p[a_1 : f(u_1) \to f(v_1)] = q$. Similarly, there exists an address a_2 in p such that $p[a_2] = g(u_2)$ for a substitution g and $p[a_2 : g(u_2) \to g(v_2)] = r$. If the two addresses a_1 and a_2 are incomparable, then we can put $s = q[a_2 : g(u_2) \to g(v_2)] = r[a_1 : f(u_1) \to f(v_1)]$ and we have $q \to s$ and $r \to s$. Now, without loss of generality, we can assume that a_2 is an extension of a_1, so that $g(u_2)$ is a subterm of $f(u_1)$.

Consider first the case when there are two addresses a, b such that $a_2 = a_1ab$ and $u_1[a] \in X$. Put $x = u_1[a]$. We can define a substitution h by $h(x) = f(x)[b : g(u_2) \to g(v_2)]$ and $h(y) = y$ for all $y \in X \setminus \{x\}$. Put $s = p[a_1 : f(u_1) \to h(v_1)]$. Clearly, q can be rewritten to s (the number of steps is $|v_1|_x$), and also r can be rewritten to s (the number of steps is $|u_1|_x$).

It remains to consider the case when $a_2 = a_1a$ for an address a and a is an occurrence of a subterm $w \notin X$ in u_1. We have $f(w) = p[a_2] = g(u_2)$. Let f_0, g_0 be a minimal unifying pair for f, g, so that $f = hf_0$ and $g = hg_0$ for a substitution h. Then $\langle f_0(v_1), f_0(u_1)[a : f_0(w) \to g_0(v_2)]\rangle$ is a critical pair and there exists a term s such that both members of the pair can be rewritten to s. Clearly, both $f(v_1)$ and $f(u_1)[a : g(u_2) \to g(v_2)]$ can be rewritten to $h(s)$. It follows easily that both q and r can be rewritten to the same term. $\square$

4. Well quasiorders

Let $\leq$ be a quasiorder on a set A, i.e., a reflexive and transitive relation. For two elements $a, b \in A$ we write $a < b$ if $a \leq b$ and $b \not\leq a$. We write $a \equiv b$ if $a \leq b$ and $b \leq a$. (If the quasiorder is denoted by $\leq'$, or $\leq_i$, etc., then the derived relations will be denoted by $<'$, $\equiv'$, $<_i$, $\equiv_i$, etc., respectedly.)

4.1. LEMMA. *The following conditions are equivalent for a quasiorder $\leq$ on a set A:*

(1) *for any infinite sequence $a_0, a_1, \dots$ of elements of A there exist two indexes i, j with $i < j$ and $a_i \leq a_j$;*
(2) *there are no infinite antichains (i.e., infinite subsets S of A such that $a \not\leq b$ whenever a, b are two distinct elements of S) and no infinite descending chains (i.e., infinite sequences $a_0, a_1, \dots$ of elements of A such that $a_j < a_i$ whenever $i < j$);*
(3) *every infinite sequence $a_0, a_1, \dots$ of elements of A has an infinite nondecreasing subsequence (i.e., there exist indexes $i_0 < i_1 < \dots$ with $a_{i_0} \leq a_{i_1} \leq \dots$).*

PROOF. (1) $\rightarrow$ (2) and (3) $\rightarrow$ (1) are clear. It remains to prove (2) $\rightarrow$ (3). This is clear if there exists an index k such that $a_k \equiv a_i$ for infinitely many indexes i. So, we can assume that for any k there are only finitely many such indexes i.

Since there are no infinite descending chains, every nonempty subset of A contains a minimal element, i.e., an element a such that there is no element b in the subset with $b < a$.

Let I_{-1} be the set of nonnegative integers. By induction on $j \geq 0$ we are going to define an index i_j and an infinite set of indexes I_j with $i_j < m$ and $a_{i_j} < a_m$ for all $m \in I_j$. Suppose that $i_0, \dots, i_{j-1}$ and I_{j-1} have been already defined. The set $Q = \{a_m : m \in I_{j-1}\}$ contains at least one minimal element; according to the above assumption, and since there are no infinite antichains, there are only finitely many elements $m \in I_{j-1}$ such that a_m is a minimal element of Q. Now for every $m \in I_{j-1}$ there exists an $m_0 \in I_{j-1}$ such that $a_{m_0} \leq a_m$ and a_{m_0} is a minimal element of Q. Since I_{j-1} is infinite, it follows that there is an $m_0 \in I_{j-1}$ such that a_{m_0} is a minimal element of Q and there are infinitely many elements $m \in I_{j-1}$ with $a_{m_0} < a_m$. Take one such m_0 and denote it by i_j; put $I_j = \{m \in I_{j-1} : a_{i_j} < a_m \text{ and } i_j < m\}$.

Now it is easy to see that $i_0 < i_1 < \dots$ and $a_{i_0} < a_{i_1} < \dots$. □

A quasiorder $\leq$ on A is called a *well quasiorder* if it satisfies one of the three equivalent conditions of Lemma 4.1.

The following observation is easy to prove: Any quasiorder on a set A which extends a well quasiorder on A, is itself a well quasiorder.

By a *bad sequence* for a quasiorder $\leq$ on A we mean an infinite sequence $a_0, a_1, \dots$ of elements of A such that $a_i \not\leq a_j$ whenever $i < j$. So, $\leq$ is a well quasiorder if and only if there is no bad sequence for $\leq$.

4.2. Lemma. *Let $\leq$ be a quasiorder on A, with respect to which there are no infinite descending chains. If $\leq$ is not a well quasiorder, then there exists a bad sequence $a_0, a_1, \dots$ for $\leq$ such that the set $\{a \in A : a < a_i$ for some $i\}$ is well quasiordered by $\leq$.*

Proof. Since there are no infinite descending chains, every nonempty subset of A contains a minimal element, i.e., an element a such that there is no element b in the subset with $b < a$. Let us define elements $a_i \in A$ by induction on $i = 0, 1, \dots$ as follows. Suppose that $a_0, \dots, a_{i-1}$ are already defined. Denote by X_i the set of the elements a for which there exists a bad sequence $b_0, b_1, \dots$ with $b_0 = a_0, \dots, b_{i-1} = a_{i-1}, b_i = a$; if X_i is nonempty, let a_i be a minimal element of X_i. By induction, it is easy to see that all the sets X_i are nonempty, so that a_i is defined for any i. Clearly, the sequence $a_0, a_1, \dots$ is bad.

Let $B = \{a \in A : a < a_i \text{ for some } i\}$. Suppose that there is a bad sequence $c_0, c_1, \dots$ of elements of B. Denote by n the least number such that $c_i < a_n$ for some i. Denote by m the least number such that $c_m < a_n$. It is easy to see that the sequence $a_0, \dots, a_{n-1}, c_m, c_{m+1}, \dots$ is a bad sequence contradicting the minimality of the element b_n. □

4.3. Lemma. *Let $\leq_1$ be a well quasiorder on A_1 and $\leq_2$ be a well quasiorder on A_2. Then the relation $\leq$ on $A_1 \times A_2$, defined by $\langle a, b\rangle \leq \langle c, d\rangle$ if and only if $a \leq_1 c$ and $b \leq_2 d$, is also a well quasiorder.*

Proof. It follows easily from condition (3) of Lemma 4.1. □

Let $\leq$ be a quasiorder on A. Denote by B the set of all finite sequences of elements of A, and define a binary relation $\leq^*$ on B by $\langle a_1, \dots, a_n\rangle \leq^* \langle b_1, \dots, b_m\rangle$ if and only if $n \leq m$ and there are indexes $1 \leq i_1 < i_2 < \dots < i_n \leq m$ with $a_1 \leq b_{i_1}, \dots, a_n \leq b_{i_n}$. Clearly, $\leq^*$ is a quasiorder on B; it is called the *sequence quasiorder* (with respect to $\leq$).

4.4. Lemma. *Let $\leq$ be a well quasiorder on a set A. Then the sequence quasiorder $\leq^*$ on the set B of finite sequences of elements of A is also a well quasiorder.*

Proof. It is easy to see that there are no infinite strictly decreasing sequences with respect to $\leq^*$. Suppose that $\leq^*$ is not a well quasiorder, so that, by Lemma 4.2, there is a bad sequence $\alpha_0, \alpha_1, \dots$ of elements of B for which the set $C = \{\alpha \in B : \alpha <^* \alpha_i \text{ for some } i\}$ is well quasiordered by $\leq^*$. Clearly, we can suppose that each α_i is a nonempty sequence; write $\alpha_i = a_i\beta_i$. Then $\beta_i <^* \alpha_i$ and the set $\{\beta_i : i \geq 0\}$ is well quasiordered by $\leq^*$. By Lemma 4.3, the set $A \times \{\beta_i : i \geq 0\}$ is well quasiordered by the quasiorder obtained from $\leq$ and $\leq^*$, and the sequence $\langle a_0, \beta_0\rangle, \langle a_1, \beta_1\rangle, \dots$ of its elements cannot be bad. Consequently, $a_i \leq a_j$ and $\beta_i \leq^* \beta_j$ for some $i < j$. But then clearly $\alpha_i \leq^* \alpha_j$, a contradiction. □

If $\leq$ is a quasiorder on a set A, then for two finite sequences $\langle a_1, \dots, a_n \rangle$ and $\langle b_1, \dots, b_m \rangle$ of elements of A we write $\langle a_1, \dots, a_n \rangle \leq \langle b_1, \dots, b_m \rangle$ (lexicographically) if one of the following two cases takes place: either $n \leq m$ and $a_i \equiv b_i$ for all $i \leq n$, or else there is an index $k \leq \min(n, m)$ such that $a_k < b_k$ and $a_i \equiv b_i$ for all $i < k$. Clearly, this is a quasiorder on the set of finite sequences of elements of A. However, it is not a well quasiorder whenever the set A contains two elements incomparable with respect to $\leq$.

5. Well quasiorders on the set of terms

For a term t, we define an element $o_1(t) \in X \cup \sigma$ and a finite sequence of terms $o_2(t)$ in this way: if $t \in X$, then $o_1(t) = t$ and $o_2(t)$ is empty; if $t = Ft_1 \dots t_n$, then $o_1(t) = F$ and $o_2(t) = \langle t_1, \dots, t_n \rangle$.

Let Y be a subset of X and $\leq_\circ$ be a quasiorder on the set $Y \cup \sigma$. Then we can define a quasiorder $\leq$ on the set of terms over Y inductively as follows: $u \leq v$ if and only if either

(1) $o_1(u) \leq_\circ o_1(v)$ and $o_2(u) \leq^* o_2(v)$, or
(2) $v = Fv_1 \dots v_n$, and $u \leq v_i$ for some i.

It is easy to see that $\leq$ is a quasiorder; it is called the *term quasiorderterm quasiorder* over $\leq_\circ$.

5.1. Lemma. *Let $\leq_\circ$ be a well quasiorder on the set $Y \cup \sigma$, where $Y \subseteq X$. Then the term quasiorder $\leq$ over $\leq_\circ$ is a well quasiorder on the set of terms over Y.*

Proof. Clearly, $u \leq v$ implies $\lambda(u) \leq \lambda(v)$. Easily, there are no infinite descending chains with respect to $\leq$. Suppose that $\leq$ is not a well quasiorder. By Lemma 4.2 there is a bad sequence $t_0, t_1, \dots$ of terms such that the set $B = \{t : t < t_i \text{ for some } i\}$ is well quasiordered by $\leq$. Observe that for each t_i, the sequence $o_2(t_i)$ is a finite sequence of elements of B. By Lemma 4.3, the sequence $\langle o_1(t_0), o_2(t_0) \rangle, \langle o_1(t_1), o_2(t_1) \rangle, \dots$ cannot be bad with respect to the product quasiorder of $\leq_\circ$ and $\leq^*$. Consequently, for some $i < j$ we have $o_1(t_i) \leq_\circ o_1(t_j)$ and $o_2(t_i) \leq^* o_2(t_j)$. But then $t_i \leq t_j$, a contradiction. □

5.2. Corollary. *Let $\leq$ be a quasiorder on the set of terms over a set $Y \subseteq X$. If there exists a well quasiorder $\leq_\circ$ on $Y \cup \sigma$ such that $\leq$ extends the term quasiorder over $\leq_\circ$, then $\leq$ is a well quasiorder.*

A term rewrite system B is said to be *compatible* with a quasiorder $\leq$ on the set of terms over a set $Y \subseteq X$ if $\mathbf{S}(u) \cup \mathbf{S}(v) \subseteq Y$ for any $\langle u, v \rangle \in B$, and $f(v) < f(u)$ for any $\langle u, v \rangle \in B$ and any substitution f over Y. (By a substitution over Y we mean one that maps the set of terms over Y into itself.)

By a *simplification quasiorder* on the set of terms over a set $Y \subseteq X$ we mean a quasiorder $\leq$ satisfying the following two conditions:

(1) $Fu_1 \dots u_n \leq Fv_1 \dots v_n$ whenever $u_i \leq v_i$ for all i;
(2) $u \leq v$ whenever u is a subterm of v.

Of course, condition (1) is equivalent to

$$u \le v \text{ implies } Fu_1 \dots u_{i-1}u\, u_{i+1} \dots u_n \le Fu_1 \dots u_{i-1}v\, u_{i+1} \dots u_n.$$

A quasiorder satisfying (1) is said to be *monotone*; in literature, such a quasiorder is also often called a *quasiorder respecting replacement.* Condition (2) is called the *subterm property*; it is equivalent to

$$u_i \le Fu_1 \dots u_n \text{ for all } i.$$

5.3. THEOREM. *Let σ be finite. Let a term rewrite system B be compatible with a simplification quasiorder $\le$ on the set of terms over a finite subset Y of X. Then B is finitely terminating.*

PROOF. The set $\sigma \cup Y$ is finite, so the identity on this set is a well quasiorder. Since $\le$ is a simplification quasiorder, it is easy to see that it is an extension of the term quasiorder over the identity on $\sigma \cup Y$. By Lemma 5.1, $\le$ is a well quasiorder.

Suppose that B is not finitely terminating, so that there exists an infinite directed path $u_0 \to u_1 \to u_2 \to \dots$ in the associated graph. Since for any substitution f also $f(u_0) \to f(u_1) \to f(u_2) \to \dots$ is an infinite path, we may assume that all the terms u_i are terms over Y. Denote by P the set of terms that are initial terms of an infinite directed path, all the members of which are terms over Y. So, P is nonempty.

Let us prove by induction on the length of t that if $t \in P$, then $t > t'$ for some $t' \in P$. If $t \in Y$, it follows easily from the definition of compatibility. So, let $t = Ft_1 \dots t_n$. Let $t = u_0 \to u_1 \to u_2 \to \dots$ be an infinite directed path, the existence of which follows from $t \in P$. For every i there are an equation $\langle p_i, q_i \rangle \in B$ and a substitution f_i such that u_{i+1} results from u_i by replacing an occurrence of a subterm $f_i(p_i)$ with $f_i(q_i)$. Now if $f_i(p_i)$ is a proper subterm of u_i for all i, then one can easily see that $t_j \in P$ for at least one number $j \in \{1, \dots, n\}$, by induction $t_j > t'_j$ for some $t'_j \in P$, and $t \ge t_j > t'_j$. So, we can assume that $f_i(p_i) = u_i$ for at least one index i. Then also $f_i(q_i) = u_{i+1}$. By the definition of compatibility, $f_i(q_i) < f_i(p_i)$, i.e., $u_{i+1} < u_i$. Since $\le$ is a monotone quasiorder, we have $t = u_0 \ge u_1 \ge \dots \ge u_i > u_{i+1} \in P$, so we can take $t' = u_{i+1}$.

From this it follows that there is an infinite decreasing sequence of terms with respect to $\le$, a contradiction, since $\le$ is a well quasiorder. □

5.4. THEOREM. *Let σ be finite and Y be a finite subset of X. Let $\le$ be a recursive quasiorder on the set of terms over Y such that $u < v$ implies $f(u) < f(v)$ for all substitutions f over Y. Then there is an algorithm for testing finite term rewrite systems for compatibility with $\le$.*

PROOF. It is clear. □

6. The Knuth-Bendix algorithm

Let σ be finite, Y be a finite subset of X and $\le$ be a recursive simplification quasiorder on the set of terms over Y such that $u < v$ implies $f(u) < f(v)$

for all substitutions f over Y. The following algorithm, called the *Knuth-Bendix algorithm* with respect to $\leq$, can be used to modify a finite term rewrite system B over Y. For some inputs, the algorithm never terminates; for other ones it halts with failure; if halting with success, the algorithm outputs a convergent term rewrite system for the equational theory based on the original term rewrite system B. Here is the algorithm:

Step 1: Modify B by replacing any equation $\langle u,v\rangle$ such that $u < v$, with the equation $\langle v,u\rangle$; if for some equation $\langle u,v\rangle \in B$ neither $u < v$ nor $v > u$ takes place, halt with failure.

Step 2: Denote by C the set of the critical pairs $\langle u,v\rangle$ for B that are not confluent (i.e., the terms u and v cannot be rewritten to a common term with respect to B); in each class of similar such critical pairs take one representant, so that C is finite. If C is empty, halt with success. Otherwise, replace B with $B \cup C$ and go to Step 1.

In order to be able to apply the Knuth-Bendix algorithm together with theorems 5.3 and 5.4, it is necessary to have a class of simplification quasiorders $\leq$ at hand, such that $u < v$ implies $f(u) < f(v)$ for all substitutions f over a given finite set of variables.

7. The Knuth-Bendix quasiorder

Let σ be finite and let Y be a finite subset of X; let $\leq_\circ$ be a quasiorder on $Y \cup \sigma$; let $\leq'$ be any quasiorder on the set of terms over Y. Let us define inductively a binary relation $\leq$ on the set of terms over Y by $u \leq v$ if and only if one of the following (mutually exclusive) cases takes place:

(1) $u <' v$;
(2) $u \equiv' v$ and $o_1(u) <_\circ o_1(v)$;
(3) $u \equiv' v$, $o_1(u) \equiv_\circ o_1(v)$ and $o_2(u) \leq o_2(v)$ (lexicographically).

It is not difficult to prove that $\leq$ is a quasiorder contained in $\leq'$; it is called the *Knuth-Bendix quasiorder* obtained from $\leq_\circ$ and $\leq'$.

A quasiorder $\leq$ on a set of terms is said to be *stable for variables* if $\mathbf{S}(u) \subseteq \mathbf{S}(v)$, whenever $u \leq v$.

7.1. Lemma. *Let $\leq$ be the Knuth-Bendix quasiorder obtained from $\leq_\circ$ and from a simplification quasiorder $\leq'$ such that $Ft_1 \dots t_n \equiv' t_i$ can hold only when F is unary and $F \geq_\circ s$ for all $s \in Y \cup \sigma$. Then:*

(1) *$\leq$ is a simplification quasiorder, and $u < v$ whenever u is a proper subterm of v;*
(2) *if either $\leq_0$ or $\leq'$ is an order, then $\leq$ is an order;*
(3) *if $\leq'$ is stable for variables and both $\leq'$ and $<'$ are fully invariant, then $\leq$ is stable for variables and both $\leq$ and $<$ are fully invariant.*

Proof. (1) The monotonicity of $\leq$ is clear. Let us prove by induction on $Ft_1 \dots t_n$ that $Ft_1 \dots t_n \equiv t_i$ can never happen. Suppose $Ft_1 \dots t_n \equiv t_i$. Then F is unary, and maximal with respect to $\leq_\circ$. Clearly, $F \equiv_\circ o_1(t_1)$ and $t_1 \equiv o_2(t_1)$ (lexicographically). This is possible only if $o_2(t_1)$ is of length 1,

$t_1 = Gu_1$ for some unary symbol G and some u_1, and $t_1 \equiv u_1$; we get a contradiction by induction.

Let us prove $t_i \le Ft_1 \dots t_n$ by induction on $Ft_1 \dots t_n$. If $t_i <' Ft_1 \dots t_n$, we are done. Otherwise, F is unary and a largest element of $Y \cup \sigma$, $i = 1$, and $t_1 \equiv' Ft_1$. If $o_1(t_1) <_\circ F$, we are done. Otherwise, $o_1(t_1) \equiv_\circ F$. By induction, every member of $o_2(t_1)$ is $\le t_1$; as we have seen above, the first member cannot be $\equiv t_1$. Hence $o_2(t_1) \le t_1$ (lexicographically), and we get $t_1 \le Ft_1$.

(2) is easy. It remains to prove (3). Since $\le$ is contained in $\le'$, the quasiorder $\le$ is stable for variables. We are going to prove by induction on the complexity of u, v that if $u \le v$ then $f(u) \le f(v)$, and if $u < v$ then $f(u) < f(v)$, for any substitution f over Y. If $u <' v$, it is clear. So, let $u \equiv' v$. If both $o_1(u)$ and $o_1(v)$ belong to σ, the proof is easy (by induction in the case (3) for $u \le v$).

Let $u \in Y$. Since $\le$ is stable for variables, u occurs in v and $f(u)$ is a subterm of $f(v)$, so that $f(u) \le f(v)$ by the subterm property; if $u < v$, then u is a proper subterm of v and $f(u) < f(v)$ by (1).

Now suppose that $u \notin Y$ and $v \in Y$. Then v is the only variable that can occur in u. If v does occur in u, then we can write $u = Fu_1 \dots u_n$ and v occurs in some u_i, so that $v \le u_i < u \le v$, a contradiction. Hence u contains no variables. Now $u \equiv' v$ and the full invariancy of $\le'$ imply that all terms over Y are equivalent with respect to $\equiv'$, a contradiction, since $\le'$ is stable for variables. □

Let Y be a finite subset of X. By a weighting function (over Y) we shall mean a mapping of $Y \cup \sigma$ into the set of nonnegative real numbers. Every weighting function α over Y can be extended to the set of all terms over Y, if we define $\alpha(t)$ to be the sum of all $\alpha(s)$, where s runs over all occurrences of variables and operation symbols in t. (Or inductively: $\alpha(Ft_1 \dots t_n) = \alpha(F) + \alpha(t_1) + \dots + \alpha(t_n)$.)

Let σ be finite and Y be a finite subset of X; let $\le_\circ$ be a quasiorder on $Y \cup \sigma$, and α be a weighting function. Let us define a quasiorder $\le'$ on the set of terms over Y by $u \le' v$ if and only if either $\alpha(u) < \alpha(v)$ and $|u|_x \le |v|_x$ for all $x \in Y$ (where $|t|_x$ is the number of occurrences of x in t), or else $\alpha(u) = \alpha(v)$ and $|u|_x = |v|_x$ for all $x \in Y$. (It is easy to see that $\le'$ is a quasiorder.) The Knuth-Bendix quasiorder obtained from $\le_\circ$ and $\le'$ will be called the Knuth-Bendix quasiorder obtained from $\le_\circ$ and α.

7.2. THEOREM. (Knuth, Bendix [70]) *Let σ be finite and Y be a finite subset of Y; let $\le_\circ$ be a quasiorder on $Y \cup \sigma$ and α be a weighting function over Y such that the following three conditions are satisfied:*

(1) $\alpha(s) > 0$ *whenever s is a constant or a variable;*
(2) *if $\alpha(F) = 0$ for a unary operation symbol F, then $F \ge_\circ s$ for all $s \in Y \cup \sigma$;*
(3) *if $x \in Y$, then $\alpha(x) = \alpha(y)$ for all $y \in Y$ and $\alpha(x) \le \alpha(c)$ for every constant $c \in \sigma$.*

Then the Knuth-Bendix quasiorder $\le$ obtained from $\le_\circ$ and α is a fully invariant simplification quasiorder; $<$ is also fully invariant, and $\le$ is stable for variables.

Proof. In order to be able to apply Lemma 7.1, we must show that $\le'$ (as defined above) is a simplification quasiorder, $Ft_1 \dots t_n \equiv' t_i$ can hold only when $n = 1$ and $F \ge_\circ s$ for all $s \in Y \cup \sigma$, $\le'$ is stable for variables and both $\le'$ and $<'$ are fully invariant.

The monotonicity of $\le'$ is clear. Let us prove $t_i \le' Ft_1 \dots t_n$. We have $\alpha(t_i) \le \alpha(Ft_1 \dots t_n)$ and everything is clear if this inequality is sharp. So, let $\alpha(t_i) = \alpha(Ft_1 \dots t_n)$. By (1) we get $n = 1$; but then, $|t_1|_x = |Ft_1|_x$ for all x.

Let $Ft_1 \dots t_n \equiv' t_i$. Clearly, $\alpha(F) = 0$ and, by (1), $n = 1$. By (2), $F \ge_\circ s$ for all $s \in Y \cup \sigma$.

Since $u \le' v$ implies $|u|_x \le |v|_x$ for all $x \in Y$, $\le'$ is stable for variables.

It is easy to see that for any term t, $\alpha(f(t)) = \alpha(t) + \sum\{\alpha(f(t[e])) - \alpha(t[e]) : e \in \mathbf{O}_X(t)\}$. It follows from (3) that $\alpha(f(x)) - \alpha(x) \ge 0$ for all $x \in Y$. From this it follows easily that both $\le'$ and $<'$ are fully invariant. □

7.3. Example. Let $\sigma = \{\cdot, {}^{-1}, e\}$ where $\cdot$ is binary, ${}^{-1}$ is unary, and e is a constant; let $Y = \{x, y, z\}$. Define $\le_\circ$ by $x \equiv_\circ y \equiv_\circ z \equiv_\circ e <_\circ \cdot <_\circ {}^{-1}$, and α by $\alpha(\cdot) = \alpha({}^{-1}) = 0$ and $\alpha(x) = \alpha(y) = \alpha(z) = \alpha(e) = 1$. Using the above results, one can see that the set consisting of the ten equations

$$\begin{array}{lllll} ex \approx x, & x^{-1}x \approx e, & (xy)z \approx x(yz), & x^{-1}(xy) \approx y, & xe \approx x, \\ e^{-1} \approx e, & (x^{-1})^{-1} \approx x, & xx^{-1} \approx e, & x(x^{-1}y) \approx y, & (xy)^{-1} \approx y^{-1}x^{-1} \end{array}$$

is a convergent base for group theory.

8. Perfect bases

Recall that for a pair u, v of terms, $u \le v$ means that a substitution instance of u is a subterm of v. If $u \not\le v$, we say that v avoids u.

Let P be a set of equations. We denote by A_P the set of all terms t such that whenever $u \approx u' \in P$, then $u \not\le t$. So A_P consists of all those terms that avoid the left sides of equations in P.

P is said to be *pre-perfect* if the following conditions are satisfied:

(pp1) if $u \approx u' \in P$, $v \approx v' \in P$ and $\langle u, u' \rangle \ne \langle v, v' \rangle$, then $u \not\le v$;

(pp2) if $u \approx u' \in P$, $v \approx v' \in P$ and $f(u) = g(v)$ for two substitutions f, g such that every proper subterm of $f(u)$ belongs to A_P, then $f(u') = g(v')$;

(pp3) if $u \approx u' \in P$ and f is a substitution such that every proper subterm of $f(u)$ belongs to A_P, then $f(u') \in A_P$.

If P is a pre-perfect set of equations, then we can define a mapping ν_P of the set of terms into A_P as follows:

$\nu_P(x) = x$ for any variable x;

$\nu_P(F(t_1, \dots, t_n)) = F(\nu_P(t_1), \dots, \nu_P(t_n))$ if the last term belongs to A_P,

$\nu_P(F(t_1, \dots, t_n)) = f(u')$ if $F(\nu_P(t_1), \dots, \nu_P(t_n)) = f(u)$ for a substitution f and an equation $u \approx u' \in P$.

In fact, the three conditions above are just a formulation of the correctness of this definition plus a little bit more.

If P is a pre-perfect set of equations, we can consider the set A_P to be an algebra of the given signature by setting $F_{A_P}(t_1, \dots, t_n) = \nu_P(F(t_1, \dots, t_n))$, where n is the arity of F.

8.1. Lemma. *Let P be a pre-perfect set of equations. Then:*

(1) *If $u \approx u' \in P$, then u is not a variable, $u' = \nu_P(u)$ and $\mathbf{S}(u') \subseteq \mathbf{S}(u)$;*
(2) *ν_P is a homomorphism of the term algebra onto A_P.*

Proof. (1) Let $u \approx u' \in P$. By (pp1), every proper subterm of u belongs to A_P, so that condition (pp3) with respect to the identical substitution says that $u' \in A_P$. In particular, A_P is nonempty; but then u cannot be a variable. It is easy to see that $u' = \nu_P(u)$. Suppose that there is a variable $y \in \mathbf{S}(u') - \mathbf{S}(u)$. Let f be the identical substitution, and g be the substitution with $g(x) = x$ for any variable $x \neq y$, and $g(y) = x$. We have $f(u) = g(u) = u$ but $f(u') \neq g(u')$, a contradiction with (pp2).

(2) We need to prove $\nu_P F(t_1, \dots, t_n) = F_A(\nu_P(t_1), \dots, \nu_P(t_n))$, i.e., we need to prove $\nu_P F(t_1, \dots, t_n) = \nu_P F(\nu_P(t_1), \dots, \nu_P(t_n))$. If $F(\nu_P(t_1), \dots, \nu_P(t_n)) \in A_P$, then both sides are equal to this term. If, on the contrary, this term is of the form $f(u)$ for a substitution f and an equation $u \approx u' \in P$, then both sides are equal to $f(u')$ according to the definition of ν_P. $\square$

By a *perfect base* we mean a pre-perfect set P of equations such that the algebra A_P satisfies all the equations from P. A subset P of an equational theory E is a perfect base for E if and only if it is pre-perfect and the algebra A_P satisfies all the equations from E.

8.2. Theorem. *Let P be a perfect base for E. Then:*

(1) *ν_P is a normal form function for E;*
(2) *A_P is the free E-algebra over the set of variables;*
(3) *E is decidable if P is recursive, with recursive domain.*

Proof. (nf2) is easy by induction on the complexity of t, and (nf3) is clear. By 8.1, ν_P is a homomorphism of the term algebra onto A_P. So, if $u \approx v \in E$, then $\nu_P(u) = \nu_P(v)$, because A_P satisfies all the equations from E. The converse follows from (nf2), so we have both implications of (nf1). It follows that A_P is isomorphic to the factor of the term algebra through E, and hence A_P is the free E-algebra over the set of variables. (3) follows from (1). $\square$

8.3. Example. Examples will be given for the signature of groupoids. The set P consisting of the two equations

$$xx \cdot yy \approx xx, \qquad (xx \cdot x) \cdot yy \approx ((xx \cdot x)x)x$$

will serve as an example of a finite perfect base, the equational theory of which has no (either finite or infinite) finitely terminating and confluent base. Denote by E the equational theory based on P and by $\circ$ the multiplication of the groupoid A_P. Based on the following observation, one can easily check that P is perfect: if $a \in A_P$, then

$$a \circ a = \begin{cases} a \text{ if } a \text{ is a square (i.e., } a = tt \text{ for a term } t), \\ aa \text{ if } a \text{ is not a square.} \end{cases}$$

In each case, $a \circ a$ is a square.

Suppose that there is a finitely terminating and confluent base Q for E. It is easy to see that if t is a term with $t \approx xx \in E$, then t contains xx as a subterm and so, because of the finite termination, xx cannot be Q-rewritten to t. It follows that xx is in Q-canonical form and the term $xx \cdot yy$ can be Q-rewritten in finitely many steps to xx. Denote by w the Q-canonical form of $(xx \cdot x) \cdot yy$. We have $w \neq (xx \cdot x) \cdot yy$, since $((xx \cdot x)x)x$ cannot be Q-rewritten to $(xx \cdot x) \cdot yy$, due to finite termination. So we have that $(xx \cdot x) \cdot yy$, as well as $xx \cdot yy$, can be Q-rewritten. This implies that w avoids both $(xx \cdot x) \cdot yy$ and $xx \cdot yy$, because w cannot be Q-rewritten, being itself in Q-canonincal form. But then $w \in A_P$ and hence $w = ((xx \cdot x)x)x$, since P is perfect. This means that $(xx \cdot x) \cdot yy$ can be Q-rewritten in finitely many steps to $((xx \cdot x)x)x$. Consequently, the term $((xx \cdot xx) \cdot xx) \cdot xx$ can be Q-rewritten in finitely many steps to $(((xx \cdot xx) \cdot xx) \cdot xx) \cdot xx$, clearly a contradiction.

So there are equational theories with finite perfect bases but without any convergent term rewriting system. On the other hand, the equational theory of semigroups serves as an example of an equational theory with a convergent term rewriting system but with no perfect base.

8.4. Lemma. *The set of the finite pre-perfect sets of equations is recursive.*

Proof. Condition (pp1) is easy to verify, and we can also easily verify that $u \approx u' \in P$ implies $\mathbf{S}(u') \subseteq \mathbf{S}(u)$, which is necessary according to 8.1. Under this assumption, conditions (pp2) and (pp3) can be equivalently reformulated in the following way:

(pp2′) if $u \approx u' \in P$, $v \approx v' \in P$ and f and g is the minimal unifying pair for u and v, then, in case that every proper subterm of $f(u)$ belongs to A_P, $f(u') = g(v')$;

(pp3′) if $u \approx u' \in P$, $v \approx v' \in P$, $s \subseteq u'$ and f and g is the minimal unifying pair for s and v, then $f(u)$ contains a proper subterm not in A_P.

The equivalence of (pp2) with (pp2′) is easy, and clearly (pp3) implies (pp3′). It remains to prove that (pp3′) implies (pp3). Let $u \approx u' \in P$ and let f be a substitution such that every proper subterm of $f(u)$ belongs to A_P. Suppose $f(u') \notin A_P$, i.e., $g(v) \subseteq f(u')$ for a substitution g and an equation $v \approx v' \in P$. If $x \in \mathbf{S}(u')$, then $x \in \mathbf{S}(u') \subseteq \mathbf{S}(u)$, $f(x)$ is a proper subterm of $f(u)$ and so, by our assumption, $g(v)$ cannot be a subterm of $f(x)$. The only other possibility for $g(v)$ to be a subterm of $f(u')$ is, that $g(v) = f(s)$ for a subterm s of u'. Let f_0, g_0 be the minimal unifying pair for s and v, so that $f = hf_0$ and $g = hg_0$ for some h. By (pp3′), $f_0(u)$ contains a proper subterm not in A_P. But then clearly $f(u) = hf_0(u)$ also contains a proper subterm not in A_P, a contradiction. $\square$

Let P be a finite pre-perfect set and let $a \approx b$ be an equation from P. We want to decide if the algebra A_P satisfies $a \approx b$. For this purpose we shall construct a finite set of substitutions which will serve as a test set. By a *permissible substitution* we shall mean one which maps variables into A_P. All our testing substitutions will be permissible. Observe that if gh is a permissible substitution, then h is also permissible, because the complement of A_P is closed under any substitution.

By induction on the complexity of a term t, we are first going to define a finite set $U(t)$ of permissible substitutions with the following property: if $f = gh$ where f is a permissible substitution, and f and h expand precisely the same substitutions from $U(t)$, then $\nu_P f(t) = g\nu_P h(t)$.

If t is a variable, let $U(t)$ consist of a single substitution, the identical one. For $f = gh$ as above, clearly both $\nu_P f(t)$ and $g\nu_P h(t)$ are equal $f(t)$.

Now let $t = F(t_1, \dots, t_n)$. Put $U_0 = U(t_1) \cup \dots \cup U(t_n)$. Consider an arbitrary nonempty subset S of U_0 which has a common expansion, and let f_S be the minimal common expansion of S. For any $u \approx u' \in P$ such that the terms $F(\nu_P f_S(t_1), \dots, \nu_P f_S(t_n))$ and u have a unifying pair, let $g_{S,u}$ and $l_{S,u}$ be the minimal unifying pair for these terms; so,

$$g_{S,u} F(\nu_P f_S(t_1), \dots, \nu_p f_S(t_n)) = l_{S,u}(u).$$

We define $U(t)$ to be the set of the permissible substitutions that either belong to U_0 or are f_S for some S or are $g_{S,u} f_S$ for some S, u.

We must prove that $U(t)$ has the property stated above. Let $f = gh$ where f is a permissible substitution, and f and h expand the same substitutions from $U(t)$. Denote by S the set of the substitutions from U_0 that can be expanded to f (and to h). Then $h = kf_S$ for some k, and all the three substitutions, f, h and f_S, expand the same substitutions from U_0. For any $i = 1, \dots, n$, $U(t_i)$ is a subset of U_0, so the three substitutions also expand the same substitutions from $U(t_i)$ and, by induction,

$$\nu_P f(t_i) = gk\nu_P f_S(t_i) \qquad \text{and} \qquad \nu_P h(t_i) = k\nu_P f_S(t_i).$$

Let us consider two cases.

The first case is when $g_{S,u}$ exists for some $u \approx u' \in P$ and f expands $g_{S,u} f_S$. Then also h expands the substitution and we can write $h = pg_{S,u} f_S$; in fact, we can suppose that $k = pg_{S,u}$. Since $f = gkf_S = gpg_{S,u} f_S$, we have

$$F(\nu_P f(t_1), \dots, \nu_P f(t_n)) = gpg_{S,u} F(\nu_P f_S(t_1), \dots, \nu_P f_S(t_n)) = gpl_{S,u}(u),$$

so that $\nu_P f(t) = gpl_{S,u}(u')$ by the definition of ν_P. Quite similarly, $\nu_P h(t) = pl_{S,u}(u')$ and we get $\nu_P f(t) = g\nu_P h(t)$ as desired.

The second case is when f (and h, as well) does not expand any $g_{S,u} f_S$. Then gk does not expand any $g_{S,u}$. By the defining property of $g_{S,u}$ this means that there is no substitution l with $gkF(\nu_P f_S(t_1), \dots, \nu_P f_S(t_n)) = l(u)$ for any

$u \approx u' \in P$. Hence the term

$$\begin{aligned} gkF(\nu_P f_S(t_1), \ldots, \nu_P f_S(t_n)) &= F(gk\nu_P f_S(t_1), \ldots, gk\nu_P f_S(t_n)) \\ &= F(\nu_P f(t_1), \ldots, \nu_P f(t_n)) \end{aligned}$$

belongs to A_P, so that, by the definition of ν_P,

$$\nu_P f(t) = F(\nu_P f(t_1), \ldots, \nu_P f(t_n)) = gkF(\nu_P f_S(t_1), \ldots, \nu_P f_S(t_n)).$$

Quite similarly $\nu_P h(t) = kF(\nu_P f_S(t_1), \ldots, \nu_P f_S(t_n))$, and we get $\nu_P f(t) = g\nu_P h(t)$ as desired.

This finishes the construction of $U(t)$ together with the proof that it has the desired property.

Clearly, it is possible to construct a finite set V of permissible substitutions such that V contains both $U(a)$ and $U(b)$ and the minimal common expansion of any subset of V belongs to V, under the assumption that it exists and is permissible. These will be our testing substitutions. If $a \approx b$ is satisfied in A_P, then $\nu_P f(a) = \nu_P f(b)$ for any $f \in V$, since $\nu_P f$ is a homomorphism of the term algebra into A_P. Conversely, suppose that $\nu_P f(a) = \nu_P f(b)$ for all $f \in V$, which can be tested in finite time. We shall show that then $a \approx b$ is satisfied in A_P, i.e., that $h(a) = h(b)$ for any homomorphism h of the term algebra into A_P; one can assume that $h(x) = x$ for any variable x not in $\mathbf{S}(a) \cup \mathbf{S}(b)$. Denote by e the substitution coinciding with h on the variables, so that $h = \nu_P e$. There is a substitution $f \in V$ such that $e = gf$ for some g, and e and f expand the same substitutions from V. We have $h(a) = \nu_P e(a) = g\nu_P f(a) = g\nu_P f(b) = \nu_P e(b) = h(b)$.

Together with 8.4, this proves the following:

8.5. Theorem. *The set of the finite sets P of equations that are a perfect base for the equational theory based on P, is recursive.*

8.6. Example. In order to describe the equational theory based on $xy \cdot zx \approx x$, one can try to prove that this single-equation base is already perfect. The test, as described above, fails and provides two more equations that should be added to a perfect base, namely, $x(y \cdot zx) \approx xz$ and $(xy \cdot z)y \approx xy$. Now the three equations together can be tested to a success; the three-element set of equations is a perfect base for the equational theory.

As demonstrated by this example, if the perfection test fails for a given finite pre-perfect base P_0, there is still a possibility to modify P_0 to obtain another finite base P_1 for the same equational theory E, which would be either perfect itself or just the next member of a sequence $P_0, P_1, \ldots$ of finite pre-perfect bases for E constructed each from the last one in the same way, the last member of which is perfect. If P_i has already been constructed, a good candidate for P_{i+1} is the union of P_i with the set of the equations $\nu_{P_i} f(a) \approx \nu_{P_i} f(b)$ such that $a \approx b \in P_0$, f is a substitution from the finite set V constructed as above, and $\nu_{P_i} f(a) \neq \nu_{P_i} f(b)$. These added equations seem to play a role similar to one played by critical pairs in the Knuth-Bendix algorithm. It may be necessary, however, to replace some of the added equations $u \approx v$ with their inverses $v \approx u$, and to delete some of the equations or in some cases to modify the set in other ways to obtain again a pre-perfect set of equations; if some old equations had to be deleted, one must then check that the new set is again a base for E, which can be done by verifying that $\nu_{P_{i+1}}(a) = \nu_{P_{i+1}}(b)$ for any $a \approx b \in P_0$. Clearly, this process of constructing the sequence $P_0, P_1, \dots$ may stop with failure, if we are not able to modify one of its members to become a pre-perfect base for E. However, it works well for many equational theories. If the sequence can be constructed, it has the property that $A_{P_{i+1}}$ is a proper subset of A_{P_i} for any i. It is natural to ask whether it can be constructed to successfully terminate always when there exists some finite perfect base Q for E such that $A_Q \subseteq A_{P_0}$. We do not know the answer to this question, and feel that it deserves a deeper study. The most usual application of the process described (in not very precise terms) above leads either to success, when the sequence can be constructed, is finite, and its last member is a finite perfect base, or to the proof that no perfect base exists (for example, one may find that a nontrivial permutational identity would have to be added), or does not terminate, producing an infinite sequence of finite pre-perfect bases for E. In the last case, it may happen that each P_i is a subset of P_{i+1}, but the union of all these bases still is not a perfect base; we then need to 'construct' a new infinite sequence of pre-perfect bases, starting with this infinite union.

8.7. Example. The equational theory based on $x(y \cdot zx) \approx x$ has an infinite perfect base consisting of the equations

$$(y_n(y_{n-1}(\dots(y_2 \cdot y_1x))))(z_m(z_{m-1}(\dots(z_2 \cdot z_1x)))) \approx y_n(y_{n-1}(\dots(y_2 \cdot y_1x)))$$

where $n, m \geq 0$ and $n - m - 1$ is divisible by 3. The proof is quite easy.

The equational theory based on $y(x \cdot xy) \approx x$ has an infinite perfect base consisting of the equations $xx \cdot x \approx x$ and $r_e s_e \approx t_e$, where e runs over all finite sequences of elements of $\{0, 1\}$ and the terms r_e, s_e and t_e are defined inductively as follows:

$$\begin{aligned} r_\emptyset &= y, & s_\emptyset &= x \cdot xy, & t_\emptyset &= x, \\ r_{e0} &= s_e, & s_{e0} &= r_e t_e, & t_{e0} &= r_e, \\ r_{e1} &= s_e \cdot s_e r_e, & s_{e1} &= t_e, & t_{e1} &= r_e. \end{aligned}$$

The proof is not so easy as in the previous case.

A set P of equations is said be *nonoverlapping* if the following are true:

(no1) if $u \approx u' \in P$, then $\mathbf{S}(u') \subseteq \mathbf{S}(u)$;
(no2) if $u \approx u' \in P$, $v \approx v' \in P$, $s \subseteq v'$, and u and s have a unifying pair, then s is a variable and $s \neq v'$;
(no3) if $u \approx u' \in P$, $v \approx v' \in P$, $s \subseteq v$, and u and s have a unifying pair, then either s is a variable or both $s = u = v$ and $u' = v'$.

8.8. Theorem. *Let P be a nonoverlapping set of equations. Then P is a perfect base for an equational theory.*

PROOF. If $u \approx u' \in P$, then u and u' do not have a unifying pair, according to (no2); in particular, u is not a variable. Conditions (pp1) and (pp2) are evidently satisfied, and instead of (pp3) it is easier to verify the condition (pp3′) formulated in the proof of 8.4. So, P is pre-perfect.

In order to prove that the algebra A_P satisfies all the equations from P, we must show that $h(a) = h(b)$ for any equation $a \approx b \in P$ and any homomorphism h of the term algebra into A_P. Denote by f the endomorphism of the term algebra which coincides with h on the set of variables.

Let us prove by induction on the complexity of t that if t is either a subterm of b or a proper subterm of a, then $h(t) = f(t)$. If t is a variable, this follows from the definition of f. Let $t = F(t_1, \dots, t_n)$. We have

$$\begin{aligned} f(t) &= F(f(t_1), \dots, f(t_n)),\\ h(t) &= F_A(h(t_1), \dots, h(t_n)) = F_A(f(t_1), \dots, f(t_n))\\ &\qquad = \nu_P(F(f(t_1), \dots, f(t_n))) = \nu_P(f(t)) \end{aligned}$$

and thus it remains to show that $f(t)$ belongs to A_P. Suppose, on the contrary, that $g(u)$ is a subterm of $f(t)$ for some substitution g and an equation $u \approx u' \in P$. For any variable x, $g(u)$ cannot be a subterm of $f(x)$, because $f(x) \in A_P$. So, $g(u) = f(s)$ for a subterm s of t which is not a variable. This means that u, s have a unifying pair, a contradiction with (no2) and (no3).

In particular, $h(b) = f(b)$. On the other hand, if $a = F(a_1, \dots, a_n)$,

$$h(a) = F_A(h(a_1), \dots, h(a_n)) = F_A(f(a_1), \dots, f(a_n))$$

which is easily seen to be equal $f(b)$ by comparing the definitions. □

An equational theory E is said to be *term finite* if every term is E-equivalent to finitely many terms only. In order to prove that the equational theory based on a given set of equations is term finite, we need to describe the equational theory. For that, one can use either the technique of term rewriting or that of perfect bases. In this case the second turns out to be more useful.

8.9. EXAMPLE. Consider the equational theory E based on $x(x(xx)) \approx (xx)x$. This equation gives both a convergent term rewrite system and a perfect base for E. In both cases, the corresponding normal form function computes the shortest term that is E-equivalent with a given term. However, in order to prove that E is term finite, we would rather need to have a normal form function computing the longest term that is E-equivalent with a given term. We need to consider a different base for E, that one consisting of the equation $(xx)x \approx x(x(xx))$. In this case, we also obtain a convergent term rewrite system, but for the proof of its finite termination we would need to know that E is term finite; the direct proof of finite termination could be quite hard. On the other hand, we immediately see that this second base is also perfect; from this it follows that E is term finite.

A quasiordering $\sqsubseteq$ on the set of terms is said to be *fully compatible* if $F(a_1, \dots, a_n) \sqsubseteq F(b_1, \dots, b_n)$ whenever $a_i \sqsubseteq b_i$ for all i, and $a \sqsubseteq b$ implies $f(a) \sqsubseteq f(b)$ for any substitution f. A quasiordering $\sqsubseteq$ such that the set

$\{u : u \sqsubseteq a\}$ is finite for any a, is called *downward finite*. A natural example of a fully compatible, downward finite quasiordering on the set of terms is the following: $u \sqsubseteq v$ if and only if every variable, and also every operation symbol, has at least as many occurrences in v as in u.

8.10. THEOREM. *Let an equational theory E have a nonoverlapping base P, such that there is a fully compatible, downward finite quasiordering $\sqsubseteq$ on the set of terms with $u \sqsubseteq u'$ whenever $u \approx u' \in P$. Then E is term finite.*

PROOF. By 8.8, P is a perfect base and thus ν_P is a normal form function for E. We have $t \sqsubseteq \nu_P(t)$ for any term t; this can be proved easily by induction on the complexity of t. Since every term u with $u \approx t \in E$ satisfies $\nu_P(u) = \nu_P(t)$, for a given term t the set of all such terms u is contained in the principal ideal of $\nu_P(t)$, which is a finite set. $\square$

8.11. EXAMPLE. The equation $((xx \cdot yy)x)x \approx xx$ is a nonoverlapping base for an equational theory E_1. Similarly, the equation $(xx \cdot x)(y \cdot yy) \approx (x(xx \cdot x))(y \cdot yy)$ is a nonoverlapping base for an equational theory E_2. While E_1 is not term finite, E_2 is term finite, which follows from 8.10, using the quasiordering described immediately preceding that theorem.

CHAPTER 14

MINIMAL SETS

1. Operations depending on a variable

An operation $f(x_1,\dots,x_n)$ on a set A is said to *depend on the variable* x_i if there exist elements $a_1,\dots,a_n$ and a_i' of A such that

$$f(a_1,\dots,a_n)\neq f(a_1,\dots,a_{i-1},a_i',a_{i+1},\dots,a_n).$$

1.1. Theorem. *Let A be an algebra and f be an n-ary polynomial of A depending on k variables ($k\geq 1$). Then for any positive integer $m\leq k$, A has an m-ary polynomial depending on all its variables.*

Proof. If we replace any variable in f on which f does not depend with a constant, we obtain a k-ary polynomial depending on all its variables. So, it remains to prove that if f is an n-ary polynomial depending on all its variables and $n>1$, then A has an $(n-1)$-ary polynomial depending on all its variables. For every $a\in A$ and every $i\in\{1,\dots,n\}$ denote by $D(a,i)$ the set of all $j\in\{1,\dots,n\}\smallsetminus\{i\}$ such that the polynomial $f(x_1,\dots,x_{i-1},a,x_{i+1},\dots,x_n)$ depends on x_j. Let us fix a pair a,i for which the set $D(a,i)$ is of maximal possible cardinality. It is enough to prove that $j\in D(a,i)$ for any $j\in\{1,\dots,n\}\smallsetminus\{i\}$. Suppose, on the contrary, that there is a $j\neq i$ such that $j\notin D(a,i)$. Since f depends on i, there exists an element b with $i\in D(b,j)$. Since $j\notin D(a,i)$, it is easy to see that $D(a,i)\subseteq D(b,j)$. But also $i\in D(b,j)$, hence $|D(b,j)|>|D(A,i)|$, a contradiction. □

Let A be an algebra and α be a congruence of A. For every polynomial f of A (or, more generally, for any α-preserving n-ary operation on A) we define an operation f_α on A/α by

$$f_\alpha(a_1/\alpha,\dots,a_n/\alpha)=f(a_1,\dots,a_n)/\alpha.$$

1.2. Lemma. *Let A be an algebra and α be a congruence of A. The polynomials of A/α are precisely the operations f_α, where f is a polynomial of A.*

Proof. Denote by H the set of the α-preserving operations f on A such that f_α is a polynomial of A/α, and by K the set of the operations f_α for a polynomial f of A. It is easy to see that H is a clone containing all the constant and all the basic operations of A. Also, it is easy to see that K is a clone containing all the constant and all the basic operations of A/α. □

1.3. Lemma. *Let A be a finite set. There exists a positive integer n such that $f^n = f^{2n}$ for all mappings f of A into A.*

Proof. Put $m = \mathbf{card}(A)$. For every $f \in A^A$ and every $a \in A$ there is a repetition in the sequence $a, f(a), \dots, f^m(a)$, i.e., there exist integers $0 \le u < m$ and $1 \le v \le m$ such that $f^u(a) = f^{u+v}(a)$; then $f^{c+u}(a) = f^{c+u+dv}(a)$ for all nonnegative integers c, d. Now if we take $n = m!$, it follows that $f^m(a) = f^{2m}(a)$ for all $f \in A^A$ and all $a \in A$. $\square$

2. Minimal algebras

By a *minimal algebra* we mean a non-trivial finite algebra A such that every unary polynomial of A is either constant or else a permutation of A.

2.1. Theorem. (Pálfy [84]) *Let A be a minimal algebra with at least three elements, having a polynomial that depends on more than one variable. Then A is polynomially equivalent with a vector space over a finite field.*

Proof. By Theorem 1.1, A has a binary polynomial depending on both its variables. Put $N = |A|$.

Claim 1. For a binary polynomial f of A and a quadruple of elements $a, b, c, d \in A$, $f(a, c) = f(a, d)$ implies $f(b, c) = f(b, d)$. Suppose $f(b, c) \ne f(b, d)$. For each $k \ge 0$ define a binary polynomial $f^{[k]}(x, y)$ of A in this way: $f^{[0]}(x, y) = y$; $f^{[k+1]}(x, y) = f(x, f^{[k]}(x, y))$. Put $g(x, y) = f^{[N!]}(x, y)$. Then g is a binary polynomial of A, and it is easy to see that $g(x, g(x, y)) = g(x, y)$ for all $x, y \in A$. Since $f(b, c) \ne f(b, d)$, the unary polynomial $h(y) = f(b, y)$ is not constant, and hence h is a permutation of A. We have $h^{N!}(y) = g(b, y) = g(b, g(b, y)) = h^{2(N!)}(y)$ for all y. Since h is a permutation, this implies $h^{N!}(y) = y$ for all $y \in A$. So, $g(b, y) = y$ for all $y \in A$. Since $g(a, c) = g(a, d)$, the mapping $y \to g(a, y)$ is constant; denote the element by e. Since $g(a, e) = e = g(b, e)$, we have $g(x, e) = e$ for all $x \in A$.

For each element $p \in A$ we have $g(p, y) = g(p, g(p, y))$, so $g(p, y)$ is either a constant or the identity.

Take an element $p \in A \setminus \{a, b\}$. (This is possible, because $|A| \ge 3$.) Also, take an element $q \in A \setminus \{e\}$. We have $g(a, q) = e$ and $g(b, q) = q$, so $g(x, q)$ is a permutation and $g(p, q) \ne e$. Since $g(p, e) = e$, we get $g(p, y) = y$ for all $y \in A$. In particular, $g(p, q) = q$. But also $g(b, q) = q$, a contradiction, since $p \ne b$.

Claim 2. If $f(a_1, \dots, a_n, a_{n+1}) = f(a_1, \dots, a_n, b_{n+1})$ for an $(n+1)$-ary polynomial f of A and elements $a_1, \dots, a_{n+1}, b_1, \dots, b_{n+1} \in A$, then $f(b_1, \dots, b_n, a_{n+1}) = f(b_1, \dots, b_n, b_{n+1})$. This follows easily from Claim 1.

Claim 3. If f is a binary polynomial of A depending on both its variables, then A is a quasigroup with respect to f. This also follows easily from Claim 1.

It follows that A has a ternary Mal'cev polynomial δ. Let us fix an element $0 \in A$. Put $x + y = \delta(x, 0, y)$ and $-x = \delta(0, x, 0)$.

Claim 4. A is an Abelian group with respect to $+, -, 0$. Put

$$p_1(x, y, z, u) = \delta(\delta(x, 0, u), 0, \delta(y, u, z)),$$

$$p_2(x,y) = \delta(x,y,\delta(y,x,0)),$$
$$p_3(x,y,z) = \delta(z,0,\delta(x,z,y)).$$

We have $(a+b)+c = p_1(a,b,c,b)$ and $a+(b+c) = p_1(a,b,c,0)$. So, by Claim 2, in order to prove $(a+b)+c = a+(b+c)$, it suffices to prove $p_1(0,b,0,b) = p_1(0,b,0,0)$. Both sides equal b.

We have $a+(-a) = p_2(a,0)$ and $0 = p_2(a,a)$. So, by Claim 1, in order to prove $a+(-a) = 0$, it suffices to prove $p_2(0,0) = p_2(0,a)$. Both sides equal 0.

We have $a+b = p_3(a,b,0)$ and $b+a = p_3(a,b,b)$. So, by Claim 2, in order to prove $a+b = b+a$, it suffices to prove $p_3(0,0,0) = p_3(0,0,b)$. But $p_3(0,0,b) = b+(-b) = 0 = p_3(0,0,0)$.

Clearly, $a+0 = a$.

Claim 5. If f is an n-ary polynomial of A, then

$$f(x_1,\dots,x_n) = \sum_{i=1}^{n} f_i(x_i) - (n-1)f(0,\dots,0)$$

where $f_i(x_i) = f(0,\dots,0,x_i,0,\dots,0)$ (x_i sitting at the i-th place). We will prove the claim by induction on n. It is clear for $n=1$. For $n=2$ we need to prove $f(x,y) = f(x,0)+f(0,y)-f(0,0)$, i.e., $f(x,y)-f(0,y) = f(x,0)-f(0,0)$. Put $g(x,y,z) = f(x,z)-f(y,z)$. So, we need to prove $g(x,0,y) = g(x,0,0)$. By Claim 2 it is sufficient to prove $g(0,0,y) = g(0,0,0)$, but this is clear.

Now let $n \geq 3$. The induction assumption applied to the $(n-1)$-ary polynomial $f(x_1,\dots,x_{n-1},x_n)$, where x_n is fixed, yields

$$\begin{aligned} f(x_1,\dots,x_{n-1},x_n) = {} & f(x_1,0,\dots,0,x_n) + \dots \\ & + f(0,\dots,0,x_{n-1},x_n) - (n-2)f(0,\dots,0,x_n). \end{aligned}$$

The desired conclusion follows by several applications of the binary case. The proof of Claim 5 is thus finished.

Denote by F the set of all unary polynomials p of A such that $p(0)=0$. For each $p \in F$, according to Claim 5 we have $p(x+y) = p(x)+p(y)$, so that p is an endomorphism of the group $(A,+,-,0)$. Since F is closed under composition and addition, it is a subring of the endomorphism ring of $(A,+,-,0)$. If $p \in F$ and p is not identically zero, then p is a permutation and so p^k is the identity for some $k \geq 1$. Hence F is a finite division ring, and thus a field. Clearly, A is a vector space over F with respect to its polynomials $+,-,0$ and $px = p(x)$.

On the other hand, according to Claim 5, every n-ary polynomial f of A can be expressed as $f(x_1,\dots,x_n) = p_1x_1+\dots+p_nx_n+c$ where $p_i(x) = f_i(x)-f_i(0)$ and $c = f(0,\dots,0)$. □

2.2. Theorem. *There are precisely seven clones on the two-element set $\{0,1\}$ that contain all constants. They are generated, respectively, by the following sets of operations (together with the constants):*

$$E_0 = \emptyset, \quad E_1 = \{'\}, \quad E_2 = \{+\}, \quad E_3 = \{\vee,\wedge,'\},$$
$$E_4 = \{\vee,\wedge\}, \quad E_5 = \{\vee\}, \quad E_6 = \{\wedge\}$$

where $'$ *is the only non-identical permutation on* $\{0,1\}$, $+$ *is addition modulo 2 and* $\vee$ *and* $\wedge$ *are the binary operations of maximum and minimum, respectively.*

PROOF. It is easy to see that the seven clones are pairwise distinct. Now let C be a clone on $\{0,1\}$ containing all constants. If C contains only essentially unary operations, then it is easy to see that C is generated by either E_0 or E_1 (and the constants). Next we assume that C contains an operation depending on at least two variables, so that, according to Theorem 1.1, C contains a binary operation f depending on both its variables.

Consider first the case when all binary operations in C satisfy Claim 1 in the proof of Theorem 2.1. Then every binary operation is either $x+y$ or $(x+y)'$. If $f(x,y)=(x+y)'$, then $x'=f(0,x)$ and $x+y=(f(x,y))'$. So, we can assume that $f(x,y)=x+y$. Proceeding similarly as in the Claims 2 and 5 of the proof of Theorem 2.1, we see that C coincides with the clone generated by $+$.

In the remaining case, we can assume that the table of f has a constant row and a non-constant row.

Suppose that C contains an operation $g(x_1,\dots,x_n)$ that is not order-preserving. There are two n-tuples $a_1,\dots,a_n$ and $b_1,\dots,b_n$ of elements of $\{0,1\}$ such that $a_i\le b_i$ for all i but $g(a_1,\dots,a_n)>g(b_1,\dots,b_n)$. Denote by I the set of all i such that $a_i=0$ and $b_i=1$, and define a unary operation h by $h(x)=g(y_1,\dots,y_n)$ where $y_i=x$ for $i\in I$ and $y_i=a_i$ for $i\notin I$. Clearly, $h\in C$ and $h(x)=x'$. Now it is easy to see that C contains either $\vee$ or $\wedge$, so that it contains E_3 and C is the clone of all operations on $\{0,1\}$.

It remains to consider the case when the table of f has a constant row and a non-constant row, and all the operations in C are order-preserving. Clearly, f is either $\vee$ or $\wedge$. Without loss of generality, we can assume that $f(x,y)=x\wedge y$. If C also contains $x\vee y$, then it is easy to see that it coincides with the clone of all order-preserving operations on $\{0,1\}$ and is generated by E_4. Let this be not the case. We are going to finish the proof by showing that every non-constant operation $h(x_1,\dots,x_n)$ of C belongs to the clone generated by $\wedge$. For every subset I of $\{1,\dots,n\}$ denote by a_I the n-tuple $a_1,\dots,a_n$ where $a_i=1$ for $i\in I$ and $a_i=0$ for $i\notin I$. Suppose that there are two incomparable minimal subsets I,J with $h(a_I)=h(a_J)=1$. Then $x\vee y$ can be derived from $h(x_1,\dots,x_n)$ by substituting x for x_i whenver $i\in I\setminus J$, y for x_i whenever $i\in J\setminus I$, 0 for x_i whenever $i\notin I\cup J$, and 1 for x_i whenever $i\in I\cap J$. But this is a contradiction, since $x\vee y$ does not belong to C. Hence there is a unique minimal subset I with $h(a_I)=1$. Then it is easy to see that $h(x_1,\dots,x_n)=\bigwedge_{i\in I}x_i$ and hence h belongs to the clone generated by $\wedge$. $\square$

A finite, non-trivial algebra A is said to be

- a *minimal algebra of type 1* (or of *unary type*) if it is polynomially equivalent to (A,G) for a subgroup G of the symmetric group on A,
- a *minimal algebra of type 2* (or of *affine type*) if it is polynomially equivalent to a vector space,

- a *minimal algebra of type 3* (or of *Boolean type*) if it is polynomially equivalent to a two-element Boolean algebra,
- a *minimal algebra of type 4* (or of *lattice type*) if it is polynomially equivalent to a two-element lattice,
- a *minimal algebra of type 5* (or of *semilattice type*) if it is polynomially equivalent to a two-element semilattice.

2.3. THEOREM. *A finite algebra is minimal if and only if it is a minimal algebra of one of the five types* $1, \dots, 5$.

PROOF. It follows from the last two theorems. $\square$

3. Minimal subsets

In this section we explain foundations of *tame congruence theory*, developed in Hobby, McKenzie [88]. The theory is much more extensive than presented here and serves also as a basis for the most modern applications of universal algebra.

3.1. THEOREM. *Let A be a finite algebra and $\langle\alpha,\beta\rangle$ be a prime quotient in the congruence lattice of A. The following two conditions are equivalent for a subset U of A:*

(1) *U is a minimal subset of A with the property $U = f(A)$ for a unary polynomial f of A such that $f(\beta) \nsubseteq \alpha$;*
(2) *U is a minimal subset of A such that $\alpha \cap U^2 \neq \beta \cap U^2$ and $U = e(A)$ for an idempotent unary polynomial e of A.*

PROOF. Clearly, it is sufficient to prove that if U is as in (1), then $U = e(A)$ for an idempotent unary polynomial e of A. Denote by K the set of all unary polynomials f of A with $f(A) \subseteq U$. Clearly, $f \in K$ implies $fg \in K$ for any unary polynomial g. Denote by α' the set of all $\langle x,y\rangle \in \beta$ such that $\langle f(x), f(y)\rangle \in \alpha$ for all $f \in K$. It is easy to see that α' is a congruence and $\alpha \subseteq \alpha' \subseteq \beta$. Since U satisfies (1), we have $U = h(A)$ for a unary polynomial h with $h(\beta) \nsubseteq \alpha$. Hence $\alpha' \neq \beta$, and we get $\alpha' = \alpha$.

There exists a pair $\langle x,y\rangle \in \beta$ such that $\langle h(x), h(y)\rangle \notin \alpha$. Suppose $\langle fg(x), fg(y)\rangle \in \alpha$ for all $f, g \in K$. By two applications of $\alpha' = \alpha$ we get $\langle x,y\rangle \in \alpha$ and hence $\langle h(x), h(y)\rangle \in \alpha$, a contradiction. This shows that there are two unary polynomials $f, g \in K$ with $\langle fg(x), fg(y)\rangle \notin \alpha$. By the minimal property of U, $fg(A) = U$ and $g(A) = U$. Hence $f(U) = U$. There is a positive integer k with $f^k = f^{2k}$ (e.g., $k = N!$, where $N = |A|$). Put $e = f^k$, so that e is an idempotent unary polynomial. Since $f(U) = U$, we have $e(U) = U$. Since $e \in K$, this implies $e(A) = U$. $\square$

For a finite algebra A and a prime quotient $\langle\alpha,\beta\rangle$ in the congruence lattice of A, by an (α,β)-*minimal subset* of A we mean any subset U satisfying the two equivalent conditions of Theorem 3.1.

3.2. LEMMA. *Let U be an (α,β)-minimal subset of A. Then β is the transitive closure of the relation $\alpha \cup R$, where R is the set of all $\langle g(x), g(y)\rangle$ such that $\langle x,y\rangle \in \beta \cap U^2$ and g is a unary polynomial of A.*

PROOF. It is easy to see that the transitive closure β' of $\alpha \cup R$ is a congruence of A and $\alpha \subseteq \beta' \subseteq \beta$. Moreover, $\beta' \neq \alpha$. $\square$

Two subsets U, V of an algebra A are said to be *polynomially isomorphic* (in A) if there are unary polynomials f, g of A such that $f(U) = V$, $g(V) = U$, $gf|_U = \mathbf{id}_U$ and $fg|_V = \mathbf{id}_V$. We then write $U \simeq V$ (if A is not clear from the context, we should write $U \simeq_A V$). We write $f : U \simeq V$ if there is a g satisfying the above conditions.

3.3. THEOREM. *Any two (α,β)-minimal subsets of a finite algebra A are polynomially isomorphic in A.*

PROOF. Let U, V be two (α,β)-minimal subsets of A. There is an idempotent unary polynomial p of A with $U = p(A)$. Denote by R the set of all $\langle q(x), q(y)\rangle$ where q is a unary polynomial and $\langle x,y\rangle \in \beta \cap V^2$. By Lemma 3.2, β is the transitive closure of $\alpha \cup R$. If $R \subseteq p^{-1}(\alpha)$, then it follows that $\beta \subseteq p^{-1}(\alpha)$, a contradiction. Hence there is an ordered pair in R not belonging to $p^{-1}(\alpha)$, i.e., there are an ordered pair $\langle a,b\rangle \in \beta \cap V^2$ and a unary polynomial q such that $\langle pq(a), pq(b)\rangle \notin \alpha$.

There is an idempotent unary polynomial e with $V = e(A)$. Put $h = pqe$. We have $h(A) \subseteq U$ and $h(\beta) \nsubseteq \alpha$ (since $\langle h(a), h(b)\rangle \notin \alpha$). By the minimality of U, $h(A) = U$. Since $h = he$, we have $h(V) = h(A) = U$.

Similarly, there is a unary polynomial f with $f(U) = f(A) = V$. Now $hf|_U$ is a permutation of U, so there exists a positive integer k with $(hf)^k|_U = \mathbf{id}_U$. Put $g = (hf)^{k-1}h$. Then $f(U) = V$, $g(V) = U$, $gf|_U = \mathbf{id}_U$ and $fg|_V = \mathbf{id}_V$. $\square$

3.4. THEOREM. *Let $\langle \alpha,\beta\rangle$ be a prime quotient in the congruence lattice of a finite algebra A. The following are true:*

(1) *For every (α,β)-minimal subset U of A and every $\langle x,y\rangle \in \beta \smallsetminus \alpha$ there is a unary polynomial f of A with $f(A) = U$ and $\langle f(x), f(y)\rangle \notin \alpha$.*

(2) *If U is an (α,β)-minimal subset of A and f is a unary polynomial such that $f(\beta \cap U^2) \nsubseteq \alpha$, then $f(U)$ is a minimal subset of A and $f : U \simeq f(U)$.*

(3) *For every unary polynomial f of A such that $f(\beta) \nsubseteq \alpha$, there is an (α,β)-minimal subset U of A with $f : U \simeq f(U)$.*

PROOF. (1) There is a unary idempotent polynomial e with $U = e(A)$. Denote by α' the set of all $\langle p,q\rangle \in \beta$ such that $\langle eg(p), eg(q)\rangle \in \alpha$ for all unary polynomials g of A. It is easy to see that α' is a congruence and $\alpha \subseteq \alpha' \subset \beta$, so that $\alpha' = \alpha$. Since $\langle x,y\rangle \notin \alpha'$, there is a unary polynomial g with $\langle eg(x), eg(y)\rangle \notin \alpha$. Put $f = eg$. We have $f(A) = U$ by the minimality of U.

(2) Take an idempotent unary polynomial e with $e(A) = U$, and a pair $\langle a,b\rangle \in \beta \cap U^2$ with $\langle f(a), f(b)\rangle \notin \alpha$. By (1) there is a unary polynomial g

with $g(A) = U$ and $\langle gf(a), gf(b)\rangle \notin \alpha$. Since $\langle gfe(a), gfe(b)\rangle \notin \alpha$, we have $gf(U) = gfe(A) = U$ by the minimality of U. Since $gf|_U$ is a permutation of U, we have $(gf)^k|_U = \mathbf{id}_U$ for some positive integer k. Then $(gf)^{k-1}g$ is the inverse of $f|_U$, so $f : U \simeq f(U)$ and this implies that $f(U)$ is a minimal subset.

(3) Take an (α, β)-minimal subset V. Proceeding similarly as at the beginning of the proof of Theorem 3.3, there are an ordered pair $\langle a, b\rangle \in \beta \cap V^2$ and a unary polynomial q with $\langle fq(a), fq(b)\rangle \notin \alpha$. Hence $\langle q(a), q(b)\rangle \notin \alpha$ and so $q(\beta \cap V^2) \not\subseteq \alpha$. By (2), the subset $U = q(V)$ is (α, β)-minimal in A. □

Let U be an (α, β)-minimal subset of A. By an (α, β)-trace in U we mean any block of $\beta \cap U^2$ that is not a block of $\alpha \cap U^2$. The union of all (α, β)-traces in U is called the *body* of U and the complement $U \setminus B$, where B is the body, is called the *tail* of U. By an (α, β)-trace in A we mean any subset that is an (α, β)-trace in some (α, β)-minimal subset of A.

3.5. Theorem. *Let U be an (α, β)-minimal subset of A and N be an (α, β)-trace in U. Then $(A|_N)/(\alpha \cap N^2)$ is a minimal algebra.*

Proof. Let h be a non-constant unary polynomial of $(A|_N)/(\alpha \cap N^2)$. According to Lemma 1.2, there is a unary polynomial g of $A|N$ with $h = g_{\alpha \cap N^2}$. Hence there is a unary polynomial f of A such that $f(N) \subseteq N$ and g is the restriction of f. Since h is non-constant, there are two elements $x, y \in N$ with $h(x/(\alpha \cap N^2)) \neq h(y/(\alpha \cap N^2))$, i.e., $\langle g(x), g(y)\rangle \notin \alpha \cap N^2$. Since N is contained in a block of β, $\langle x, y\rangle \in \beta$ and $\langle f(x), f(y)\rangle \notin \alpha$. By Theorem 3.4(2), the restriction of f to U is injective. Consequently, g is injective. But then, g is a permutation of N and h is a permutation of $N/(\alpha \cap N^2)$. □

The type of the minimal algebra $(A|_N)/(\alpha \cap N^2)$ will be called the type of the (α, β)-trace N.

3.6. Proposition. *Let U be an (α, β)-minimal subset of A and N be an (α, β)-trace of type 5 in U. Then N is the only (α, β)-trace in U and there exist an element $1 \in N$ and a binary polynomial p of A with the following properties:*

(1) *The two blocks of α contained in N are $\{1\}$ and $O = N \setminus \{1\}$*
(2) *Both U and N are closed under p and N/α is a two-element semilattice with neutral element $\{1\}$ with respect to the restriction of p_α*
(3) *$p(x, 1) = p(1, x) = p(x, x) = x$ for all $x \in U$*
(4) *$\langle p(x, u), x\rangle \in \alpha$ and $\langle p(u, x), x\rangle \in \alpha$ for all $x \in U \setminus \{1\}$ and $u \in O$*
(5) *$p(x, p(x, y)) = p(x, y)$ for all $x, y \in U$*

Proof. Since N is of type 1, there is a binary polynomial g of A such that U and N are closed under g and N/α is a two-element semilattice with respect to the restriction of g_α. Denote by I the neutral and by O the annihilating element of this semilattice. So, I and O are the two blocks of α contained in N and we have $g(I \times I) \subseteq I$ and $g(I \times O) \cup g(O \times I) \cup g(O \times O) \subseteq O$. Put $d(x) = g(x, x)$. Since d is a unary polynomial mapping I into I and O into O, d is a permutation of U. Take $k > 1$ with $d^k(x) = x$ for all $x \in U$ and put

$h(x,y) = d^{k-1}g(x,y)$. Then h has the same properties as g and, moreover, $h(x,x) = x$ for all $x \in U$. Put $h^{[0]}(x,y) = x$ and $h^{[i+1]}(x,y) = h(h^{[i]}(x,y),y)$. By 1.3 there exists an $m > 0$ such that $h^{[2m]} = h^{[m]}$; put $f(x,y) = h^{[m]}(x,y)$. Then f is a binary polynomial with the same properties as g and h and, moreover, $f(f(x,y),y) = f(x,y)$. For $z \in I$ the polynomial $r(x) = f(x,z)$ is a permutation on U, since $\langle r(z), r(u)\rangle \notin \beta$ for $u \in O$. But $rr(x) = r(x)$, so r is the identity and we get $f(x,z) = x$ for all $x \in U$ and $z \in I$. By an iteration of the second variable in $f(x,y)$ we can obtain, similarly as f was obtained from h by iterating the first variable, a binary polynomial p such that $p(x,p(x,y)) = p(x,y)$. For $x \in U$ and $z \in I$ we have evidently $p(x,z) = p(x,x) = x$; we also have $p(z,x) = x$ (by the same argument that was used to prove $f(x,z) = x$). For $z_1, z_2 \in I$ we get $z_1 = p(z_1,z_2) = z_2$. Thus $I = \{1\}$ for an element 1.

Suppose that U contains an (α,β)-trace K different from N. Since $f(x,x) = x = f(x,1)$ and $1 \in N$, it follows that $f(K \times K) \cup f(K \times N) = K$. Take an element $u \in O$. Since $f(u,u) = f(1,u)$, we have $f(x,u) = f(y,u)$ whenever $x,y \in U$ and $(x,y) \in \beta$. In particular, $f(x,0) = f(y,0)$ for all $x,y \in K$. Consequently, there is an element $a \in K$ with $f(a,u) \neq a$. Hence $f(a,u) \neq f(a,1)$. Put $s(x) = f(a,x)$. Then s is a permutation of U. Since $a \in K$, we have $p(K \cup N) \subseteq K$. But $|K \cup N| > |K|$, a contradiction.

It remains to prove (4). Let $x \in U \setminus \{1\}$ and $u \in O$. If $x \in O$ then both $p(x,u)$ and $p(u,x)$ are in O. Let $x \notin O$, so that $x \in U \setminus N$ and (since N is the only trace in U) $x/\alpha = x/\beta$. We have $\langle p(x,u), p(x,1)\rangle \in \beta$, i.e., $\langle p(x,u), x)\rangle \in \beta$ and thus $\langle p(x,u), x\rangle \in \alpha$. We can get $\langle p(u,x), x\rangle \in \alpha$ similarly. □

3.7. Proposition. *Let U be an (α,β)-minimal subset of A and N be an (α,β)-trace of type either 3 or 4 in U. Then N is the only (α,β)-trace in U and there exist two elements $0,1 \in N$ and two binary polynomials p,q of A with the following properties:*

(1) $N = \{0,1\}$
(2) *Both U and N are closed under p and q and N/α is a two-element lattice with respect to the restriction of p_α and q_α*
(3) $p(x,1) = p(1,x) = p(x,x) = x = q(x,x) = q(x,0) = q(0,x)$ *for all* $x \in U$
(4) $\langle p(x,0), x\rangle \in \alpha$, $\langle p(0,x), x\rangle \in \alpha$, $\langle q(x,1), x\rangle \in \alpha$ *and* $\langle q(1,x), x\rangle \in \alpha$ *for all* $x \in U \setminus N$
(5) $p(x,p(x,y)) = p(x,y)$ *and* $q(x,q(x,y)) = q(x,y)$ *for all* $x,y \in U$

Proof. There are two polynomials g_1 and g_2 under which N/α is a two-element lattice. Now repeat the proof of 3.6 for each of them. □

Under the assumptions of 3.6, p is called a pseudo-meet operation of U (with respect to α,β). Under the assumptions of 3.7, p,q are called pseudo-meet and pseudo-join operations of U, respectively.

It follows that if an (α,β)-minimal subset of a finite algebra A contains two distinct (α,β)-traces, then all its (α,β)-traces are of type 1 or 2.

3.8. LEMMA. *Let $\langle\alpha,\beta\rangle$ be a prime quotient in the congruence lattice of a finite algebra A and let γ be a congruence with $\gamma\subseteq\alpha$, so that $\langle\alpha/\gamma,\beta/\gamma\rangle$ is a prime quotient in the congruence lattice of A/γ. The $(\alpha/\gamma,\beta/\gamma)$-minimal subsets of A/γ are just the sets U/γ, where U is an (α,β)-minimal subset of A. Moreover, for an (α,β)-minimal subset U, the $(\alpha/\gamma,\beta/\gamma)$-traces in U/γ are just the sets N/γ, where N is an (α,β)-trace in U, and the corresponding traces are of the same types.*

PROOF. Let U be an (α,β)-minimal subset. There is an idempotent unary polynomial e of A with $U=e(A)$. Clearly, e_γ (see 1.2) is an idempotent polynomial of A/γ, $e_\gamma(A/\gamma)=U/\gamma$ and $e_\gamma(\beta/\gamma)\not\subseteq\alpha/\gamma$. Let f_γ (where f is a unary polynomial of A) be a unary polynomial of A/γ such that $f_\gamma(A/\gamma)\subseteq U/\gamma$ and $f_\gamma(\beta/\gamma)\not\subseteq\alpha/\gamma$. We have $f_\gamma=e_\gamma f_\gamma=(ef)_\gamma$ and $ef(\beta)\not\subseteq\alpha$, so that $ef(A)=U$. But then, $f_\gamma(A/\gamma)=(ef)_\gamma(A/\gamma)=U/\gamma$. We see that U/γ is an $(\alpha/\gamma,\beta/\gamma)$-minimal subset of A/γ.

Now let W be an $(\alpha/\gamma,\beta/\gamma)$-minimal subset of A/γ. Take an arbitrary (α,β)-minimal subset U of A. Then U/γ is $(\alpha/\gamma,\beta/\gamma)$-minimal in A/γ. By Theorem 3.3 there is a unary polynomial f_γ of A/γ with $f_\gamma: U/\gamma\simeq W$ (and f is a polynomial of A). Since $f_\gamma(\beta/\gamma)\cap(U/\gamma)^2\not\subseteq\alpha/\gamma$, we have $f(\beta)\cap U^2\not\subseteq\alpha$. By Theorem 3.4(2) it follows that $f(U)$ is an (α,β)-minimal subset of A. Clearly, $W=f(U)/\gamma$.

The traces part can be proved easily. □

3.9. LEMMA. *Let U be an (α,β)-minimal subset of a finite algebra A. Then $(\beta\cap U^2)/(\alpha\cap U^2)$ is an Abelian congruence of $U/(\alpha\cap U^2)$ if and only if every (α,β)-trace in U is of type either 1 or 2.*

PROOF. Clearly, it is sufficient to prove the statement under the assumptions $\alpha=\mathbf{id}_A$ and $U=A$.

Let β be an Abelian congruence of U and N be an $(\mathbf{id}_U,\beta)$-trace. If f is an n-ary polynomial of U such that $f(N^n)\subseteq N$, then the condition in the above recalled definition of Abelian congruence is true for all $u,v,x_i,y_i\in N$, which means that the induced algebra N is Abelian. Consequently, N is of type either 1 or 2.

Now let β be not Abelian. Take n minimal such that for an n-ary polynomial f of U, there are pairs $\langle u,v\rangle\in\beta$ and $\langle x_i,y_i\rangle\in\beta$ $(i=2,\dots,n)$ with

$$f(u,x_2,\dots,x_n)=f(u,y_2,\dots,y_n)\text{ and }f(v,x_2,\dots,x_n)\neq f(v,y_2,\dots,y_n).$$

Clearly, $n>1$, $u\neq v$ and $x_i\neq y_i$ for all i. Put $N_1=u/\beta$, $N_i=x_i/\beta$ $(i=2,\dots,n)$ and $K=f(u,x_2,\dots,x_n)/\beta$. Then $N_1,\dots,N_n$ are traces, and $f(N_1\times\dots\times N_n)\subseteq K$.

It follows from the minimality of n that for each $j=1,\dots,n$ there are elements $c_1,\dots,c_{j-1},c_{j+1},\dots,c_n$ such that the unary polynomial

$$h_j(x)=f(c_1,\dots,c_{j-1},x,c_{j+1},\dots,c_n)$$

is not constant on N_j. Let us fix one such unary polynomial h_j for each j. Since U is $(\mathbf{id}_U,\beta)$-minimal, it follows that h_j is a permutation of U. Then also the inverse h_j^{-1} is a polynomial of U. Now h_j must permute the blocks of β, and so $h_j(N_j) = K$. Put

$$g(z_1, \dots, z_n) = f(h_1^{-1}(z_1), \dots, h_n^{-1}(z_n)),$$

so that g is a polynomial of U. Clearly, $g(K^n) = K$. We have

$$g(h_1(u), h_2(x_2), \dots, h_n(x_n)) = g(h_1(u), h_2(y_2), \dots, h_n(y_n))$$

while

$$g(h_1(v), h_2(x_2), \dots, h_n(x_n)) = g(h_1(v), h_2(y_2), \dots, h_n(y_n)).$$

Thus the induced algebra $U|_K$ is not Abelian, and the type of the trace K is neither 1 nor 2. □

3.10. Proposition. *Let U be an (α, β)-minimal subset of A and N be an (α, β)-trace of type 2 in U; denote by B the body of U. Then all (α, β)-traces of A are polynomially isomorphic in A (and so of type 2) and there exists a ternary polynomial d of A with the following properties:*

(1) *U is closed under d and $d(x, x, x) = x$ for all $x \in U$*
(2) *$d(x, x, y) = y = d(y, x, x)$ for all $x \in B$ and $y \in U$*
(3) *for any $a, b \in B$, the unary polynomials $d(x, a, b)$, $d(a, x, b)$ and $d(a, b, x)$ are permutations of U*
(4) *B is closed under d*

Moreover, every ternary polynomial d satisfying (1) and (2) also satisfies (3) and (4).

Proof. Let N be an (α, β)-trace in U. Since $(A|_N)/(\alpha \cap N^2)$ is a vector space, there exists a ternary polynomial f of A such that U and N are closed under f and $f_\alpha(x/\alpha, y/\alpha, z/\alpha) = x/\alpha - y/\alpha + z/\alpha$ for all $x, y, z \in N$. Denote by Φ the set of all ternary polynomials f with this property (so that Φ is nonempty) and put

$$\Phi_1 = \{f \in \Phi : f(x, x, x) = x \text{ for all } x \in U\},$$
$$\Phi_2 = \{f \in \Phi_1 : f(x, x, y) = y \text{ for } x \in B \text{ and } y \in U\},$$
$$\Phi_3 = \{f \in \Phi_2 : f(y, x, x) = y \text{ for } x \in B \text{ and } y \in U\}.$$

Claim 1. If $f \in \Phi$ then (restrictions of) the unary polynomials $f(x, a, b)$, $f(a, x, b)$ and $f(a, b, x)$ are permutations of U for any $a, b \in N$. This follows easily from the minimality of U.

Claim 2. Φ_1 is nonempty. Take $f \in \Phi$ and put $p(x) = f(x, x, x)$. Then p is a permutation of U, since $p(N)$ is not contained in a block of α. Clearly, the ternary polynomial $p^{-1}f(x, y, z)$ belongs to Φ_1.

Claim 3. If $f \in \Phi_1$ and $a \in B$ then (restrictions of) $f(x, a, a)$ and $f(a, a, x)$ are permutations of U. As the two cases are symmetric, we will give the proof only for $f(x, a, a)$. If $a \in B$, it follows from Claim 1. Let $a \in N'$ where N' is a trace in U different from (and thus disjoint with) N. For y, z fixed put $r_{y,z}(x) = f(x, y, z)$. By 1.3 there exists a positive integer n such that $r_{y,z}^n = r_{y,z}^{2n}$ for all y, z. So, the ternary polynomial $g(x, y, z) = r_{y,z}^n(x)$ satisfies $g(g(x, y, z), y, z) = g(x, y, z)$. For $b, c \in N$ the unary polynomial $g(x, b, c)$ is a permutation of U, since f is; hence $g(x,b,c)=x$ for $x\in U$ and $b,c\in N$. Also, clearly $g(x,x,x)=x$. Thus $g(N'\times N\times N)\cup g(N'\times N'\times N')\subseteq N'$, since N and N' are blocks of β. For $a'\in N'$ the unary polynomial $G(x)=g(a',x,x)$ satisfies $G(N\cup N')\subseteq N'$, so that G is not a permutation of U and hence $G(\beta\cap(U^2))\subseteq\alpha$. Hence for all $a'\in N'$ and $v\in N'$ we have $\langle g(a',v,v),g(a',a',a')\rangle\in\alpha$, i.e., $\langle g(a',v,v),a'\rangle\in\alpha$. Take $a'\in N'$ such that $\langle a',a\rangle\notin\alpha$. We have $\langle g(a',a,a),g(a,a,a)\rangle=\langle a',a\rangle\notin\alpha$, so that $g(x,a,a)$ must be a permutation of U and thus $f(x,a,a)$ is a permutation of U.

Claim 4. Φ_2 is nonempty. Take $f\in\Phi_1$. Put $r_x(y)=f(x,x,y)$ and take $n>1$ such that $r_x^n=r_x^{2n}$ for all x, so that the binary polynomial $u(x,y)=r_x^n(y)$ satisfies $u(x,u(x,y))=u(x,y)$. Put $f'(x,y,z)=r_x^{n-1}(f(x,y,z))$. For $x,y,z\in N$ we get

$$y/\alpha=r_x^2(y)/\alpha=\cdots=r_x^{n-1}(y)/\alpha$$

by computing it in the vector space, so that $\langle f'(x,y,z),f(x,y,z)\rangle\in\alpha$. Thus $f'\in\Phi$. Clearly $f'(x,x,x)=x$ for all $x\in U$, so that $f'\in\Phi_1$. We have $f'(x,x,y)=u(x,y)$ for all $x,y\in U$. Let $x\in B$. By Claim 3, r_x is a permutation of U and so $u(x,y)=y$ for all y. Thus $f'(x,x,y)=y$ and $f'\in\Phi_2$.

Claim 5. Φ_3 is nonempty. Take $f\in\Phi_2$. Put $r_y(x)=f(x,y,y)$ and take $n>1$ such that $r_y^n=r_y^{2n}$ for all y, so that the binary polynomial $v(x,y)=r_y^n(x)$ satisfies $v(v(x,y),y)=v(x,y)$. Put $f'(x,y,z)=r_z^{n-1}(f(x,y,z))$. Similarly as in Claim 4, $f'\in\Phi_1$ and $f'(y,x,x)=y$ for all $x\in B$ and $y\in U$. Let $x\in B$. We have $f'(x,x,y)=r_y^{n-1}(f(x,x,y))=y$, since $f(x,x,y)=y$. Thus $f'\in\Phi_3$.

We have proved the existence of a ternary polynomial d with properties (1) and (2). Let $a,b\in B$, so that $a\in N_0$ and $b\in N_1$ where N_0,N_1 are (not necessarily distinct) (α,β)-traces contained in U. Define unary polynomials f_0,f_1,f_2 by $f_0(x)=d(x,a,b)$, $f_1(x)=d(a,x,b)$ and $f_2(x)=d(a,b,x)$. Clearly, for $i=1,2,3$, either f_i is a permutation of U or $f_i(\beta\cap U^2)\subseteq\alpha$ and these two possibilitites exclude each other.

Claim 6. Either f_i are permutations of U for all $i=0,1,2$ or $f_i(\beta\cap U^2)\subseteq\alpha$ for all $i=1,2,3$. Assume that f_0 is a permutation of U. Then $\langle x,y\rangle\in\alpha$ if and only if $\langle f_0(x),f_0(y)\rangle\in\alpha$ for all $x,y\in U$. Take $u\in N_0$ with $\langle u,a\rangle\notin\alpha$. Then $\langle f_0(a),f_0(u)\rangle\notin\alpha$, i.e., $\langle d(u,u,b),d(u,a,b)\rangle=\langle b,d(u,a,b)\rangle=\langle d(a,a,b),d(u,a,b)\rangle\notin\alpha$. By 3.9, $\beta\cap U^2$ is Abelian over $\alpha\cap U^2$, so it follows that $\langle d(a,u,b),d(a,a,b)\rangle\notin\alpha$, i.e., $\langle f_1(u),f_1(a)\rangle\notin\alpha$ and f_1 is a permutation of U. All the steps were reversible, so f_0 is a permutation of U if and only if f_1 is a permutation of U. Quite similarly, f_1 is a permutation of U if and only if f_2 is a permutation of U.

Now suppose that f_0,f_1,f_2 all fail to be a permutation of U. Put $w(x)=d(a,d(a,x,b),x)$. If $x\in N_0$ then $\langle d(a,x,b),b\rangle=\langle f_1(x),f_1(a)\rangle\in\alpha$ and so $\langle w(x),f_2(x)\rangle\in\alpha$. It means that all elements of $w(N_0)$ are congruent modulo α and thus w is not a permutation of U and $(w(N_1))^2\subseteq\alpha$. For $x\in N_1$ we have $\langle d(a,x,b),a\rangle=\langle f_1(x),f_1(b)\rangle\in\alpha$ and so $\langle w(x),x\rangle=\langle w(x),d(a,a,x)\rangle\in\alpha$. Thus, for $x,y\in N_1$, $x\equiv w(x)\equiv w(y)\equiv y$ modulo α. But N_1 is a trace and we get a contradiction.

We have proved (3). If $a, b \in B$ then the polynomial $f_2(x) = d(a, b, x)$ is a permutation of U, so it maps traces onto traces and B onto itself. We have proved (4). It remains to prove that any two (α, β)-traces N_0, N_1 of A are polynomially isomorphic. By 3.3 it is sufficient to assume that they are contained in the same (α, β)-minimal set U. Take $a \in N_0$ and $b \in N_1$. Since $f_2(b) = a$ and f_2^{-1} is a polynomial, we have $f_2(N_1) = N_0$ and $f_2^{-1} : N_0 \simeq N_1$. □

Under the assumptions of 3.10, d is called a pseudo-Mal'cev operation of U (with respect to α, β).

It follows from the above results that for any finite algebra A and any prime quotient $\langle \alpha, \beta \rangle$ in the congruence lattice of A, all (α, β)-traces of A are of the same type. This type is called the type of the prime quotient $\langle \alpha, \beta \rangle$.

3.11. LEMMA. *Let $\langle \alpha, \beta \rangle$ be a prime quotient in the congruence lattice of a finite algebra A. Then β is the transitive closure of $\alpha \cup R$, where*

$$R = \cup\{N^2 : N \textit{ is an } (\alpha, \beta)\textit{-trace in } A\}.$$

PROOF. Take any (α, β)-minimal set U. Denote by P the set of the ordered pairs $\langle g(x), g(y) \rangle$ such that $\langle x, y \rangle \in \beta \cap U^2$ and g is a unary polynomial of A. By Lemma 3.2, β is the transitive closure of $\alpha \cup P$. So, it is enough to show that $\alpha \cup P \subseteq \alpha \cup R$. Let $\langle x, y \rangle \in \beta \cap U^2$ and g be a unary polynomial with $\langle g(x), g(y) \rangle \notin \alpha$. Then $\langle x, y \rangle \notin \alpha$ and the set $N = (x/\beta) \cap U$ is an (α, β)-trace in U containing both x and y. By Theorem 3.4(2), $g(U)$ is a minimal subset of A and $g : U \simeq g(U)$. Hence $g(N)$ is an (α, β)-trace, and we get $\langle g(x), g(y) \rangle \in R$. □

3.12. THEOREM. *A prime quotient $\langle \alpha, \beta \rangle$ in the congruence lattice of a finite algebra A is of unary type if and only if β is strongly Abelian over α.*

PROOF. By 9.1.5 and 3.8 it is sufficient to consider the case when $\alpha = \mathbf{id}_A$. If β is strongly Abelian over $\mathbf{id}_A$ then it follows from 3.6, 3.7 and 3.10 that $\langle \mathbf{id}_A, \beta \rangle$ cannot be of any of the types 2 through 5, so that it is of type 1. Let $\langle id_A, \beta \rangle$ be of type 1.

Claim 1. If N, N_0 and N_1 are $(\mathbf{id}_A, \beta)$-traces of A and f is a binary polynomial such that $f(N_0 \times N_1) \subseteq N$, then $f \restriction (N_0 \times N_1)$ depends on at most one variable. Suppose that it depends on both variables, so that $f(a_1, b) \neq f(a_2, b)$ and $f(c, d_1) \neq f(c, d_2)$ for some $a_1, a_2, c \in N_0$ and $b, d_1, d_2 \in N_1$. Put $g_0(x) = f(x, b)$ and $g_1(x) = f(c, x)$. It follows from 3.4(2) that $g_i : N_i \simeq N$ $(i = 0, 1)$, so that there are unary polynomials h_0, h_1 with $g_i h_i \restriction N = \mathbf{id}_N$ and $h_i g_i \restriction N_i = \mathbf{id}_{N_i}$ $(i = 0, 1)$. Put $p(x, y) = f(h_0(x), h_1(y))$. Then p restricted to N is a polynomial of the minimal algebra $A \restriction N$ of type 1, so that $p \restriction (N \times N)$ depends on at most one variable which clearly gives a contradiction.

Claim 2. If N is an $(\mathbf{id}_A, \beta)$-trace, T_0, T_1 are blocks of β and f is a binary polynomial such that $f(T_0 \times T_1) \subseteq N$, then $f \restriction (T_0 \times T_1)$ depends on at most one variable. Suppose, on the contrary, that there are elements $a \in T_0$ and $b \in T_1$ such that $f(x, b)$ is non-constant on T_0 and $f(a, y)$ is non-constant on T_1. By an easy application of 3.11, there are $(\mathbf{id}_A, \beta)$-traces $N_0 \subseteq T_0$ and $N_1 \subseteq T_1$ such that $f(x, b)$ is non-constant on N_0 and $f(a, x)$ is non-constant on N_1. Let y be an arbitrary element of T_1. By 3.11 there are elements $b_0, \dots, b_k$ and $(\mathbf{id}_A, \beta)$-traces $M_0, \dots, M_k$ such that $b_0 = b$, $b_k = y$ and $\{b_i, b_{i+1}\} \subseteq M_i$ for $i < k$. It is easy to prove by induction on i that $f(x, b) = f(x, b_i)$ for all $x \in N_0$. In particular, $f(x, b) = f(x, y)$ for all $x \in N_0$. Hence $f(a, b) = f(a, y)$ for all $y \in T_1$, a contradiction.

Claim 3. If N is an $(\mathbf{id}_A, \beta)$-trace, p is an n-ary polynomial of A and $f(T_1 \times \dots \times T_n) \subseteq N$ where T_i are blocks of β then $f \restriction (T_1 \times \dots \times T_n)$ depends on at most one variable. This follows easily from Claim 2 by induction, using 1.1.

Let f be an $(n+1)$-ary polynomial of A and $c_0 \overset{\beta}{\equiv} d_0$, $c_i \overset{\beta}{\equiv} d_i \overset{\beta}{\equiv} e_i$ $(i = 1, \dots, n)$ be elements such that $f(c_0, e_1, \dots, e_n) \neq f(d_0, e_1, \dots, e_n)$. We must prove $f(c_0, \dots, c_n) \neq f(d_0, \dots, d_n)$. Put $T_i = c_i/\beta$. By 3.4(1) there exist an $(\mathbf{id}_A, \beta)$-minimal set U and a unary polynomial h such that $h(A) = U$ and $hf(c_0, e_1, \dots, e_n) \neq hf(c_0, e_1, \dots, e_n)$. Let N be the $(\mathbf{id}_A, \beta)$-trace containing $hf(c_0, e_1, \dots, e_n)$. Clearly, $hf(T_0 \times \dots \times T_n) \subseteq N$. By Claim 3, $hf \restriction (T_0 \times \dots \times T_n)$ depends on at most one variable; but it depends on the first variable, so it does not depend on the other ones. Hence $hf(c_0, \dots, c_n) = hf(c_0, e_1, \dots, e_n)$ and $hf(d_0, \dots, d_n) = hf(d_0, e_1, \dots, e_n)$. It follows that $hf(c_0, \dots, c_n) \neq hf(d_0, \dots, d_n)$ and consequently $f(c_0, \dots, c_n) \neq f(d_0, \dots, d_n)$. □

3.13. THEOREM. *A prime quotient $\langle \alpha, \beta \rangle$ in the congruence lattice of a finite algebra A is of affine type if and only if β is Abelian but not strongly Abelian over α.*

PROOF. It is sufficient to consider the case when $\alpha = \mathbf{id}_A$. If β is Abelian but not strongly Abelian over $\mathbf{id}_A$ then it follows from 3.6 and 3.7 that $\langle \mathbf{id}_A, \beta \rangle$ cannot be of any of the types 3 through 5 and it follows from 3.12 that it cannot be of type 1, so that it is of type 2. Let $\langle \mathbf{id}_A, \beta \rangle$ be of type 2. By 3.12, β is not strongly Abelian. It remains to prove that β is Abelian. Suppose, on the contrary, that there exist an $(n+1)$-ary polynomial f and pairs $\langle a, b \rangle \in \beta$, $\langle c_i, d_i \rangle \in \beta$ $(i = 1, \dots, n)$ such that $f(a, c_1, \dots, c_n) = f(a, d_1, \dots, d_n)$ but $f(b, c_1, \dots, c_n) \neq f(b, d_1, \dots, d_n)$. It follows from 3.11 that there exists such a situation with $\{a, b\}$ contained in one $(\mathbf{id}_A, \beta)$-trace N; denote by U the $(\mathbf{id}_A, \beta)$-minimal set with $N \subseteq U$.

We can assume that $f(A^{n+1}) \subseteq U$ and that there is an $(\mathbf{id}_A, \beta)$-trace $N' \subseteq U$ such that the elements $f(a, c_1, \dots, c_n)$, $f(b, c_1, \dots, c_n)$ and $f(b, d_1, \dots, d_n)$ all belong to N'. Indeed, by 3.4(1) there is a unary polynomial h with $h(A) = U$ and $hf(b, c_1, \dots, c_n) \neq hf(b, d_1, \dots, d_n)$; we could replace f by hf.

Also, we can assume that $N' = N$. Indeed, by 3.10 N, N' are polynomially equivalent, so that there is a unary polynomial g such that $g(U) \subseteq U$ and a restriction of g is a bijection of N' onto N; we could replace f by gf.

For $i = 1, \dots, n$ put $T_i = c_i/\beta$. Clearly, $f(N \times T_1 \times \dots \times T_n) \subseteq N$. By 3.11 for every $i = 1, \dots, n$ there are $(\mathbf{id}_A, \beta)$-traces $N_{i,0}, \dots, N_{i,k_i} \subseteq T_i$ such that $c_i \in N_{i,0}$, $d_i \in N_{i,k_i}$ and $N_{i,j} \cap N_{i,j+1} \neq \emptyset$ for $0 \leq j < k_i$. We can assume that $k_1 = \dots = k_n$; denote this number by k. By 3.10 $N_{i,j} = g_{i,j}(N)$ bijectively for some unary polynomials $g_{i,j}$ ($i = 1, \dots, n$ and $j = 0, \dots, k$). For $j = 0, \dots, k$ put $f_j(x, x_1, \dots, x_n) = f(x, g_{1,j}(x_1), \dots, g_{n,j}(x_n))$, so that f_j is an $(n+1)$-ary polynomial of A and $f_j(N^{n+1}) \subseteq N$. Now $A|_N$ is a vector space over a finite field F, so there exist elements $r_{i,j} \in F$ and $e_j \in N$ such that

$$f_j(x_0, \dots, x_n) = r_{0,j}x_0 + \cdots + r_{n,j}x_n + e_j$$

for all $x_0, \dots, x_n \in N$.

Let $0 \le j < k$. For $i = 1, \dots, n$ there are elements $u_i, v_i \in N$ with $g_{i,j}(u_i) = g_{i,j+1}(v_i)$. For all $x \in N$ we have $f_j(x, u_1, \dots, u_n) = f_{j+1}(x, v_1, \dots, v_n)$, i.e.,

$$r_{0,j}x + r_{1,j}u_1 + \cdots + r_{n,j}u_n + e_j = r_{0,j+1}x + r_{1,j+1}v_1 + \cdots + r_{n,j+1}v_n + e_{j+1}$$

from which we get $r_{0,j} = r_{0,j+1}$.

Consequently, $r_{0,0} = r_{0,k}$. For $i = 1, \dots, n$ take elements $c'_i, d'_i \in N$ with $g_{i,0}(c'_i) = c_i$ and $g_{i,k}(d'_i) = d_i$. We have

$$f_0(a, c'_1, \dots, c'_n) = f(a, c_1, \dots, c_n) = f(a, d_1, \dots, d_n) = f_k(a, d'_1, \dots, d'_n),$$

i.e.,

$$r_{0,0}a + r_{1,0}c'_1 + \cdots + r_{n,0}c'_n + e_0 = r_{0,0}a + r_{1,k}d'_1 + \cdots + r_{n,k}d'_n + e_k.$$

This remains valid if a is replaced by b and from that we get $f(b, c_1, \dots, c_n) = f(b, d_1, \dots, d_n)$. $\square$

CHAPTER 15

THE LATTICE OF EQUATIONAL THEORIES

1. Intervals in the lattice

1.1. LEMMA. *A lattice is isomorphic to an interval in the lattice of equational theories of groupoids if and only if it is an algebraic lattice containing at most countably many compact elements.*

PROOF. An equational theory is a compact element in the lattice of all equational theories if and only if it is finitely based. Thus the lattice of equational theories of groupoids (or of algebras of any fixed at most countable signature) has at most countably many compact elements. Every interval of an algebraic lattice with countably many elements is itself an algebraic lattice with countably many elements. We have obtained the direct implication.

Conversely, let L be an algebraic lattice with at most countably many compact elements. According to Theorem 3.12.14, L is isomorphic to the congruence lattice of an algebra with only unary operations. The proof of that theorem yields a countable algebra if there are only countably many compact elements in L. It is also easy to see that it is sufficient to take only countably many unary operations. Thus we may assume that L is isomorphic to the congruence lattice of an algebra A with the underlying $A = \omega - \{0\}$ (the set of positive integers), and with unary operations f_i ($i \in A$). Clearly, we can also assume that $f_i = \mathbf{id}_A$ for all even numbers $i \in A$.

Let X be an infinite countable set of variables and T be the groupoid of terms over X. For any term t and any finite sequence $z = \langle x_1, \dots, x_n\rangle$ ($n \geq 0$) of variables define two terms $t\alpha z$ and $t\beta z$ in this way: if $n = 0$ then $t\alpha z = t\beta z = t$; if $t \geq 1$ then $t\alpha z = (t\alpha\langle x_1, \dots, x_{n-1}\rangle)x_n$ and $t\beta z = x_n(t\beta\langle x_1, \dots, x_{n-1}\rangle)$.

Let x be a variable and $s = \langle z_1, \dots, z_k\rangle$ be a finite sequence of finite sequences of variables. We put $h(xx, s) = ((((xx)\alpha z_1)\beta z_2)\alpha z_3)\dots\varepsilon z_k$ where $\varepsilon = \alpha$ if k is odd and $\varepsilon = \beta$ if k is even. By a defining pair for a term t we mean a pair $x, \langle z_1, \dots, z_k\rangle$ such that x is a variable, $\langle z_1, \dots, z_k\rangle$ is a finite sequence of finite sequences of variables, $t = h(xx, \langle z_1, \dots, z_k\rangle)$, $k \geq 2$, k is even, $z_1, \dots, z_{k-1}$ are nonempty, and z_2 is nonempty. Of course, every term has at most one defining pair.

If $x, \langle\langle x_{1,1}, \dots, x_{1,n_1}\rangle, \dots, \langle x_{k,1}, \dots, x_{k,n_k}\rangle\rangle$ is a defining pair for a term t, then we define a positive integer $p(t)$ as follows: if $n_k \neq 0$ then $p(t) = f_{n_{k-1}}(f_{n_{k-3}}(\dots(f_{n_3}(n_1))))$; if $n_k = 0$ then $p(t) = f_{n_{k-3}}(f_{n_{k-5}}(\dots(f_{n_3}(n_1))))$.

The finite sequence $\langle x, \dots, x\rangle$ with i members will be denoted by $\langle x, \dots, x\rangle_i$. If $x \in X$ and $n \in A$, we put $h_n(x) = x((x)\alpha\langle x, \dots, x\rangle_n)$. Clearly, $p(h_n(x)) = n$.

Denote by U the set of all terms that have a subterm $ab.cd$ for some terms a, b, c, d. If t has a defining pair then evidently $t \notin U$.

For every congruence r of A define a binary relation r^* on T as follows. For two terms u, v let $\langle u, v\rangle \in r^*$ if and only if either $u = v$ or $\{u, v\} \subseteq U$ or the following holds: $\mathbf{S}(u) = \mathbf{S}(v)$; u has a defining pair $x, \langle z_1, \dots, z_k\rangle$ and v has a defining pair $y, \langle z'_1, \dots, z'_m\rangle$; z_k, z'_m are either both empty or both nonempty; if they are empty, then $z_{k-1} = z'_{m-1}$; finally, $\langle p(u), p(v)\rangle \in r$.

Evidently, r^* is an equivalence. Let us prove that it is a congruence of T. Let $\langle u, v\rangle \in r^*$ where $u \neq v$ and let w be a term. If $w \notin X$ then $uw, vw, wu, wv \in U$, so that $\langle uw, vw\rangle \in r^*$ and $\langle wu, wv\rangle \in r^*$. Let $w \in X$. Then $\langle uw, vw\rangle \in r^*$ is easy; $\langle wu, wv\rangle \in r^*$ is easy if z_k, z'_m are nonempty; if z_k, z'_m are empty then $\langle wu, wv\rangle \in r^*$ follows from the fact that r is a congruence of A.

Let us prove that r^* is a fully invariant congruence of T. Let $\langle u, v\rangle \in r^*$ and let g be an endomorphism of T. If $g(\mathbf{S}(u)) \subseteq X$, then $\langle g(u), g(v)\rangle \in r^*$ follows immediately from the definition of r^*; in the opposite case evidently $g(u), g(v) \in U$, so that $\langle g(u), g(v)\rangle \in r^*$ as well.

It is easy to see that $\langle n, m\rangle \in r$ if and only if $\langle h_n(x), h_m(x)\rangle \in r^*$ for some (or any) variable x. If r_1, r_2 are two congruences of A, then $r_1 \subseteq r_2$ if and only if $r_1^* \subseteq r_2^*$. Denote by I the least and by J the largest congruence of A. Let S be an arbitrary fully invariant congruence of T such that $I^* \subseteq S \subseteq J^*$. Define a binary relation R on A by $\langle n, m\rangle \in R$ if and only if $\langle h_n(x), h_m(x)\rangle \in S$ for some (and consequently any) variable x. We are going to prove that R is a congruence of A and $S = R^*$.

Evidently, R is an equivalence. Let $\langle n, m\rangle \in R$ and let $i \in A$. Since $\langle h_n(x), h_m(x)\rangle \in S$ and S is a congruence, we have $\langle x(h_n(x)\alpha\langle x, \dots, x\rangle_i), x(h_m(x)\alpha\langle x, \dots, x\rangle_i)\rangle \in S$. The first member of this pair is congruent modulo I^* (and consequently modulo S) with $h_{f_i(n)}(x)$ and the second member with $h_{f_i(m)}(x)$. Thus $\langle f_i(n), f_i(m)\rangle \in R$.

Let us prove $R^* \subseteq S$. Let $\langle u, v\rangle \in R^*$. If either $u = v$ or $u, v \in U$ then $\langle u, v\rangle \in S$ is clear. In the remaining case we have $\langle p(u), p(v)\rangle \in R$, so that $\langle h_{p(u)}(x), h_{p(v)}(x)\rangle \in S$. Denote by $y_1, \dots, y_q$ the variables contained in $\mathbf{S}(u)$. Put $u' = y_1(y_2(\dots(y_q(h_{p(u)}(x)))))$ and $v' = y_1(y_2(\dots(y_q(h_{p(v)}(x)))))$. Since S is a congruence, we have $\langle u', v'\rangle \in S$. If the last sequence of variables in the defining pair for u is nonempty, then $\langle u, u'\rangle \in I^* \subseteq S$ and $\langle v, v'\rangle \in I^* \subseteq S$, so that $\langle u, v\rangle \in S$. If the last sequence is empty, denote by $\langle x_1, \dots, x_c\rangle$ the last nonempty sequence. Since S is a congruence, we have $\langle((u'x_1)\dots)x_c, ((v'x_1)\dots)x_c\rangle \in S$. The first member of this pair is congruent modulo I^* with u and the second member with v, so that $\langle u, v\rangle \in S$ again.

Let us prove $S \subseteq R^*$. Let $\langle u, v\rangle \in S$. Then $\langle u, v\rangle \in J^*$. If $u = v$, then $\langle u, v\rangle \in R^*$ evidently. If $u, v \in U$, then $\langle u, v\rangle \in I^* \subseteq R^*$. Let u have a defining pair $x, \langle\langle x_{1,1}, \dots, x_{1,n_1}\rangle, \dots, \langle x_{k,1}, \dots, x_{k,n_k}\rangle\rangle$ and let $y, \langle\langle y_{1,1}, \dots, y_{1,m_1}\rangle, \dots, \langle y_{d,1}, \dots, y_{d,m_d}\rangle\rangle$ be a defining pair for v. Denote by g the endomorphism of T sending every variable to x. Since S is fully invariant, $\langle g(u), g(v)\rangle \in S$. If $n_k \neq 0$, then $\langle g(u), h_{p(u)}(x)\rangle \in I^* \subseteq S$ and similarly for v, so that

$\langle h_{p(u)}(x), h_{p(v)}(x)\rangle \in S$, i.e., $\langle p(u), p(v)\rangle \in R$; by the definition of R^* this means that $\langle u, v\rangle \in R^*$. If $n_k = 0$, then $\langle g(u), h_{p(u)}\alpha\langle x, \dots, x\rangle_j\rangle \in I^* \subseteq S$ where $j = n_{k-1} = m_{d-1}$ and similarly for v. Put $i = j$ if j is even and $i = j+1$ if j is odd. Since S is a congruence, we have $\langle x(h_{p(u)}\alpha\langle x, \dots, x\rangle_i), x(h_{p(v)}\alpha\langle x, \dots, x\rangle_i)\rangle \in S$. The first member of this pair is congruent modulo I^* with $h_{f_i(p(u))}(x)$ and the second member with $h_{f_i(p(v))}(x)$. Since $f_i = \mathbf{id}_A$, we get $\langle h_{p(u)}(x), h_{p(v)}(x)\rangle \in S$, so that $\langle p(u), p(v)\rangle \in R$. By the definition of R^* this means that $\langle u, v\rangle \in R^*$ again.

Thus the interval determined by I^* and J^* in the lattice of equational theories of groupoids is isomorphic to the congruence lattice of A, which is isomorphic to the given lattice L. □

We have actually proved that every algebraic lattice with at most countably many compact elements is isomorphic to a principal ideal in the lattice of equational theories of groupoids satisfying $(xy \cdot zu)v = v(xy \cdot zu) = xy \cdot zu$.

1.2. Lemma. *A lattice is isomorphic to an interval in the lattice of equational theories of algebras with two unary operations if and only if it is an algebraic lattice containing at most countably many compact elements.*

Proof. Denote the two unary operation symbols by F and G. Let T be the algebra of terms over X. Take an algebra A with unary operations $f_1, f_2, \dots$ similarly as in the proof of 1.1. Let U denote the set of the terms $h_1 h_2 \dots h_n(x)$ such that $x \in X$, $h_i \in \{F, G\}$ for all i and there exists an $i \in \{2, \dots, n-1\}$ with $h_{i-1} = h_i = F$. For every congruence r of A define a fully invariant congruence r^* of T as follows. For two terms u, v let $\langle u, v\rangle \in r^*$ if and only if either $u = v$ or $u, v \in U$ or there exist positive integers $c, d, n_1, \dots, n_c, m_1, \dots, m_d$, a nonnegative integer k and a variable x with $u = G^k F G^{n_c} F G^{n_{c-1}} \dots F G^{n_1} F F(x)$, $v = G^k F G^{m_d} F G^{m_{d-1}} \dots F G^{m_1} F F(x)$ and $\langle f_{n_c} f_{n_{c-1}} \dots f_{n_2}(n_1), f_{m_d} f_{m_{d-1}} \dots f_{m_2}(m_1)\rangle \in r$. The proof can be completed similarly as the proof of 1.1. □

1.3. Theorem. *Let σ be a rich signature containing at most countably many operation symbols. A lattice is isomorphic to an interval in the lattice of equational theories of signature σ if and only if it is an algebraic lattice with at most countably many compact elements.*

Proof. It follows from 1.1 and 1.2. □

2. Zipper theorem

2.1. Lemma. *Let E be an equational theory. The lattice of all equational theories of the given signature that extend E is isomorphic to the congruence lattice of an algebra A of a signature containing a binary operation symbol G (and perhaps some other operation symbols) such that, with respect to G, A contains a left zero (an element o and a left unit u (i.e., $G(o, a) = 0$ and $G(u, a) = a$ for all $a \in A$).*

PROOF. Denote by $x_0, x_1, \dots$ the free generators of the algebra T of terms. $For t, u, v \in T$ denote by $t[u, v]$ the term $f(t)$ where f is the substitution with $f(x_0) = u$, $f(x_1) = v$ and $f(y) = y$ for all the other variables y. Let A be the algebra T with the following additional operations: all substitutions, considered as unary operations, and the binary operation $G(x, y) = x[x, y]$. Clearly, congruences of A are just the equational theories. The element x_0 is a zero element and the element x_1 is the unit of this algebra. The lattice of equational theories extending E is isomorphic to the congruence lattice of A/E. □

2.2. THEOREM. (Lampe [86]) *Let L be a lattice isomorphic to the lattice of all equational theories extending a given equational theory of some signature; denote by 1_L the largest element of L. If a, b and a_i ($i \in I$, I being a nonempty set) are elements of L such that $\bigvee\{a_i : i \in I\} = 1_L$ and $a_i \wedge a = b$ for all $i \in I$ then $a = b$.*

PROOF. According to 2.1, it is sufficient to assume that L is the congruence lattice of an algebra with a binary operation G, such that A contains a left zero 0 and a left unit u. We have $1_L = A^2$. Since $\langle o, u \rangle \in A^2 = \bigvee\{a_i : i \in I\}$, for some positive integer n there are elements $c_0, \dots, c_n$ of A and elements $i_1, \dots, i_n$ of I such that $c_0 = o$, $c_n = u$ and $\langle c_{j-1}, c_j \rangle \in a_{i_j}$ for $j = 1, \dots, n$. We are going to prove $a \subseteq b$. Let $\langle x, y \rangle \in a$. We are going to prove by induction on $j = 0, \dots, n$ that $\langle G(c_j, x), G(c_j, y) \rangle \in b$. For $j = 0$ we have $G(c_0, x) = o = G(c_0, y)$. Let $j > 0$. We have $\langle G(c_j, x), G(c_{j-1}, x) \rangle \in a_j$, $\langle G(c_{j-1}, x), G(c_{j-1}, y) \rangle \in b \subseteq a_j$ by the induction assumption and $\langle G(c_{j-1}, y), G(c_j, y) \rangle \in a_j$, so that $\langle G(c_j, x), G(c_j, y) \rangle \in a_j$; but this pair also belongs to a and so it belongs to $a_i \cap a = b$. We are done with the induction. In particular, $\langle G(c_n, x), G(c_n, y) \rangle \in b$, i.e., $\langle x, y \rangle \in b$. We have proved $a \subseteq b$ and we get $a = b$. □

2.3. COROLLARY. *The lattice $\mathbf{M}_5$ is not isomorphic to the lattice of all equational theories extending E, for any equational theory E.*

CHAPTER 16

MISCELLANEOUS

1. Clones: The Galois correspondence

Let a nonempty set A be given.We denote by $\mathcal{O}_A$ the set of all operations of arbitrary positive arities on A. For a positive integer n, the set of n-ary operations on A is denoted by $O_A^{(n)}$; for a set F of operations on A we put $F^{(n)} = F \cap O_A^{(n)}$.

The expression $a_i, \dots, a_j$ will be abbreviated as a_i^j.

For integers $n \geq 1$ and $i \in \{1, \dots, n\}$, the i-th n-ary projection on A is the operation $e_{n,i}$ defined by $e_{n,i}(a_1^n) = a_i$ for all $a_1^n \in A$. If f is an n-ary operation and $g_1, \dots, g_n$ are k-ary operations on A, we define a k-ary operation $f(g_1, \dots, g_n)$ on A, called the superposition of $f, g_1, \dots, g_n$, by $f(g_1, \dots, g_n)(a_1^k) = f(g_1(a_1^k), \dots, g_n(a_1^k))$. By a *clone* on A we mean a set of operations containing all the projections and closed under superposition.

The intersection of an arbitrary set of clones on A is again a clone. From this it follows that for an arbitrary subset F of $\mathcal{O}_A$ there exists a least clone containing F; it will be denoted by $[F]$; instead of $[\{f_1, f_2, \dots\}]$ we write simply $[f_1, f_2, \dots]$. Further, it follows that the set of clones on A is a complete lattice with respect to inclusion. Its greatest element is the clone $\mathcal{O}_A$, and its least element is the clone J_A of all projections on A. The clone J_A is called trivial, while any other clone is called nontrivial.

1.1. THEOREM. *The lattice of clones on A is an algebraic lattice; a clone C is a compact element of this lattice if and only if it is a finitely generated clone.*

PROOF. It is evident. □

We fix an infinite countable set of variables X=$\{x_1, x_2, \dots\}$. For every operation f on A we fix an operation symbol $\bar{f}$ of the same arity. We denote by W_A (or just W) the set of terms over X of the signature consisting of the symbols $\bar{f}$ with $f \in \mathcal{O}_A$. For a given set F of operations on A we denote by $W(F)$ the subset of W consisting of the terms t such that whenever $\bar{f}$ occurs in t then f belongs to F. By an at most n-ary term we mean a term containing no other variables than $x_1, \dots, x_n$. For an at most n-ary term t we define an n-ary operation $[t]_n$ as follows: if $t = x_i$ then $[t]_n = e_{n,i}$; if $t = \bar{f}(t_1^m)$ then $[t]_n(a_1^n) = f([t_1]_n(a_1^n), \dots, [t_m]_n(a_1^n))$.

1.2. THEOREM. *For a set F of operations on A, the clone generated by F consists exactly of the operations $[t]_n$ with n running over the positive integers and t running over the at most n-ary terms from $W(F)$.*

PROOF. It is easy. $\square$

We define three unary operations ζ, τ, Δ and one binary operation $\circ$ on $\mathcal{O}_A$ as follows:

for $f \in \mathcal{O}_A^{(1)}$, $\quad \zeta f = \tau f = \Delta f = f$;

for $f \in \mathcal{O}_A^{(n)}$ with $n \geq 2$, $\quad \zeta f = g \in \mathcal{O}_A^{(n)}$ where $g(a_1^n) = f(a_2^n, a_1)$;

for $f \in \mathcal{O}_A^{(n)}$ with $n \geq 2$, $\quad \tau f = g \in \mathcal{O}_A^{(n)}$ where $g(a_1^n) = f(a_2, a_1, a_3^n)$;

for $f \in \mathcal{O}_A^{(n)}$ with $n \geq 2$, $\quad \Delta f = g \in \mathcal{O}_A^{(n-1)}$ where $g(a_1^{n-1}) = f(a_1, a_1^{n-1})$;

for $f \in \mathcal{O}_A^{(n)}$ and $g \in \mathcal{O}_A^{(m)}$, $\quad f \circ g = h \in \mathcal{O}_A^{(n+m-1)}$ where $h(a_1^{n+m-1}) = f(g(a_1^m), a_{m+1}^{n+m-1})$.

1.3. THEOREM. *The following are equivalent for a set C of operations on A:*

(1) *C is a clone;*
(2) *C contains $e_{2,1}$ and is closed under ζ, τ, Δ and $\circ$;*
(3) *$[t]_n \in C$ for any $n \geq 1$ and any at most n-ary term $t \in W(C)$.*

PROOF. (1) $\rightarrow$ (2) and (3) $\rightarrow$ (1) are evident. Let (2) be satisfied. Since $e_{1,1} = \Delta e_{2,1}$, $e_{2,2} = \tau e_{2,1}$, $e_{n,i} = e_{2,1} \circ e_{n-1,i}$ if $i \leq n-1$ and $e_{n,n} = e_{2,2} \circ e_{n-1,1}$, C contains all projections. Let us prove by induction on the length of an n-ary term $t \in W(C)$ that $[t]_n$ belongs to C. If t is a variable then $[t]_n$ is a projection. Let $t = f(t_1^m)$. By induction, the operations $[t_1]_n, \dots, [t_m]_n$ belong to C; we have $f \in C$ and so the mn-ary operation $g(a_1^{mn}) = f([t_1]_n(a_1^n), \dots, [t_m]_n(a_{(m-1)n+1}^{mn}))$ belongs to C, since it can be expressed as $\zeta(\zeta(\dots(\zeta(\zeta f \circ [t_m]_n) \circ [t_{m-1}]_n)\dots) \circ [t_2]_n) \circ [t_1]_n$. Now, in order to prove that $[t]_n$ belongs to C, it is evidently enough to show that if h is a k-ary operation belonging to C and $1 \leq i < j \leq k$ then the operation $h'(a_1^{k-1}) = h(a_1^{j-1}, a_i, a_j^{k-1})$ belongs to C; but this is clear from $h' = \zeta^{i-1}\Delta(\tau\zeta)^{j-i}\zeta^{l-j+1}h$. $\square$

We denote by $\mathcal{R}_A$ the set of all relations on A; for $n \geq 1$, $\mathcal{R}_A^{(n)}$ denotes the set of n-ary relations on A.

Let f be an n-ary operation and r be an m-ary relation on A. We say that f preserves r if the following is true: if $(a_{1,1}^{1,m}) \in r, \dots, (a_{n,1}^{n,m}) \in r$ then $(f(a_{1,1}^{n,1}), \dots, f(a_{1,m}^{n,m})) \in r$. For a set R of relations on A we denote by $\mathcal{P}(R)$ the set of the operations preserving all the relations from R; this set is a clone and its elements are called polymorphisms of R. For a set F of operations on A we denote by $\mathcal{I}(F)$ the set of the relations preserved by all the operations from F; the elements of $\mathcal{I}(F)$ are called the invariants of F.

1.4. THEOREM. *The mappings $\mathcal{P}$ and $\mathcal{I}$ define a Galois correspondence between the subsets of $\mathcal{O}_A$ and the subsets of $\mathcal{R}_A$.*

PROOF. It is evident. □

For a clone C on A and a finite relation $r = \{(a_{1,1}^{k,1}), \dots, (a_{1,n}^{k,n})\}$ on A put

$$\Gamma_C(r) = \{(f(a_{1,1}^{1,n}), \dots, f(a_{k,1}^{k,n})); f \in C^{(n)}\}.$$

1.5. LEMMA. *Let r be a finite k-ary relation on A and C be a clone on A. Then $\Gamma_C(r)$ is the least invariant of C containing r.*

PROOF. It is easy. □

1.6. LEMMA. *Let f be an n-ary operation on A, C be a clone on A and U be a finite subset of A. Let $(a_{1,1}^{1,n}), \dots, (a_{k,1}^{k,n})$ be all the n-tuples of elements of U and denote by r the k-ary relation $\{(a_{1,1}^{k,1}), \dots, (a_{1,n}^{k,n})\}$. Then f preserves $\Gamma_C(r)$ if and only if it coincides with an operation from C on U.*

PROOF. It is easy. □

If the set A is finite, of cardinality k, then for any positive integer n we fix an ordering $(a_{1,1}^{1,n}), \dots, (a_{k^n,1}^{k^n,n})$ of all the n-tuples of elements of A and denote by χ_n the k^n-ary relation $\{(a_{1,1}^{k^n,1}), \dots, (a_{1,n}^{k^n,n})\}$.

1.7. LEMMA. *Let A be finite; let C be a clone on A and n be a positive integer. An n-ary operation belongs to C if and only if it preserves $\Gamma_C(\chi_n)$.*

PROOF. It follows from 1.6. □

Let C be a clone on A. We say that an operation $g \in \mathcal{O}_A^{(n)}$ can be interpolated by operations from C if for every finite subset S of A there exists an operation $f \in C$ such that $f|_S = g|_S$. The clone C is called locally closed if it contains every operation that can be interpolated by operations from C.

1.8. THEOREM. *The sets of the form $\mathcal{P}(R)$, for a set R of relations on A, are exactly the locally closed clones on A.*

PROOF. It is easy to see that $\mathcal{P}(R)$ is always a locally closed clone. Let C be a locally closed clone on A. It follows from 1.5 and 1.6 that an operation belongs to C if and only if it preserves all the relations $\Gamma_C(r)$, for r running over all finite relations on A. □

1.9. COROLLARY. *In the case when A is finite, the sets of operations closed in the Galois correspondence $\mathcal{P}$ - $\mathcal{I}$ are exactly the clones.*

For a positive integer n and an equivalence e on the set $\{1, \dots, n\}$ we denote by $\delta_{n,e}$ the n-ary relation on A defined by $(a_1, \dots, a_n) \in \delta_{n,e}$ if and only if $a_i = a_j$ for all $(i, j) \in e$. The relations obtained in this way are called diagonal.

For every relation r on A we fix a relation symbol $\bar{r}$ of the same arity. By a formula over A we mean a formula of the language consisting of the symbols $\bar{r}$ with $r \in \mathcal{R}_A$. By an at most n-ary formula we mean a formula whose every free variable belongs to $\{x_1, \dots, x_n\}$. For an at most n-ary formula f we define an n-ary relation $[f]_n$ on A as follows: $(a_1^n) \in [f]_n$ if and only if (a_1^n) satisfies f. By a $\{\&, \exists\}$-formula we mean a formula not containing $\neg, \vee, \rightarrow, \forall$. By a formula in R, for a set R of relations on A, we mean a formula f such that whenever $\bar{r}$ occurs in f then $r \in R$.

By a relation system on A we mean a set R of relations (of arbitrary arities) on A such that $[f]_n \in R$ for any positive integer n and any at most n-ary $\{\&, \exists\}$-formula f in R.

1.10. Theorem. *The set of relation systems on A is an algebraic lattice; its least element is the relation system of the diagonal relations and its greatest element is the relation system $\mathcal{R}_A$. For any set F of operations on A, the set $\mathcal{I}(F)$ is a relation system.*

Proof. It is easy. □

We define three unary operations ζ, τ, Δ and one binary operation $\circ$ on $\mathcal{R}_A$ as follows:

for $r \in \mathcal{R}_A^{(1)}$, $\zeta r = \tau r = \Delta r = r$;

for $r \in \mathcal{R}_A^{(n)}$ with $n \geq 2$, $\zeta r = s \in \mathcal{R}_A^{(n)}$ where $(a_1^n) \in s$ if and only if $(a_2^n, a_1) \in r$;

for $r \in \mathcal{R}_A^{(n)}$ with $n \geq 2$, $\tau r = s \in \mathcal{R}_A^{(n)}$ where $(a_1^n) \in s$ if and only if $(a_2, a_1, a_3^n) \in r$;

for $r \in \mathcal{R}_A^{(n)}$ with $n \geq 2$, $\Delta r = s \in \mathcal{R}_A^{(n-1)}$ where $(a_1^{n-1}) \in s$ if and only if $(a_1, a_1^{n-1}) \in r$;

for $r \in \mathcal{R}_A^{(n)}$ and $s \in \mathcal{R}_A^{(m)}$, $r \circ s = t \in \mathcal{R}_A^{(n+m-2)}$ where $(a_1^{n+m-2}) \in t$ if and only if there exists an u with $(a_1^{n-1}, u) \in r$ and $(u, a_n^{n+m-2}) \in s$; if $n = m = 1$, put $r \circ s = \emptyset$.

1.11. Theorem. *Let the set A be finite. Then a set R of relations on A is a relation system if and only if it is closed with respect to ζ, τ, Δ and contains the diagonal relation $\delta_{3,e}$ where e is the equivalence on $\{1, 2, 3\}$ identifying 2 with 3.*

Proof. Excercise. □

1.12. Lemma. *Let C be a clone on a finite set A. The relation system $\mathcal{I}(C)$ is generated by the relations $\Gamma_C(\chi_t)$ $(t = 1, 2, \dots)$.*

Proof. Let $r \in \mathcal{I}(C)$ be m-ary; denote the elements of r by $(a_{1,1}^{m,1}), \dots, (a_{1,t}^{m,t})$ and the elements of χ_t by $(b_{1,1}^{k^t,1}), \dots, (b_{1,t}^{k^t,t})$. For every $i \in \{1, \dots, m\}$ denote by $h(i)$ the number with $(a_{i,1}^{i,t}) = (b_{h(i),1}^{h(i),t})$. It is easy to prove that an m-tuple (a_1^m) belongs to r if and only if there exist elements $y_1, \dots, y_{k^t}$ such that $(y_1^{k^t}) \in \Gamma_C(\chi_t)$ and $a_1 = y_{h(1)}, \dots, a_m = y_{h(m)}$. Hence r belongs to the relation system generated by $\Gamma_C(\chi_t)$. □

1.13. Theorem. *Let Q be a set of relations on a finite set A. Then Q is a relation system if and only if $Q = \mathcal{I}(F)$ for some set of operations F.*

PROOF. Let Q be a relation system and put $F = \mathcal{P}(Q)$; we have to prove $Q = \mathcal{I}(F)$. By 1.12 it suffices to show $\Gamma_F(\chi_t) \in Q$ for any $t \geq 1$. Denote by γ the intersection of all k^t-ary relations from Q containing χ_t; we have $\gamma \in Q$. Evidently, $\Gamma_F(\chi_t) \subseteq \gamma$. It is enough to prove $\Gamma_F(\chi_t) = \gamma$. Suppose, on the contrary, that $\Gamma_F(\chi_t) \subset \gamma$; fix a sequence $(u_1^{k^t}) \in \gamma \smallsetminus \Gamma_F(\chi_t)$. Denote the elements of χ_t by $(b_{1,1}^{k^t,1}), \dots, (b_{1,t}^{k^t,t})$ and define a t-ary operation f by $f(b_{i,1}^{i,t}) = u_i$ for all $i \in \{1, \dots, k^t\}$. By 1.5 we have $f \notin F = \mathcal{P}(Q)$, so that there exists an $m \geq 1$ and a relation $\varrho \in Q^{(m)}$ which is not preserved by f. There exist $(a_{1,1}^{m,1}), \dots, (a_{1,t}^{m,t})$ in ϱ such that $(f(a_{1,1}^{1,t}), \dots, f(a_{m,1}^{m,t})) \notin \varrho$. For every $j \in \{1, \dots, m\}$ denote by $h(j)$ the number from $\{1, \dots, k^t\}$ with $(a_{j,1}^{j,t}) = (b_{h(j),1}^{h(j),t})$. Denote by e the least equivalence on $\{1, \dots, k^t + m\}$ such that $(h(j), k^t + j) \in e$ for any $j \in \{1, \dots, m\}$. Put $\varrho' = (\gamma \times \varrho) \cap \delta_{k^t+m,e}$. Define a k^t-ary relation ϱ'' by $(z_1^{k^t}) \in \varrho''$ if and only if there exist $y_1, \dots, y_m$ with $(z_1^{k^t}, y_1^m) \in \varrho'$. Evidently, $\varrho'' \in Q$. Since $(b_{1,1}^{k^t,1}, a_{1,1}^{m,1}) \in \varrho', \dots, (b_{1,t}^{k^t,t}, a_{1,t}^{m,t}) \in \varrho'$, we have $\chi_t \subseteq \varrho''$. Hence $\gamma \subseteq \varrho''$. From this we get $(u_1^{k^t}) \in \varrho''$, i.e., $(u_1^{k^t}, a_1^m) \in \varrho'$ for some $(a_1^m) \in \varrho$. We have $(a_1^m) = (u_{h(1)}^{h(m)}) = (f(b_{h(1),1}^{h(1),t}), \dots, f(b_{h(m),1}^{h(m),t})) = (f(a_{1,1}^{1,t}), \dots, f(a_{m,1}^{m,t}))$, a contradiction, since this sequence does not belong to ϱ. □

An operation $f \in \mathcal{O}_A^{(n)}$ is said to depend essentially on the i-th variable ($i \in \{1, \dots, n\}$) if there exist elements $a_1^n, b, c \in A$ such that $f(a_1^{i-1}, b, a_{i+1}^n) \neq f(a_1^{i-1}, c, a_{i+1}^n)$. An operation which depends on at most one variable is called essentially unary.

1.14. THEOREM. *The clone $\mathcal{O}_A^{(1)}$ generated by the unary operations on A consists exactly of the essentially unary operations. The lattice of subclones of this clone is isomorphic to the lattice of submonoids of the transformation monoid of A; the mapping $C \to C \cap \mathcal{O}_A^{(1)}$ establishes the isomorphism; inversely, the clone generated by a transformation monoid M consists of the operations f such that there exist an $i \in \{1, \dots, n\}$ (n being the arity of f) and a $g \in M$ with $f(a_1^n) = g(a_i)$ for all $a_1^n \in A$.*

Similarly, the lattice of subclones of the clone generated by the permutations of A is isomorphic to the lattice of subgroups of the permutation group of A.

PROOF. It is evident. □

1.15. LEMMA. *Let A be a finite set of at least three elements and $f \in \mathcal{O}_A^{(n)}$ be an operation depending on at least two variables and taking m dif and only iferent values, where $m \geq 3$. Then:*

1. *There are subsets $K_1, \dots, K_n$ of A of cardinalities ≤ 2 such that f restricted to $K_1 \times \dots \times K_n$ takes at least three dif and only iferent values.*
2. *There are subsets $K'_1, \dots, K'_n$ of A of cardinalities $\leq m - 1$ such that f restricted to $K'_1 \times \dots \times K'_n$ takes all the m values.*

Proof. (1) We can assume that f depends essentially on the first place. Suppose that there are no such sets $K_1, \dots, K_n$.

Let (a_1^n) and (b_1^n) be two n-tuples such that $a_i = b_i$ for all $i \neq 1$ and $f(a_1^n) \neq f(b_1^n)$. If, for some n-tuple (c_1^n), either $f(a_1, c_2^n)$ or $f(b_1, c_2^n)$ does not belong to $\{f(a_1^n), f(b_1^n)\}$, we can put $K_1 = \{a_1, b_1\}, K_2 = \{a_2, c_2\}, \dots, K_n = \{a_n, c_n\}$. So, we have proved that if $f(a_1^n) \neq f(b_1^n)$ and $a_i = b_i$ for all $i \neq 1$ then, for any n-tuple (c_1^n), the elements $f(a_1, c_2^n)$ and $f(b_1, c_2^n)$ both belong to $\{f(a_1^n), f(b_1^n)\}$.

Let $(a_1^n), (b_1^n)$ be two n-tuples such that $f(a_1^n) \neq f(b_1^n)$ and $a_i = b_i$ for $i \neq 1$. There exists an n-tuple (c_1^n) such that the elements $f(a_1^n), f(b_1^n), f(c_1^n)$ are pairwise distinct. The element $f(a_1, c_2^n)$ equals either $f(a_1^n)$ or $f(b_1^n)$. If it were equal to $f(b_1^n)$, we could put $K_i = \{a_i, c_i\}$ for all i, a contradiction. Hence $f(a_1, c_1^n) = f(a_1^n)$. Since $f(c_1^n) \neq f(a_1^n)$, we get $f(a_1, d_2^n) \in \{f(a_1^n), f(b_1^n)\}$ for any d_2^n; however, we have $f(a_1, d_2^n) \in \{f(a_1^n), f(b_1^n)\}$ and thus $f(a_1, d_2^n) = f(a_1^n)$ for any d_2^n. We have proved that if $a_i = b_i$ for all $i \neq 1$ and $f(a_1^n) \neq f(b_1^n)$ then $f(a_1, d_2^n) = f(a_1^n)$ and $f(b_1, d_2^n) = f(b_1^n)$ for all $d_2^n \in A$. From this it follows that f depends on the first variable only, a contradiction.

(2) Let $K_1, \dots, K_n$ be as in (1); let a, b, c be three dif and only iferent values of f on $K_1 \times \dots \times K_n$; let $f(a_{1,1}^{1,n}), \dots, f(a_{m-3,1}^{m-3,n})$ be the remaining values of f. We can put $K_i' = K_i \cup \{a_{1,i}, \dots, a_{m-3,i}\}$. □

Define two relations π_4 and ν on A as follows:

π_4 is the set of the quadruples (a, b, c, d) such that either $a = b$ or $c = d$.

If A has at least three elements, put $\nu = \{(a, b); a \neq b\}$; if A has only two elements, put $\nu = \{(a, b, c, d, e, f); (a, b, c, d) \in \pi_4 \& e \neq f\}$.

1.16. Theorem. *The clone of essentially unary operations is equal to $\mathcal{P}(\pi_4)$. The clone generated by the permutations of A is equal to $\mathcal{P}(\nu)$.*

Proof. Denote the first clone by C and the second by D. Clearly, $C \subseteq \mathcal{P}(\pi_4)$. Let $f \in \mathcal{P}(\pi_4)$ be n-ary and suppose that f depends essentially on two dif and only iferent variables, the i-th and the j-th. We have $f(a_1^n) \neq f(b_1^n)$ and $f(c_1^n) \neq f(d_1^n)$ for four n-tuples such that if $k \neq i$ then $a_k = b_k$ and if $k \neq j$ then $c_k = d_k$. For every $k \in \{1, \dots, n\}$ we have $(a_k, b_k, c_k, d_k) \in \pi_4$ and so $(f(a_1^n), f(b_1^n), f(c_1^n), f(d_1^n)) \in \pi_4$, a contradiction.

Let A contain at least three elements. Clearly, $D \subseteq \mathcal{P}(\nu)$. Let $f \in \mathcal{P}(\nu)$ be n-ary. Evidently, it is sufficient to prove $f \in C$. Suppose, on the contrary, that f depends essentially on two dif and only iferent variables. Denote by k the cardinality of A. Evidently, f takes k dif and only iferent values on the n-tuples $(c, \dots, c)$ with $c \in A$. By 1.15 there exists an n-tuple (c_1^n) such that f takes all the k values on the set $M = \{(a_1^n); a_i \neq c_i \text{ for all } i\}$. There is an n-tuple $(a_1^n) \in M$ with $f(a_1^n) = f(c_1^n)$. We have $(a_1, c_1) \in \nu, \dots, (a_n, c_n) \in \nu$ but $(f(a_1^n), f(c_1^n)) \notin \nu$, a contradiction. □

We define a binary operation $\vee$ and a unary operation $\neg$ on $\mathcal{R}_A$ as follows:

If $r \in \mathcal{R}_A^{(n)}$ and $s \in \mathcal{R}_A^{(m)}$ are nonempty then $r \vee s \in \mathcal{R}_A^{(N)}$ where $N = \max(n,m)$ and $(a_1^N) \in r \vee s$ if and only if either $(a_1^n) \in r$ or $(a_1^m) \in s$; put $r \vee \emptyset = \emptyset \vee r = r$ for all r.

If $r \in \mathcal{R}_A^{(n)}$ is nonempty then let $\neg r$ be the complement of r in $\mathcal{R}_A^{(n)}$; put $\neg\emptyset = \emptyset$.

1.17. THEOREM. *Let R be a set of relations on a finite set A. Then $R = \mathcal{I}(F)$ for a set of unary operations F if and only if R is a relation system closed for $\vee$ if and only if $[f]_n \in R$ for any $n \geq 1$ and any $\{\&, \vee, \exists\}$-formula f in R. Further, $R = \mathcal{I}(F)$ for a set of permutations F if and only if R is a relation system closed for $\vee$ and $\neg$ if and only if $[f]_n \in R$ for any $n \geq 1$ and any formula f in R.*

PROOF. It is easy. □

1.18. THEOREM. *Let $k = \mathbf{card}(A)$ be finite and let C be a clone on A generated by its n-ary operations. Then there are at most $2^{k^{k^n}}$ clones covered by C in the lattice of clones on A.*

PROOF. Let H be a clone covered by C and let C be generated by n-ary operations $f_1, \dots, f_s$. If it were $f_1, \dots, f_s \in \mathcal{P}\Gamma_H(\chi_n)$ then by 1.7 we would have $f_1, \dots, f_s \in H$, so that $C \subseteq H$, a contradiction. Hence $H \subseteq C \cap \mathcal{P}\Gamma_C(\chi_n) \subset C$; since H is covered by C, we get $H = C \cap \mathcal{P}\Gamma_C(\chi_n)$. From this it follows that H is uniquely determined by a k^n-ary relation. The number of k^n-ary relations on A is $2^{k^{k^n}}$. □

1.19. COROLLARY. *The following are equivalent for a clone C on a finite set A:*

(1) *C is finitely generated;*
(2) *C is not the union of an increasing infinite chain of its subclones;*
(3) *the lattice of subclones of C is coatomic and contains a finite number of coatoms only.* □

1.20. THEOREM. *Let R be a relation system on a finite set A of cardinality k. If R is finitely generated then R is one-generated. If R is generated by a relation r containing n tuples then there are at most k^{k^n} relation systems covered by R in the lattice of relation systems on A.*

PROOF. If R is generated by $r_1, \dots, r_s$ and $r_1, \dots, r_s$ are nonempty then R is generated by the relation $r_1 \times \dots \times r_s$. Let R be generated by r and denote by n the number of the tuples in r; let $n \geq 1$. Let S be a relation system covered by R. We have $\mathcal{P}^{(n)}R \subseteq \mathcal{P}^{(n)}S$. In the case of equality 1.5 would yield $\Gamma_{\mathcal{P}S}(r) = \Gamma_{\mathcal{P}R}(r) = r$, so that $r \in \mathcal{IP}(S) = S$ and consequently $R \subseteq S$, a contradiction. Hence we can take an operation $f \in \mathcal{P}^{(n)}(S) \smallsetminus \mathcal{P}^{(n)}(R)$. We have $S \subseteq R \cap \mathcal{I}(f_S) \subset R$ and so $S = R \cap \mathcal{I}(f_S)$, since S is covered by R. Hence S is uniquely determined by an n-ary operation. The number of n-ary operations on A is k^{k^n}. □

1.21. Corollary. *The following are equivalent for a relation system R on a finite set A:*

(1) *R is finitely generated;*
(2) *R is not the union of an increasing infinite chain of its relation subsystems;*
(3) *the lattice of relation subsystems of C is coatomic and contains a finite number of coatoms only.* □

1.22. Theorem. *The clone of all operations on a finite set A is finitely generated. The following are some examples of finite generating systems of operations for $\mathcal{O}_{\{1,\dots,k\}}$:*

(1) *$\{\min, \max, c_1^k, j_1^k\}$ where $\min$ and $\max$ are binary and c_i, j_i are unary operations defined by $c_i(x) = i$ for all x, $j_i(x) = k$ for $x = i$ and $j_i(x) = 1$ for $x \neq i$;*
(2) *$\{\min, g\}$ where g is the unary operation defined by $g(1) = 2, g(2) = 3, \dots, g(k) = 1$;*
(3) *$\{h\}$ where h is the binary operation defined by $h(x, y) = 1$ for $x = y = k$ and $h(x, y) = \min(x, y) + 1$ otherwise;*
(4) *$\{f, g, c_1, \dots, c_k, d_1, \dots, d_k\}$ where, for some pair e, o of distinct elements of A, f, g are two binary operations satisfying $f(x, e) = x$, $f(x, o) = o$, $g(o, x) = g(x, o) = x$, c_i are the constants and d_i are the unary operations defined by $d_i(a_i) = e$ and $d_i(a_j) = o$ for $i \neq j$;*
(5) *$\{f, h\}$ where, for an element $o \in A$, f is an arbitrary binary operation such that a restriction of f is a group operation on $A \setminus \{o\}$, $f(x, o) = f(o, x) = o$ for all x and h is a cyclic permutation of A.*

Proof. (1) We shall prove by induction on n that an arbitrary n-ary operation h on A belongs to the clone generated by $\{\min, \max, c_1^k, j_1^k\}$. For $n = 1$ we have

$$f(x) = \max(\min(j_1(x), c_{f(1)}(x)), \dots, \min(j_k(x), c_{f(k)}(x))).$$

For $n \geq 2$ we have

$$f(x_1^n) = \max(\min(j_1(x_n), f(x_1^{n-1}, 1)), \dots, \min(j_k(x_n), f(x_1^{n-1}, k))).$$

(2) We have $c_1(x) = \min(g(x), g^2(x), \dots, g^k(x))$, $c_i(x) = g^i(c_1(x))$, $j_i(x) = g^{k-1}(\min(c_2(x), g^{k-i}(x)))$. Put $f_{s,i}(x) = g^{k-1}(\min(j_i(x), c_{g(s)}(x)))$, so that $f_{s,i}(x) = s$ if $x = i$ and $f_{s,i}(x) = k$ if $x \neq i$. Put $h(x) = \min(f_{k,1}(x), f_{k-1,2}(x), \dots, f_{1,k}(x)) = k - 1 - x$. We have $\max(x, y) = h(\min(h(x), h(y)))$ and we can use (1).

(3) This follows from (2), since $g(x) = h(x, x)$ and $\min(x, y) = g^{k-1}h(x, y)$.

(4) Write $\wedge$ instead of f and $\vee$ instead of g. If h is an n-ary operation then

$$h(x_1^n) = \bigvee_{(a_1^n) \in A^n} (c_{h(a_1^n)}(x_1) \wedge d_{a_1}(x_1) \wedge \dots \wedge d_{a_n}(x_n))$$

with the bracketings and with the order of the n-tuples (a_1^n) arbitrary.

(5) For $i = 0, \dots, k-1$ put $a_i = h^i(o)$, so that $A = \{a_0^{k-1}\}$ and $a_0 = o$. Denote by s the number from $\{1, \dots, k-1\}$ such that a_s is the unit of the group $(A \setminus \{o\}, f)$; put $e = a_s$. Define c_i and d_i as in (4) and write xy instead of $f(x, y)$. We have $c_o(x) = xh(x) \dots h^{k-1}(x)$, $c_{a_i}(x) = h^i(c_o(x))$, $d_{a_i}(x) = h^s(h^{k-s}(c_o(x))(h^{k-i}(x))^{k-1})$. Put $g(x, y) = h^{k-s}(h^s(x)h^s(y))$. We can apply (4). □

1.23. Theorem. *Let A be a finite set with k elements, $k \geq 2$. The relation system $\mathcal{R}_A$ is finitely generated; if $k = 2$ then it is generated by the set of its ternary relations; if $k \geq 3$ then it is generated by the two binary relations ϱ and ν where $(x, y) \in \varrho$ if and only if $x \leq y$ and $(x, y) \in \nu$ if and only if $x \neq y$.*

Proof. Let $k \geq 3$ and let f be an n-ary operation preserving both ϱ and ν; it is enough to prove that f is a projection. By 1.16, there exist an i and a permutation g of A such that $f(x_1^n) = g(x_i)$. Since f preserves ϱ, we get $g = e_{1,1}$. □

1.24. Theorem. *Let A be a finite set of cardinality $k \geq 3$. Then the lattice of clones on A is uncountable; it contains a subposet isomorphic to the lattice of all subsets of an infinite countable set.*

Proof. Let a, b, c be three distinct elements of A. Put $N = \{2, 3, 4, \dots\}$. For every $n \in N$ define an n-ary operation g_n on A as follows: $g_n(x_1^n) = a$ if there exists an i such that $x_i = a$ and $x_1 = \dots x_{i-1} = x_{i+1} = \dots x_n = b$; $g_n(x_1^n) = c$ in all other cases. For every $I \subseteq N$ denote by C_I the clone generated by $\{g_n; n \in I\}$. It is easy to prove that $I \subseteq I'$ if and only if $C_I \subseteq C_{I'}$. □

2. Categorical embeddings

A relation r on a set A is said to be rigid if there is no mapping f of A into itself, except $\mathbf{id}_A$, such that whenever $\langle x, y \rangle \in r$ then $\langle f(x), f(y) \rangle \in r$.

2.1. Theorem. (Vopěnka, Pultr, Hedrlín [65]) *For every set A there exists a rigid, antireflexive relation r on A such that r is contained in a well ordering of A.*

Proof. If A is finite then we can take $r = s \setminus \mathbf{id}_a$ where s is any well ordering of A.

Let A be infinite. Put $k = \mathbf{card}(A)$ and denote by D the set of all ordinal numbers that are less or equal $k+1$. Since $\mathbf{card}(A) = \mathbf{card}(D)$, it is sufficient to prove that a rigid relation r, satisfying the above formulated requirements, exists on D.

Denote by D_0 the set of all the limit ordinal numbers from D that are the union of a countable set of smaller ordinal numbers; denote by D_1 the set of all the other limit ordinal numbers from D; and denote by D_2 the set of the non-limit ordinal numbers from D. (We have $0 \in D_1$.) For every $a \in D_0$ there exists, and we will fix one, non-decreasing sequence $a'_2, a'_3, a'_4, \dots$ such that $a = (a'_2 + 2) \cup (a'_3 + 3) \cup (a'_4 + 4) \cup \dots$; put $a_i = a'_i + i$, so that a is the union of the increasing sequence $a_2, a_3, a_4, \dots$. Define a relation r on D in this way:

(1) $\langle 0, 2\rangle \in r$
(2) $\langle a, a+1\rangle \in r$ for all $a \in D \smallsetminus \{k+1\}$
(3) if $b \in D_1$ then $\langle a, b\rangle \in r$ if and only if $a \in b$ and a is a limit ordinal number
(4) if $a \in D_0$ then $\langle c, a\rangle \in r$ if and only if $c = a_n$ for some $n \geq 2$
(5) $\langle a, k+1\rangle \in r$ if and only if either $a = k$ or $a \in D_2$ and $a \neq k+1$

It remains to show that r is rigid; the other requirements are clearly satisfied. Let f be a mapping of D into D such that $\langle a, b\rangle \in r$ implies $\langle f(a), f(b)\rangle \in r$.

Claim 1. If $a, b \in D$ and $a \in b$ then $f(a) \in f(d)$. Suppose that this is not true and take the least ordinal number $b \in D$ for which there exists an $a \in b$ with $f(a) \geq f(b)$. If $b \in D_1$ then, where c is the union of all $a+n$ for $n \in \omega$, c is a limit ordinal number such that $a \in c \in b$; by the minimality of b we have $f(a) \in f(c)$; by (3) we have $\langle c, b\rangle \in r$, so that $\langle f(c), f(b)\rangle \in r$ and thus $f(a) \in f(c) \in f(b)$, hence $f(a) \in f(b)$, a contradiction. If $b \in D_0$ then $a \in b_n \in b$ for some $n \geq 2$; by the minimality of b we have $f(a) \in f(b_n)$; by (4) we have $\langle b_n, b\rangle \in r$, so that $\langle f(b_n), f(b)\rangle \in r$ and hence $f(a) \in f(b_n) \in f(b)$, a contradiction. Finally, if $b \in D_2$ then there exists a c such that $b = c+1$, and we have $a \leq c \in b$; then $f(a) \leq f(c)$ by the minimality of b; by (2) we have $\langle c, b\rangle \in r$, so that $\langle f(c), f(b)\rangle \in r$ and thus $f(a) \leq f(c) \in f(b)$, a contradiction again.

Claim 2. $a \leq f(a)$ for all $a \in D$. Suppose that this is not true and let a be the least ordinal number from D such that $f(a) \in a$. By Claim 1 we have $f(f(a)) \in f(a)$, a contradiction with the minimality of a.

In particular, $f(k+1) = k+1$ and $f(k) = f(k)$.

Claim 3. $f(a) \in D_2$ for all $a \in D_2$. For $a = k+1$ we have it, so let $a \neq k+1$. Then $\langle a, k+1\rangle \in r$ and hence $\langle f(a), k+1\rangle \in r$. Now $f(a) \in D_2$ follows from (5).

Claim 4. $f(n) = n$ for all natural numbers n. Since f is injective, we have $f(2) \neq k+1$. If $f(2) \neq 2$ then it follows from the definition of r and from $\langle f(0), f(2)\rangle \in r$ and $\langle f(1), f(2)\rangle \in r$ that $f(0)+1 = f(2) = f(1)+1$, so that $f(0) = f(1)$, which is not possible. Hence $f(2) = 2$. Then Also $f(0) = 0$ and $f(1) = 1$. Let $f(n) = n$ where $n > 1$. We have $\langle n, n+1\rangle \in r$, so that $\langle f(n), f(n+1)\rangle \in r$. Since $f(n+1) \neq k+1$ and $2 \in f(n+1)$, by Claim 3 and from the definition of r we get $f(n+1) = f(n)+1 = n+1$.

Claim 5. If $a+n \in D$ where n is a natural number then $f(a+n) = f(a)+n$. For a finite it follows from Claim 4. For $a = k$ and for $a = k+1$ it is clear. Let a be infinite and $a < k$. Clearly, it is sufficient to prove $f(a+1) = f(a)+1$. We have $\langle a, a+1\rangle \in r$ and hence $\langle f(a), f(a+1) \in r$. By Claim 3 we have $f(a+1) \in D_2$; since $f(a+1)$ is neither 2 nor $k+1$, it follows from the definition of r that $f(a+1) = f(a)+1$.

Claim 6. If $a \in D$ is a limit ordinal number then $f(a)$ is also a limit ordinal number. This is clear for $a = 0$. Let $a \neq 0$. Clearly, either $\langle 0, a\rangle \in r$ or there exist infinitely many ordinal numbers c such that $\langle c, a\rangle \in r$. So, either $\langle 0, f(a)\rangle \in r$ (and certainly $f(a) \neq 1$ and $f(a) \neq 2$) or there exist infinitely many ordinal numbers d such that $\langle d, f(a)\rangle \in r$; in both cases it is clear that $f(a)$ is a limit ordinal number.

Claim 7. If $a \in D_0$ and $b = f(a)$ then $b \in D_0$ and $b_n = f(a_n)$ for all natural numbers $n \geq 2$. Since $\langle a_n, a\rangle \in r$, we have $\langle f(a_n), f(a)\rangle \in r$; by Claim 5 we have $\langle f(a'_n) + n, b\rangle \in r$.If $b \in D_1$ then $f(a'_n) + n$ is a limit ordinal number, a contradiction. By Claim 6 we get $b \in D_0$. Let $n \geq 2$ be a natural number. Since $\langle f(a'_n) + n, b\rangle \in r$, there exists a natural number $k \geq 2$ such that $f(a'_n)+n = b'_k+k$. Since $f(a'_n)$ and b'_k are limit ordinal numbers, we have $n = k$ and hence $f(a_n) = b_n$.

Suppose $f(a) \neq a$ for some $a \in D$. Then $a \in f(a)$. By Claim 1 we have $f^n(a) \in f^{n+1}(a)$ for all natural numbers n. Let b be the union of all the ordinal numbers $f^n(a)$, so that $b \in D_0$. Suppose $b \in f(b)$. Put $c = f(b)$, so that $c \in D_0$. There exists a natural number $n \geq 2$ such that $b \in c_n \in c$. Also, there exists a natural number $i \geq 2$ such that $b_n \in f^i(a) \in b$. We have $c_n = f(b_n) \in f(f^i(a)) = f^{i+1}(a) \in b$, a contradiction. Hence $f(b) = b$. By Claim 7 we get $f(b_n) = b_n$ for all $n \geq 2$. There exists an $n \geq 2$ such that $a \in b_n \in b$. Then $f(a) \in f(b_n) \in b$; from this we get $f^2(a) \in f(b_n) = b_n \in b$, and so on. Hence $b \leq b_n \in b$, a contradiction. □

2.2. Lemma. *For every nonempty set A there exist two unary operations f, g on A such that the algebra with these two operations has no nonidentical endomorphism.*

Proof. If $A = \{a_1, \dots, a_n\}$ is finite, one can put $f(a_i) = a_{i+1}$ for $i < n$, $f(a_n) = a_n$, $g(a_1) = a_1$ and $g(a_i) = a_{i-1}$ for $i > 1$. Let A be infinite. By 2.2 there exists a rigid relation r on A. Denote by B the disjoint union $A \cup r \cup \{0, 1\}$ and define two binary operations f, g on B in this way:

(1) for $a \in A$ put $f(a) = 0$ and $g(a) = 1$
(2) for $\langle a, b\rangle \in r$ put $f(\langle a, b\rangle) = a$ and $g(\langle a, b\rangle) = b$
(3) put $f(0) = f(1) = 1$ and $g(0) = g(1) = 0$

Consider B as an algebra with respect to these two unary operations. Since $\mathbf{card}(A) = \mathbf{card}(B)$, it is sufficient to prove that the algebra B has no nonidentical endomorphism. Let h be an endomorphism of B. Since $g(h(0)) = h(g(0)) = h(0)$, we have $h(0) = 0$. Since $f(h(1)) = h(f(1)) = h(1)$, we have $h(1) = 1$. If $a \in A$ then $f(h(a)) = h(f(a)) = h(0) = 0$, so that $h(a) \in A$. If $\langle a, b\rangle \in r$ then $f(h(\langle a, b\rangle)) = h(f(\langle a, b\rangle)) = h(a) \in A$ and $g(h(\langle a, b\rangle)) = h(g(\langle a, b\rangle)) = h(b)$, so that $\langle h(a), h(b)\rangle \in r$ and $h(\langle a, b\rangle) = \langle h(a), h(b)\rangle$. Since r is rigid, it follows that h restricted to A is the identity. Then also $h(\langle a, b\rangle) = \langle h(a), h(b)\rangle = \langle a, b\rangle$ for all $\langle a, b\rangle \in r$. We have proved that $h(x) = x$ for all $x \in B$. □

Let K be a class of algebras of a signature σ and L be a class of algebras of a signature τ. We can consider K and L as categories. A functor J of K into L is said to be a categorical embedding if it is injective on objects (i.e., $J(A) = J(B)$ implies $A = B$ for $A, B \in K$) and for every homomorphism

g of $J(A)$ into $J(B)$ (where $A, B \in K$) there exists precisely one homomorphism f of A into B with $J(\langle f, A, B\rangle) = g$ (more precisely, we should write $J(\langle f, A, B\rangle) = \langle g, J(A), J(B)\rangle$; if there is no confusion, we will write $J(f)$ instead of $J(\langle f, A, B\rangle)$.

Clearly, if there are a categorical embedding of K into L and a categorical embedding of L into M then there is also a categorical embedding of K into M.

2.3. LEMMA. *Let σ, τ be two signatures such that there is an injective mapping z of $\mathbf{Dom}(\sigma)$ into $\mathbf{Dom}(\tau)$ with $\tau(z(F)) \geq \sigma(F)$ for all $F \in \mathbf{Dom}(\sigma)$ and such that if σ is without constants then τ is without constants. Then there is a categorical embedding of the class of all σ-algebras into the class of all τ-algebras.*

PROOF. For every σ-algebra A let $J(A)$ be the τ-algebra with the underlying set A defined in this way: if G is an m-ary operation symbol of τ and $G = z(F)$ for some n-ary operation symbol F of σ then for $a_1, \dots, a_m \in A$ put $G_{J(A)}(a_1, \dots, a_m) = F_A(a_1, \dots, a_n)$; if G is not in the range of z and $m > 0$, put $G_{J(A)}(a_1, \dots, a_m) = a_1$; if a constant G of τ is not in the range of z, put $G_{J(A)} = F_A$ where F is a fixed constant of σ. For a homomorphism f of A into B, where $A, B \in K$, put $J(f) = f$. □

2.4. LEMMA. *For every signature σ there exists a signature τ containing only unary operation symbols, such that there is a categorical embedding of the class of all σ-algebras into the class of all τ-algebras.*

PROOF. It follows from 2.3 that it is sufficient to consider the case when σ is nonempty and contains no constants. Denote by τ the signature with domain $\{\langle F, i\rangle : F \in \mathbf{Dom}(\sigma),\ 0 \leq i \leq \sigma(F)\}$, where each symbol has arity 1. For every σ-algebra A define a τ-algebra $J(A)$ in this way: the underlying set of $J(A)$ is the union of $A \cup \{u, v\}$ (u, v are two different elements not belonging to A) with the set of all finite sequences $\langle F, a_1, \dots, a_{\sigma(F)}\rangle$ where $F \in \mathbf{Dom}(\sigma)$ and $a_i \in A$ for all i; the unary operations are defined by

(1) $\langle F, 0\rangle_{J(A)}(a) = u$ for $a \in A$
(2) $\langle F, 0\rangle_{J(A)}(u) = \langle F, 0\rangle_{J(A)}(v) = v$
(3) $\langle F, 0\rangle_{J(A)}(\langle F, a_1, \dots, a_n\rangle) = F_A(a_1, \dots, a_n)$
(4) $\langle F, 0\rangle_{J(A)}(\langle G, a_1, \dots, a_n\rangle) = u$ for $G \neq F$
(5) $\langle F, i\rangle_{J(A)}(a) = v$ for $i \geq 1$ and $a \in A$
(6) $\langle F, i\rangle_{J(A)}(u) = \langle F, i\rangle_{J(A)}(v) = u$ for $i \geq 1$
(7) $\langle F, i\rangle_{J(A)}(\langle F, a_1, \dots, a_n\rangle) = a_i$ for $i \geq 1$
(8) $\langle F, i\rangle_{J(A)}(\langle G, a_1, \dots, a_n\rangle) = a_1$ for $i \geq 1$ and $G \neq F$

For every mapping f of a σ-algebra A into a σ-algebra B define a mapping $J(f)$ of $J(A)$ into $J(B)$ in this way:

(1) $J(f)(a) = f(a)$ for $a \in A$
(2) $J(f)(\langle F, a_1, \dots, a_n\rangle) = \langle F, f(a_1), \dots, f(a_n)\rangle$
(3) $J(f)(u) = u$ and $J(f)(v) = v$

One can easily check that if f is a homomorphism of A into B then $J(f)$ is a homomorphism of $J(A)$ into $J(B)$. It remains to show that for every homomorphism g of $J(A)$ into $J(B)$ there exists a homomorphism f of A into B with $g = J(f)$. Let us fix a symbol S of σ. Since $g(v) = g(\langle S, 0\rangle_{J(A)}(v)) = \langle S, 0\rangle_{J(B)}(g(v))$, we have $g(v) = v$. Since $g(u) = g(\langle S, 1\rangle_{J(A)}(u)) = \langle S, 1\rangle_{J(B)}(g(u))$, we have $g(u) = u$. Let $a \in A$. Since $\langle S, 1\rangle_{J(B)}(g(a)) = g(\langle S, 1\rangle_{J(A)}(a)) = g(v) = v$, we have $g(a) \in B$. Denote by f the restriction of g to A, so that f is a mapping of A into B.

We are going to prove that $g = J(f)$. Let F be an n-ary operation symbol of σ. Since $\langle F, 0\rangle_{J(B)}(g(\langle F, a_1, \dots, a_n\rangle)) = g(\langle F, 0\rangle_{J(A)}(\langle F, a_1, \dots, a_n\rangle)) = g(F_A(a_1, \dots, a_n)) \in B$, we have $g(\langle F, a_1, \dots, a_n\rangle) = \langle F, b_1, \dots, b_n\rangle$ for some $b_1, \dots, b_n \in B$. For $i = 1, \dots, n$ we have $b_i = \langle F, i\rangle_{J(B)}(\langle F, b_1, \dots, b_n\rangle) = \langle F, i\rangle_{J(B)}(g(\langle F, a_1, \dots, a_n\rangle)) = g(\langle F, i\rangle_{J(A)}(\langle F, a_1, \dots, a_n\rangle)) = g(a_i) = f(a_i)$. Hence $g(\langle F, a_1, \dots, a_n\rangle) = \langle F, f(a_1), \dots, f(a_n)\rangle$.

It remains to prove that f is a homomorphism of A into B. Let F be an n-ary operation symbol of σ and let $a_1, \dots, a_n \in A$. We have $f(F_A(a_1, \dots, a_n)) = g(F_A(a_1, \dots, a_n)) = g(\langle F, 0\rangle_{J(A)}(\langle F, a_1, \dots, a_n\rangle)) = \langle F, 0\rangle_{J(B)}(g(\langle F, a_1, \dots, a_n\rangle)) = \langle F, 0\rangle_{J(B)}(\langle F, f(a_1), \dots, f(a_n)\rangle) = F_B(f(a_1), \dots, f(a_n))$. □

2.5. Lemma. *Let σ be a signature containing only unary operation symbols and let τ be the signature containing just one ternary operation symbol S. There exists a categorical embedding of the class of all σ-algebras into the class of all τ-algebras.*

Proof. Put $T = \mathbf{Dom}(\sigma)$. By 2.2 there exist two mappings p, q of T into T such that whenever f is a mapping of T into T satisfying $fp = pf$ and $fq = qf$ then $f = \mathbf{id}_T$. For every σ-algebra A define a τ-algebra $J(A)$ in this way: its underlying set is the disjoint union $A \cup T \cup (A \times T) \cup \{u, v\}$ where u, v are two distinct elements;

(1) if $x, y, z \in J(A)$ and either $x = y = z = u$ or at most one of x, y, z equals u then $S_{J(A)}(x, y, z) = v$, with the exception $S_{J(A)}(v, v, v) = u$; moreover, $S_{J(A)}(v, u, u) = S_{J(A)}(u, v, u) = S_{J(A)}(u, u, v) = v$
(2) for $a \in A$ and $F \in T$ put $S_{J(A)}(\langle a, F\rangle, u, u) = \langle F_A(a), F\rangle$, $S_{J(A)}(u, \langle a, F\rangle, u) = F$, $S_{J(A)}(u, u, \langle a, F\rangle) = a$
(3) for $a \in A$ put $S_{J(A)}(a, u, u) = S_{J(A)}(u, a, u) = S_{J(A)}(u, u, a) = u$
(4) for $F \in T$ put $S_{J(A)}(F, u, u) = p(F)$, $S_{J(A)}(u, F, u) = q(F)$, $S_{J(A)}(u, u, F) = v$

For every mapping f of a σ-algebra A into a σ-algebra B define a mapping $J(f)$ of $J(A)$ into $J(B)$ by

(1) $J(f)(a) = f(a)$ for $a \in A$
(2) $J(f)(\langle a, F\rangle) = \langle f(a), F\rangle$ for $a \in A$ and $F \in T$
(3) $J(f)(x) = x$ for $x \in T \cup \{u, v\}$

It is easy to check that if f is a homomorphism of A into B then $J(f)$ is a homomorphism of $J(A)$ into $J(B)$. Clearly, we will be done if we show that for any two σ-algebras A and B and any homomorphism g of $J(A)$ into $J(B)$ there exists a homomorphism f of A into B with $g = J(f)$.

We have $g(u) = g(S_{J(A)}(v,v,v)) = S_{J(B)}(g(v),g(v),g(v))$ which is either u or v. Also, $g(v) = g(S_{J(A)}(u,u,u)) = S_{J(B)}(g(u),g(u),g(u))$ which is either v or u according to whether $g(u) = u$ or $g(v) = v$. If $g(u) = v$ then $g(v) = g(S_{J(A)}(u,v,v)) = S_{J(B)}(g(u),g(v),g(v)) = S_{J(B)}(v,u,u) = v$, a contradiction. Hence $g(u) = u$ and $g(v) = v$.

Let $a \in A$. We have $u = g(u) = g(S_{J(A)}(a,u,u)) = S_{J(B)}(g(a),g(u),g(u)) = S_{J(B)}(g(a),u,u)$ which is possible only if $g(a) \in B$. Denote by f the restriction of g to A, so that f is a mapping of A into B.

Let $a \in A$ and $F \in T$. We have $S_{J(B)}(u,u,g(\langle a,F\rangle)) = S_{J(B)}(g(u),g(u), g(\langle a,F\rangle)) = g(S_{J(A)}(u,u,\langle a,F\rangle)) = g(a) = f(a) \in B$ which is possible only if $g(\langle a,F\rangle) = \langle b,G\rangle$ for some $b \in B$ and $G \in T$.

Let us choose an element $a \in A$. If $F \in T$ then $g(F) = g(S_{J(A)}(u,\langle a,F\rangle,u)) = S_{J(B)}(g(u),g(\langle a,F\rangle),g(u)) = S_{J(B)}(u,g(\langle a,F\rangle),u) \in T$. Denote by h the restriction of g to T, so that h is a mapping of T into itself. For $F \in T$ we have $h(p(F)) = g(S_{J(A)}(F,u,u)) = S_{J(B)}(g(F),g(u),g(u)) = S_{J(B)}(h(F),u,u) = p(h(F))$ and similarly $h(q(F)) = g(q(F)) = g(S_{J(A)}(u,F,u)) = S_{J(B)}(g(u), g(F),g(u)) = S_{J(B)}(u,h(F),u) = q(h(F))$; by the choice of p,q we get $h = \mathbf{id}_T$. Hence $g(F) = F$ for all $F \in T$.

Let $a \in A$ and $F \in T$, so that $g(\langle a,F\rangle) = \langle b,G\rangle$ for some $b \in B$ and $G \in T$. We have $G = S_{J(B)}(u,\langle b,G\rangle,u) = S_{J(B)}(g(u),g(\langle a,F\rangle),g(u)) = g(S_{J(A)}(u,\langle a,F\rangle,u)) = g(F) = F$; also, $b = S_{J(B)}(u,u,\langle b,G\rangle) = S_{J(B)}(g(u), g(u),g(\langle a,F\rangle)) = g(S_{J(A)}(u,u,\langle a,F\rangle)) = g(a) = f(a)$. Hence $g(\langle a,F\rangle) = \langle f(a),F\rangle$.

We have proved $g = J(f)$ and it remains to prove that f is a homomorphism of A into B. Let $a \in A$ and $F \in T$. We have $\langle f(F_A(a)),F\rangle = g(\langle F_A(a),F\rangle) = g(S_{J(A)}(\langle a,F\rangle,u,u)) = S_{J(B)}(g(\langle a,F\rangle),g(u),g(u)) = S_{J(B)}(\langle f(a),F\rangle,u,u) = \langle F_B(f(a)),F\rangle$, so that $f(F_A(a)) = F_B(f(a))$. □

2.6. LEMMA. *Let σ be the signature containing just one ternary operation symbol S and τ be the signature containing just two unary operation symbols f,G. There exists a categorical embedding of the class of all σ-algebras into the class of all τ-algebras.*

PROOF. For every σ-algebra A define a τ-algebra $J(A)$ in this way: the underlying set of $J(A)$ is the disjoint union $(A\times A\times A\times\{0,1\})\cup(A\times\{2,3\})$;

(1) $F_{J(A)}(\langle a,b,c,0\rangle) = \langle b,c,a,1\rangle$
(2) $F_{J(A)}(\langle a,b,c,1\rangle) = \langle a,b,c,0\rangle$
(3) $F_{J(A)}(\langle a,2\rangle) = F_{J(A)}(\langle a,3\rangle) = \langle a,2\rangle$
(4) $G_{J(A)}(\langle a,b,c,0\rangle) = \langle a,2\rangle$
(5) $G_{J(A)}(\langle a,b,c,1\rangle) = \langle S_A(a,b,c),3\rangle$
(6) $G_{J(A)}(\langle a,2\rangle) = G_{J(A)}(\langle a,3\rangle) = \langle a,3\rangle$

If A,B are two σ-algebras then for any mapping f of A into B we define a mapping $J(f)$ of $J(A)$ into $J(B)$ in this way:

(1) $J(f)(\langle a,b,c,i\rangle) = \langle f(a), f(b), f(c), i\rangle$
(2) $J(f)(\langle a,j\rangle = \langle f(a), j\rangle$

It is easy to check that if f is a homomorphism of A into B then $J(f)$ is a homomorphism of $J(A)$ into $J(B)$. Let A, B be two σ-algebras and g be a homomorphism of $J(A)$ into $J(B)$. It remains to show that $g = J(f)$ for a homomorphism f of A into B.

If $a \in A$ then $g(\langle a,3\rangle = g(G_{J(B)}(\langle a,3\rangle)) = G_{J(B)}(g(\langle a,3\rangle))$, so that $g(\langle a,3\rangle) = \langle b,3\rangle$ for an element $b \in B$, since only elements of $B \times \{3\}$ can be fixed points of $G_{J(B)}$. So, we can define a mapping f of A into B by $g(\langle a,3\rangle) = \langle f(a),3\rangle$. For $a \in A$ we have $g(\langle a,2\rangle) = g(F_{J(A)}(\langle a,3\rangle)) = F_{J(B)}(g(\langle a,3\rangle)) = F_{J(B)}(\langle f(a),3\rangle) = \langle f(a),2\rangle$.

Let $a,b,c \in A$. We have $G_{J(B)}(g(\langle a,b,c,0\rangle)) = g(G_{J(A)}(\langle a,b,c,0\rangle)) = g(\langle a,2\rangle) = \langle f(a),2\rangle$, so that $g(\langle a,b,c,0\rangle) = \langle a',b',c',0\rangle$ for some $a',b',c' \in B$. We have $g(\langle a,b,c,1\rangle) = g(F^5_{J(A)}(\langle a,b,c,0\rangle)) = F^5_{J(B)}(g(\langle a,b,c,0\rangle)) = F^5_{J(B)}(\langle a',b',c',0\rangle) = \langle a',b',c',1\rangle$. Also, $\langle a',2\rangle = G_{J(B)}(\langle a',b',c',0\rangle) = G_{J(B)}(g(\langle a,b,c,0\rangle)) = g(G_{J(A)}(\langle a,b,c,0\rangle)) = g(\langle a,2\rangle) = \langle f(a),2\rangle$, so that $a' = f(a)$; we have $\langle b',2\rangle = G_{J(B)}(F^2_{J(B)}(\langle a',b',c',0\rangle)) = G_{J(B)}(F^2_{J(B)}(g(\langle a,b,c,0\rangle))) = g(G_{J(A)}(F^2_{J(A)}(\langle a,b,c,0\rangle))) = g(\langle b,2\rangle) = \langle f(b),2\rangle$, so that $b' = f(b)$; we have $\langle c',2\rangle = G_{J(B)}(F^4_{J(B)}(\langle a',b',c',0\rangle)) = G_{J(B)}(F^4_{J(B)}(g(\langle a,b,c,0\rangle))) = g(G_{J(A)}(F^4_{J(A)}(\langle a,b,c,0\rangle))) = g(\langle c,2\rangle) = \langle f(c),2\rangle$, so that $c' = f(c)$. Hence $g(\langle a,b,c,0\rangle) = \langle f(a),f(b),f(c),0\rangle$. Then also $g(\langle a,b,c,1\rangle) = \langle f(a),f(b),f(c),1\rangle$. We have proved $g = J(f)$.

It remains to prove that f is a homomorphism of A into B. Let $a,b,c \in A$. We have $\langle f(S_A(a,b,c)),3\rangle = g(\langle S_A(a,b,c),3\rangle) = g(G_{J(A)}(\langle a,b,c,1\rangle)) = G_{J(B)}(g(\langle a,b,c,1\rangle)) = G_{J(B)}(\langle f(a),f(b),f(c),1\rangle) = \langle S_B(f(a),f(b),f(c)),3\rangle$, so that $f(S_A(a,b,c)) = S_B(f(a),f(b),f(c))$. □

2.7. Lemma. *Let σ be the signature containing just two unary operation symbols F, G and τ be the signature containing F, G and, moreover, one constant* 0. *There exists a categorical embedding of the class of all σ-algebras into the class of all τ-algebras.*

Proof. For every set A take an element u_A not belonging to A and put $A' = A \cup \{u_A\}$. For every σ-algebra A define a τ-algebra $J(A)$ with the underlying set $(A' \times A' \times \{0,1\}) \cup (A' \times \{2,3\}) \cup \{v\}$ (where v is a fixed element that is not an ordered pair);

(1) $0_{J(A)} = v$
(2) $F_{J(A)}(\langle x,y,0\rangle) = \langle y,x,1\rangle$ for $x,y \in A'$
(3) $F_{J(A)}(\langle x,y,1\rangle) = \langle x,y,0\rangle$ for $x,y \in A'$
(4) $F_{J(A)}(\langle a,2\rangle) = F_{J(A)}(\langle a,3\rangle) = \langle a,2\rangle$ for $a \in A$
(5) $F_{J(A)}(\langle u_A,2\rangle) = F_{J(A)}(\langle u_A,3\rangle) = v$
(6) $F_{J(A)}(v) = \langle u_A,2\rangle$
(7) $G_{J(A)}(v) = v$
(8) $G_{J(A)}(\langle x,2\rangle) = \langle x,3\rangle$ for $x \in A'$

(9) $G_{J(A)}(\langle x,3\rangle) = v$ for $x \in A'$
(10) $G_{J(A)}(\langle x,y,0\rangle) = \langle x,2\rangle$ for $x,y \in A'$
(11) $G_{J(A)}(\langle a,u_A,1\rangle) = \langle F_A(a),3\rangle$ for $a \in A$
(12) $G_{J(A)}(\langle u_A,a,1\rangle) = \langle G_A(a),3\rangle$ for $a \in A$
(13) $G_{J(A)}(\langle x,y,1\rangle) = v$ in all other cases

If A,B are two σ-algebras then for every mapping f of A into B we define a mapping f' of A' into B' by $f'(a) = f(a)$ for $a \in A$ and $f'(u_A) = u_B$; we define a mapping $J(f)$ of $J(A)$ into $J(B)$ by

(1) $J(f)(\langle x,y,i\rangle) = \langle f'(x),f'(y),i\rangle$
(2) $J(f)(\langle x,j\rangle) = \langle f'(x),j\rangle$
(3) $J(f)(v) = v$

It is easy to check that if f is a homomorphism of A into B then $J(f)$ is a homomorphism of $J(A)$ into $J(B)$. Let A,B be two σ-algebras and g be a homomorphism of $J(A)$ into $J(B)$. It remains to show that $g = J(f)$ for some homomorphism f of A into B.

Clearly, $g(v) = v$.

We have $g(\langle u_A,2\rangle) = g(F_{J(A)}(v)) = F_{J(B)}(g(v)) = F_{J(B)}(g(\langle a,2\rangle))$ and hence $g(\langle u_A,3\rangle) = g(G_{J(A)}(\langle u_A,2\rangle)) = G_{J(B)}(g(\langle u_A,2\rangle)) = G_{J(B)}(\langle u_B,2\rangle) = \langle u_B,3\rangle$.

For $a \in A$ we have $g(\langle a,2\rangle) = g(F_{J(A)}(\langle a,2\rangle)) = F_{J(B)}(g(\langle a,2\rangle))$, from which clearly $g(\langle a,2\rangle) = \langle b,2\rangle$ for some $b \in B$. We can define a mapping f of A into B by $g(\langle a,2\rangle) = \langle f(a),2\rangle$.

For $a \in A$ we have $g(\langle a,3\rangle) = g(G_{J(A)}(\langle a,2\rangle)) = G_{J(B)}(g(\langle a,2\rangle)) = G_{J(B)}(\langle f(a),2\rangle) = \langle f(a),3\rangle$.

If $x,y \in A'$ then $G_{J(B)}(g(\langle x,y,0\rangle)) = g(G_{J(A)}(\langle x,y,0\rangle)) = g(\langle x,2\rangle) = \langle f'(x),2\rangle$, hence $g(\langle x,y,0\rangle) = \langle x',y',0\rangle$ for some $x',y' \in B'$. We have $\langle x',2\rangle = G_{J(B)}(\langle x',y',0\rangle) = G_{J(B)}(g(\langle x,y,0\rangle)) = g(G_{J(A)}(\langle x,y,0\rangle)) = g(\langle x,2\rangle) = \langle f'(x),2\rangle$ and $\langle y',2\rangle = G_{J(B)}(F^2_{J(B)}(\langle x',y',0\rangle)) = G_{J(B)}(F^2_{J(B)}(g(\langle x,y,0\rangle))) = g(G_{J(A)}(F^2_{J(A)}(\langle x,y,0\rangle))) = g(\langle y,2\rangle) = \langle f'(y),2\rangle$; hence $g(\langle x,y,0\rangle) = \langle f'(x), f'(y),0\rangle$. Also, $g(\langle x,y,1\rangle) = g(F^3_{J(A)}(\langle x,y,0\rangle)) = F^3_{J(B)}(g(\langle x,y,0\rangle)) = F^3_{J(B)}(\langle f'(x),f'(y),0\rangle) = \langle f'(x),f'(y),1\rangle$.

We have proved $g = J(f)$. It remains to show that f is a homomorphism of A into B. Let $a \in A$. We have $\langle f(F_A(a)),3\rangle = g(\langle F_A(a),3\rangle) = g(G_{J(A)}(\langle a,u_A,1\rangle)) = G_{J(B)}(g(\langle a,u_A,1\rangle)) = G_{J(B)}(\langle f(a),u_B,1\rangle) = \langle F_B(f(a)),3\rangle$, so that $f(F_A(a)) = F_B(f(a))$. Similarly $\langle f(G_A(a)),3\rangle = g(\langle G_A(a),3\rangle) = g(G_{J(A)}(\langle u_A,a,1\rangle)) = G_{J(B)}(g(\langle u_A,a,1\rangle)) = G_{J(B)}(\langle u_B,f(a),1\rangle) = \langle G_B(f(a)),3\rangle$, so that $f(G_A(a)) = G_B(f(a))$. □

2.8. LEMMA. *Let σ be the signature containing just two unary operation symbols F,G. There exists a categorical embedding of the class of all σ-algebras into the class of all groupoids.*

PROOF. For every σ-algebra A define a groupoid $J(A)$ with the underlying set $A \cup \{p,q\}$ (where p,q are two distinct elements not in A) in this way:

(1) $a \cdot p = F_A(a)$ for $a \in A$
(2) $p \cdot a = G_A(a)$ for $a \in A$
(3) $q \cdot q = p$
(4) $x \cdot y = q$ in all other cases

If A, B are two σ-algebras then for every mapping f of A into B we define a mapping $J(f)$ of $J(A)$ into $J(B)$ by

(1) $J(f)(a) = f(a)$ for $a \in A$
(2) $J(f)(p) = p$ and $J(f)(q) = q$

It is easy to check that if f is a homomorphism of A into B then $J(f)$ is a homomorphism of $J(A)$ into $J(B)$. Let A, B be two σ-algebras and g be a homomorphism of $J(A)$ into $J(B)$. We need to show that $g = J(f)$ for a homomorphism f of A into B.

Since for every $b \in B$ we have $b \cdot b \in \{p, q\}$ and since $p \cdot p = q$ and $q \cdot q = p$, it is clear that g maps $\{p, q\}$ onto itselt. Hence $g(q) = g(p \cdot q) = g(p) \cdot g(q) = q$; then also $g(p) = p$.

Let $a \in A$. If $g(a) = q$ then $q = g(q) = g(a \cdot q) = g(a) \cdot g(q) = q \cdot q = p$, a contradiction. If $g(a) = p$ then $q = p \cdot p = g(p) \cdot g(a) = g(p \cdot a) \neq q$, a contradiction. Hence $g(a) \in B$. Denote by f the restriction of g to A, so that f is a mapping of A into B and $g = J(f)$.

It remains to prove that f is a homomorphism of A into B. For $a \in A$ we have $f(F_A(a)) = g(F_A(a)) = g(a \cdot p) = g(a) \cdot g(p) = f(a) \cdot p = F_B(f(a))$ and similarly $f(G_A(a)) = g(G_A(a)) = g(G_A(a)) = g(p \cdot a) = g(p) \cdot g(a) = p \cdot f(a) = G_B(f(a))$. □

2.9. LEMMA. *Let σ be the signature containing just two unary operation symbols; let τ be the signature containing one binary operation symbol and one constant. There exists a categorical embedding of the class of all σ-algebras into the class of all τ-algebras.*

PROOF. The proof of 2.8 can be repeated with the only modification that the constant should be interpreted by the element p. □

2.10. THEOREM. (Hedrlín, Pultr [66]) *Let σ be any signature and τ be a large signature. There exists a categorical embedding of the class of all σ-algebras into the class of all τ-algebras.*

PROOF. It follows from the above lemmas. □

2.11. COROLLARY. *Every monoid is isomorphic to the endomorphism monoid of a groupoid. Every group is isomorphic to the automorphism group of a groupoid.*

OPEN PROBLEMS

Problem 1. *Is every finite lattice isomorphic to the congruence lattice of a finite algebra?*

Problem 2. *Characterize those lattices* (*or at least those finite lattices*) *that are isomorphic to the lattice of all subvarieties of some variety.*

(See 15.2.1.)

Problem 3. *Let A be a finite algebra of a finite signature and V be the variety generated by A; let V be residually very finite. Must V be finitely based?*

Problem 4. *Is there an algorithm deciding for every finite algebra A of a finite signature whether the quasivariety generated by A is finitely axiomatizable?*

Problem 5. *Characterize the varieties V such that whenever V is properly contained in a variety W then there is a subvariety of W that covers V.*

Clearly, every finitely based variety has this property. It follows from Theorem 6.12.2 that also every balanced variety has this property. However, not every variety has this property.

Problem 6. *Let V be a locally finite variety of a finite signature. We say that V is finitely based at finite level if there exists a finitely based variety W such that V and W have the same finite algebras. Is it true that V is finitely based whenever it is finitely based at finite level?*

This problem is due to S. Eilenberg and M.P. Schützenberger [76]. In 1993, R. Cacioppo proved that if V is finitely generated and finitely based at finite level but not finitely based then V is inherently nonfinitely based.

Problem 7. *Find an algorithm* (*or prove that no such algorithm exists*) *deciding for any equation $\langle u, v\rangle$ of a given signature whether the variety V based on $\langle u, v\rangle$ has an upper semicomplement in the lattice of all varieties* (*i.e., whether $V \vee W = U$ for some variety $W \subset U$, where U is the variety of all algebras*)*.*

Problem 8. *Characterize the equations $\langle u, v\rangle$ such that the variety based on $\langle u, v\rangle$ is a meet-irreducible element of the lattice of all varieties of the given signature.*

A conjecture is that (if the signature contains at least one operation symbol of positive arity) this is the case if and only if $\langle u, v\rangle$ is regular and the terms u, v are either incomparable or one can be obtained from the other by a permutation of variables of prime power order. For details see Ježek [83].

PROBLEM 9. *Can one effectively construct, for any nonempty finite set B of nontrivial equations of a finite signature, another finite set B' of equations such that the variety V_B based on B is covered by the variety $V_{B'}$ based on B'?*

By Ježek and McNulty [95a], one can effectively construct a finite set B'' of equations such that V_B is properly contained in $V_{B''}$ and the number of varieties between these two varieties is finite and can be effectively estimated by an upper bound. However, this still does not yield positive solution to our problem.

PROBLEM 10. *Is the equational theory based on any finite number of equations of the form $\langle F(t_1, \dots, t_n), t_1 \rangle$ (t_i are any terms) always decidable?*

PROBLEM 11. *Is there an algorithm deciding for any equation $\langle u, v \rangle$ in a single variable, such that the two terms u and v are incomparable, whether the equational theory based on $\langle u, v \rangle$ is term finite? Is the answer always yes?*

References

Baker K.A.

[77] *Finite equational bases for finite algebras in a congruence-distributive equational class.* Advances in Math. **24**, 207-243.

Baker K., McNulty G., Werner H.

[87] *The finitely based varieties of graph algebras.* Acta Sci. Math. Szeged **51**, 3–15.

[89] *Shift-automorphism methods for inherently nonfinitely based varieties of algebras.* Czechoslovak Math. J. **39**, 53–69.

Baker K.A., Pixley A.F.

[75] *Polynomial interpolation and the Chinese remainder theorem for algebraic systems.* Math. Z. **43**, 165-174.

Berman J.

[80] *A proof of Lyndon's finite basis theorem.* Discrete Math. **29**, 229-233.

Birkhoff G.

[35] *On the structure of abstract algebras.* Proc. Cambridge Philos. Soc. **31**, 433-454.

[44] *Subdirect unions in universal algebra.* Bull. Amer. Math. Soc. **50**, 764-768.

Bryant R.

[82] *The laws of finite pointed groups.* Bull. London Math. Soc. **14**, 119-123.

Burris S., Sankappanavar H.P.

[81] *A course in universal algebra.* Graduate Texts in Mathematics, Springer-Verlag, New York.

Cohen P.J.

[66] *Set theory and the continuum hypothesis.* W.A. Benjamin, Inc. New York, Amsterdam.

Day A.

[69] *A characterization of modularity for congruence lattices of algebras.* Canadian Math. Bull. **12**, 167-173.

Dershowitz N., Jouannaud J.-P.

[90] *Rewrite systems.* Chapter 6, 243–320 in J. van Leeuwen, ed., Handbook of Theoretical Computer Science, B: Formal Methods and Semantics. North Holland, Amsterdam 1990.

Evans T.

[51] *On multiplicative systems defined by generators and relations. I. Normal form theorems.* Proc. Cambridge Phil. Soc. **47**, 637-649.

[53] *Embeddability and the word problem.* J. London Math. Soc. **28**, 76-80.

Fleischer I.

[55] *A note on subdirect products.* Acta Math. Acad. Sci. Hungar. **6**, 463-465.

Foster A.L.

[53] *Generalized "Boolean" theory of universal algebras. Part I: Subdirect sums and normal representation theorem.* Math. Z. **58**, 306-336.

[53a] *Generalized "Boolean" theory of universal algebras. Part II: Identities and subdirect sums in functionally complete algebras.* Math. Z. **59**, 191-199.

Foster A.L., Pixley A.F.

[64] *Semi-categorical algebras I. Semi-primal algebras.* Math. Z. **83**, 147-169.

[64a] *Semi-categorical algebras II.* Math. Z. **85**, 169-184.

Freese R., McKenzie R.

[87] *Commutator theory for congruence modular varieties.* London Mathematical Society Lecture Notes **125**, Cambridge Univ. Press, 227pp.

Fried E., Pixley A.F.

[79] *The dual discriminator function in universal algebra.* Acta Sci. Math. Szeged **41**, 83-100.

García O.C., Taylor W.

[84] *The lattice of interpretability types of varieties.* Memoirs AMS **50**, v+125 pp.

Gödel K.

[40] *The consistency of the axiom of choice and of the generalized continuum-hypothesis with the axioms of set theory.* Princeton University Press.

Gorbunov V.A.

[99] *Algebraic theory of quasivarieties.* (Algebraicheskaya teoriya kvazimnogoobrazij.) (Russian) Sibirskaya Shkola Algebry i Logiki. 5. Novosibirsk: Nauchnaya Kniga, xii+368 p. English translation by Plenum, New York, 1998, xii+298 p.

Grätzer G.

[79] *Universal algebra, Second edition.* Springer-Verlag, New York.

Grätzer G., Schmidt E.T.

[63] *Characterizations of congruence lattices of abstract algebras.* Acta Sci. Math. Szeged **24**, 34-59.

Gumm H.P.

[83] *Geometrical methods in congruence modular algebras.* Memoirs of the AMS **45**, no.286.

Hedrlín Z., Pultr A.

[66] *On full embeddings of categories of algebras.* Illinois J. Math. **10**, 392-406.

Hobby D., McKenzie R.

[88] *The Structure of Finite Algebras.* Contemporary Mathematics, AMS, Providence, RI.

Ježek J.

[69] *Primitive classes of algebras with unary and nullary operations.* Colloquium Math. **20**, 159-179.

[85a] *Nonfinitely based three-element idempotent groupoids.* Algebra Universalis **20**, 292-301.

[83] *On join-indecomposable equational theories.* Lecture Notes in Math. **1004**, 159-165.

Ježek J., McNulty G.F.

[95a] *The existence of finitely based lower covers for finitely based equational theories.* J. Symbolic Logic **60**, 1242-1250.

[95b] *Perfect bases for equational theories.* J. Symbolic Computation **19**, 489-505.

Jónsson B.

[67] *Algebras whose congruence lattices are distributive.* Math. Scand. **21**, 110-121.

[79] *Congruence varieties.* Appendix 3 in: G. Grätzer, Universal Algebra (2nd ed.), Springer-Verlag.

[95] *Congruence distributive varieties.* Math. Japonica **42**, 353-401.

Kiss E.W., Valeriote M.

[93] *Abelian Algebras and the Hamiltonian Property.* J. Pure and Appl. Alg. **87**, 37–49.

Klukovits L.

[75] *Hamiltonian varieties of universal algebras.* Acta Sci. Math. Szeged **37**, 11-15.

Knuth D.E., Bendix P.B.

[70] *Simple word problems in universal algebras.* In: J. Leech, ed., Computational Problems in Abstract Algebra (Proc. Conf., Oxford, 1967). Pergamon Press, Oxford, 263-297.

Lampe W.

[86] *A property of the lattice of equational theories.* Algebra Universalis **23**, 61-69.

Lyndon R.C.

[51] *Identities in two-valued calculi.* Trans. Amer. Math. Soc. **71**, 457-465.

Mal'cev A.I.

[54] *On the general theory of algebraic systems.* (Russian) Mat. Sbornik **35**, 3-20.

McKenzie R.

[70] *Equational bases for lattice theories.* Mathematica Scandinavica **27**, 24-38.

[75] *On spectra, and the negative solution of the decision problem for identities having a finite non-trivial model.* J. Symbolic Logic **40**, 186-196.

[78] *Para primal varieties: a study of finite axiomatizability and definable principal congruences in locally finite varieties.* Algebra Universalis **8**, 336-348.

[87] *Finite equational bases for congruence modular algebras.* Algebra Universalis **24**, 224-250.

[96] *The residual bounds of finite algebras.* International J. of Algebra and Computation **6**, 1-28.

[96a] *The residual bound of a finite algebra is not computable.* International J. of Algebra and Computation **6**, 29-48.

[96b] *Tarski's finite basis problem is undecidable.* International J. of Algebra and Computation **6**, 49-104.

McKenzie R., Hobby D.

[88] *The structure of finite algebras (tame congruence theory).* Contemporary Mathematics, AMS, Providence, R.I.

McKenzie R., McNulty G.F., Taylor W.

[87] *Algebras, lattices, varieties. Vol. I.* Wadsworth & Brooks/Cole, Monterey, Calif.

McNulty G.F.

[85] *How to construct finite algebras which are not finitely based.* Lecture Notes in Math. **1149**, 167-174.

Murskij V.L.

[65] *The existence in three-valued logic of a closed class with finite basis, without a finite complete set of identities.* (Russian) DAN SSSR **163**, 815-818.

[75] *Finite basedness of identities and other properties of 'almost all' finite algebras.* (Russian) Problemy Kibernet. **30**, 43-56.

Oates-Williams S., Powell M.B.

[64] *Identical relations in finite groups.* J. Algebra **1**, 11-39.

Pálfy P.P.

[84] *Unary polynomials in algebras I.* Algebra Universalis **18**, 262-273.

Perkins P.

[84] *Basic questions for general algebras.* Algebra Universalis **19**, 16-23.

Pixley A.F.

[71] *The ternary discriminator function in universal algebra.* Math. Ann. **191**, 167-180.

Pudlák P.

[76] *A new proof of the congruence lattice characterization theorem.* Algebra Universalis **6**, 269-275.

Trakhtman A.N.

[74] *On covers in the lattice of varieties of algebras.* (Russian) Matem. Zametki **15**, 307-312.

Vopěnka P., Hájek P.

[72] *The theory of semisets.* Academia, Praha.

Vopěnka P., Pultr A., Hedrlín Z.

[65] *A rigid relation exists on any set.* Commentationes Math. Univ. Carolinae **6**, 149-155.

Werner H.

[74] *Congruences on products of algebras and functionally complete algebras.* Algebra Universalis **4**, 99-105.

Willard R.

[00] *A finite basis theorem for residually finite, congruence meet-semidistributive varieties.* J. Symbolic Logic **65**, 187-200.

Index