Notes on CSPs and Polymorphisms

Contents

1 Introduction / Advertisement 2
2 The Inv-Pol Galois connection 16
3 Three basic examples 20
4 Varieties, Birkhoff’s HSP theorem, and the hardness proof 23
5 Cores and Idempotent Reducts 27
   5.1 Reflections and Height 1 Identities 31
6 Taylor Algebras 33
7 Two simple algorithms (width 1 and bounded strict width) 37
   7.1 The Basic LP relaxation of a CSP 44
8 Mal’cev algebras 48
9 Mal’cev algorithm and compact representations 53
   9.1 Near-subgroups 57
10 Abelian Mal’cev algebras are affine 61
   10.1 Commutators 71
11 Generalized Majority-Minority operations (motivating Few Subpowers) 77
12 Algebras with Few Subpowers 83
   12.1 Some connections with congruence modularity 91
13 Parallelogram terms 98
   13.1 Critical rectangular relations in congruence modular varieties 102
14 Learnability of relations encoded by compact representations 107
15 Algebras with few subpowers are finitely related 116
16 Fourth basic example: the Rock-Paper-Scissors algebra 122
1 Introduction / Advertisement

In this section we’ll state many of the results and motivating questions that we’ll try to understand in these notes. If you don’t understand something written here right away, don’t despair - we’ll go over everything in more detail later. The impatient reader can safely skip ahead to Section 2.

The story starts with a result of Schaefer [103] on a problem he called “Generalized Satisfiability”.

**Definition 1.1.** If $\Gamma$ is a set of relations on $\{0, 1\}$, then GenSAT($\Gamma$) is the decision problem which takes as input a set of variables $V$ and a collection of constraints, where each constraint is of the form “the relation $R(v_1, ..., v_k)$ must be satisfied” where $(v_1, ..., v_k)$ is a tuple of variables of $V$ and $R$ is a relation from $\Gamma$ of arity $k$, and where the desired output is whether or not it is possible to assign values in $\{0, 1\}$ to the variables such that the assignment satisfies all of the given constraints.

**Theorem 1.2 (Schaefer [103]).** If GenSAT($\Gamma$) is not NP-complete, then $\Gamma$ is contained in one of the following sets of relations:

- the set of relations containing the all-0s vector,
- the set of relations containing the all-1s vector,
- the set of relations which can be written as an intersection of Horn clauses, where a Horn clause is a disjunction of literals such that at most one variable appears positively,
- the set of relations which can be written as an intersection of dual-Horn clauses, where a dual-Horn clause is a disjunction of literals such that at most one variable appears negatively,
- the set of relations which can be written as an intersection of relations involving at most two variables,
• the set of relations which can be written as solution sets to systems of linear equations over \( \mathbb{F}_2 \).

In each of these cases, GenSAT(\( \Gamma \)) can be solved in polynomial time.

The next result of this form is due to Hell and Nešetril [61], on a generalization of \( n \)-coloring which they call “\( H \)-coloring”.

**Definition 1.3.** If \( H \) is a graph, then \( H \)-coloring is the decision problem which takes a graph \( G \) as input, and where the desired output is whether or not there is a graph homomorphism from \( G \) to \( H \).

Note that if we take \( H = K_n \), then \( K_n \)-coloring is equivalent to \( n \)-coloring.

**Theorem 1.4** (Hell, Nešetril [61]). \( H \)-coloring is in \( P \) if \( H \) is bipartite, and it is NP-complete otherwise.

These two results led Feder and Vardi [53] to ask whether there is a general dichotomy between \( P \) and \( NP \). However, any such dichotomy has to avoid Ladner’s [82] anti-dichotomy result.

**Theorem 1.5** (Ladner [82]). If \( P \neq NP \), then there are problems in \( NP \) which are neither in \( P \) nor NP-complete.

In order to find a subclass of \( NP \) which avoids Ladner’s Theorem, Feder and Vardi used a result of Fagin [51] which showed an equivalence between problems in \( NP \) and problems in finitary existential second order logic. I hope I can state these results correctly, as I’m not very familiar with the “strict \( NP \)” point of view.

**Theorem 1.6** (Fagin [51]). A set \( S \) of finite models of a language \( \mathcal{L} \) is in \( NP \) if and only if there exists a first order sentence \( \phi \) in a language \( \mathcal{L}' \) containing \( \mathcal{L} \) such that \( S = \{ M : \exists R_1, \ldots, R_k (M, R_1, \ldots, R_k) \models \phi, |M| < \infty \} \), where \( R_1, \ldots, R_k \) are the new function and relation symbols of \( \mathcal{L}' \).

**Definition 1.7.** A set of finite models of a relational language \( \mathcal{L} \) is in SNP if it can be characterized by a second order logic formula \( \varphi \) of the form \( \exists R_1, \ldots, R_k \forall x_1, \ldots, x_n \phi \), where \( R_1, \ldots, R_k \) are relations and \( \phi \) is quantifier-free. The formula \( \varphi \) is called **monadic** if each relation \( R_i \) is unary, \( \varphi \) is called **monotone** if each relation of \( \mathcal{L} \) appears in the formula \( \phi \) only negatively, and we say that \( \varphi \) is **without inequality** if the formula \( \phi \) has no negative occurrences of \( \neq \) (or positive occurrences of \( = \)). We say that a set of finite models of a relational language is in MMSNP if it can be characterized by a second order existential formula \( \varphi \) as above which is monotone, monadic, and without inequality.

Feder and Vardi [53] showed that for any two of “monotone”, “monadic”, and “without inequality”, if one restricts to the subclass of SNP defined by formulas satisfying these two properties, one can find a polynomially equivalent problem in this subclass to any problem in \( NP \), so Ladner’s Theorem still applies. This motivated the restriction to MMSNP. Using randomized constructions of hypergraphs of large girth, Feder and Vardi then showed that problems in MMSNP are equivalent to constraint satisfaction problems.

**Definition 1.8.** A CSP-template \( T \) consists of a finite set \( D \) together with a finite collection \( \Gamma = (R_1, \ldots, R_n) \) of relations on \( D \) - equivalently, we can think of \( T \) as a relational structure \( (D, R_1, \ldots, R_n) \). The decision problem CSP(\( T \)) takes as input a list of variables \( V \) and for each \( i \leq n \) a list of tuples \( C_i \) of variables of \( V \) which are required to satisfy the constraint \( R_i \), and accepts if there exists an assignment of variables to values in the set \( D \) satisfying the given constraints.
**Theorem 1.9** (Feder, Vardi [53]). Every problem in MMSNP is equivalent under randomized polynomial time reductions to a CSP, and vice versa.

Having justified restricting our studies to CSPs, we now give some examples of particularly famous CSP templates.

**Example 1.1.** The problem $k$-COLORING (given a graph, determine if it can be colored with $k$ colors) is equivalent to CSP$(\{1,\ldots,k\},\neq) = \text{CSP}(K_k)$, where $K_k$ is the complete graph of $k$ vertices (considered as a relational structure).

**Example 1.2.** The problem 2SAT is equivalent to CSP$(\{0,1\}, \leq, \neq)$. This problem is in P - in fact, it is known to be NL-complete (NL stands for nondeterministic logspace), and it can be solved in linear time.

**Example 1.3.** The problem 3SAT can be thought of as CSP$(\{0,1\}, R_{(0,0,0)}, \ldots, R_{(1,1,1)})$, where $R_{(i,j,k)} = \{0,1\}^3 \setminus \{(i,j,k)\}$. We can also simplify this to the equivalent problem CSP$(\{0,1\}, \{0,1\}^3 \setminus \{(0,0,0)\}, \neq)$.

**Example 1.4.** The problem NAE-SAT is CSP$(\{0,1\}, \text{NAE})$, where $\text{NAE} = \{0,1\}^3 \setminus \{(0,0,0),(1,1,1)\}$ is the relation that states that the three variables in question are not all equal. This CSP template is known to be NP-complete.

**Example 1.5.** The problem 1-IN-3 SAT is CSP$(\{0,1\}, \{(0,0,1),(0,1,0),(1,0,0)\})$. This CSP template is known to be NP-complete.

**Example 1.6.** The problem HORN-SAT is CSP$(\{0,1\}, \{0\}, \{1\}, \{0,1\}^3 \setminus \{(1,1,0)\})$ (the third constraint is $(x \land y) \implies z$). This problem is known to be P-complete, and it can be solved in linear time.

**Example 1.7.** The problem XOR-SAT is CSP$(\{0,1\}, \{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\}, \neq)$. This problem is in P - in fact, it can be solved in deterministic time $n^{\log_2(7)}$ and randomized quadratic time [109] (it is unknown if it can be solved in linear time).

Generalizing the XOR-SAT example to a larger domain, we have the following very general family of problems which can be thought of as the natural generalization of systems of linear equations, over a possibly noncommutative group.

**Example 1.8.** Let $G$ be a finite group, and consider the CSP template with domain $G$, and with a relation $gH$ for every subgroup $H \leq G^n$ and every element $g \in G^n$, for every $n$. Note that strictly speaking, this is not a CSP (as we have defined it) since the set of relations is infinite. Feder and Vardi [53] prove that this general subgroup problem is polynomially solvable.

Based on the examples they knew at the time, Feder and Vardi guessed that tractable CSPs fall into two types: “bounded width” problems, which are solved by local propagation of information, and problems with “the ability to count” such as the subgroup problems above. They further divided the bounded width problems into two main subclasses: problems with “width 1” (such as HORN-SAT) and problems with “bounded strict width” (such as 2-SAT).

The bounded width problems can be defined formally in terms of a logic programming language called Datalog (a simple subset of the programming language Prolog), where a program consists of rules for updating a database of known facts about tuples of variables by adding new facts if certain preconditions are met. For instance, a program to determine whether a graph is connected might have two predicates, one for the edges of the graph and another for connectivity, and a rule that says “if connected(a,b) and edge(b,c), then add connected(a,c) to the database”. This
Example program maintains facts about pairs of variables, but has rules that involve examining three variables at a time.

**Definition 1.10.** A CSP has width \((l, k)\), \(k \geq l\) if it can be solved by a Datalog program which keeps track of facts about tuples of at most \(l\) variables, and updates its database by using rules that examine at most \(k\) variables at a time. We say that it has width \(l\) if there exists any \(k\) such that it has width \((l, k)\).

In some cases we want to consider CSPs with relations of arbitrarily large arities. In these cases, one uses the concept of *relational width*, introduced by Bulatov [34], where our Datalog program is also allowed to update its database of facts about \(l\)-tuples of variables by using rules that examine any set of variables which is contained in the scope of some constraint relation, and to shrink our constraint relations based on facts about \(l\)-tuples of variables.

As it turns out, there is a canonical Datalog program for solving problems of width \((l, k)\), which correctly solves every instance of a CSP if and only if the CSP has width \((l, k)\). This program just keeps track of the set of all possible assignments to each tuple of at most \(l\) variables, and eliminates possibilities from these lists by brute-forcing the set of possible assignments to each \(k\)-tuple of variables in turn (checking for consistency with each subset of these variables of size \(\leq l\)), until it can no longer eliminate any further possible assignments from its database. If there are \(n\) variables, this runs in time \(O(n^k)\) and space \(O(n^l)\).

A slight weakening of the above canonical Datalog program with width 1, in which we only consider one relation at a time in order to remove potential values for the variables, is called “arc-consistency”, or sometimes “generalized arc-consistency” if the relations have arity greater than 2. CSPs which can be solved by arc-consistency have a special property called “tree-duality”, which says that an instance has a solution if and only its “universal cover” has a solution (the universal cover is an instance with variables and constraints forming an infinite tree that corresponds to the universal cover of the (hyper-)graph of variables and constraints of the original instance).

The width of a CSP can also be defined in terms of a two player game (see [5]), in which one player (the Prover) tries to convince the other player (the Verifier) that an instance of the CSP has a solution. The game goes as follows: in each round of the game, the Prover has assigned values to a certain tuple of at most \(l\) variables (at the beginning of the game, this tuple is empty). The Verifier then picks a superset of the previous tuple of size at most \(k\), and challenges the Prover to extend their assignment to this larger collection of variables. After this the Verifier selects any subset of the variables of size at most \(l\), restricting the assignment to that subset, and the next round begins. The Verifier wins if at any point the Prover’s assignment fails to satisfy some constraint of the CSP. Then a CSP has width \((l, k)\) if the Prover has a winning strategy only when the problem has a valid global solution.

**Definition 1.11.** A CSP has strict width \(l\) if, whenever a partial solution to an instance of the CSP has no extension to a full solution, there exists a subset of the partial solution of size at most \(l\), such that this subset already has no extension to a full solution. Equivalently, for every instance of the CSP, the projection of the solution set onto any set of \(k > l\) variables is completely determined by the projections of the solution set onto subsets of those variables of size \(l\).

As a consequence of the above definition, if a CSP has strict width \(l\), then any constraint having arity greater than \(l\) must be expressible as a conjunction of constraints involving at most \(l\) variables. Feder and Vardi [53] prove that one can check whether a CSP has strict width \(l\) in time polynomial...
in the size of the domain and the constraints (for a fixed $l$), and give a necessary and sufficient criterion in terms of the existence of a near-unanimity operation of arity $l + 1$ which “preserves” the constraints of the CSP.

In trying to understand the set of CSPs which do not have bounded width, Feder and Vardi introduced the concept of the ability to count.

Definition 1.12. A CSP has the ability to count if there are elements 0, 1 in the domain and there are relations $C$, $Z$ in the library of constraints such that $C$ is ternary, $Z$ is unary, $(0, 0, 1), (0, 1, 0), (1, 0, 0) \in C$, 0 \in Z, and in any instance of the CSP which only uses the constraints $C, Z$, such that the constraints of the instance can be partitioned into two parts $A, B$ such that each variable of the instance shows up in exactly one constraint from $A$ and exactly one constraint from $B$, and such that $A$ contains exactly one more copy of the constraint $C$ than $B$ does] there is no solution to that instance.

Following an argument of Razborov for bipartite matching, Feder and Vardi prove the following.

Theorem 1.13 (Feder, Vardi [53]). Any CSP with the ability to count can't be solved by polynomial size monotone circuits. A CSP with the ability to count can never have bounded width.

They then make the following two outrageous conjectures.

Conjecture 1.1. Any CSP which can't “simulate” a CSP which has the ability to count does have bounded width.

Conjecture 1.2. Any CSP which can't “simulate” 1-IN-3 SAT can be solved in polynomial time.

Shockingly, despite seeming hopelessly vague and intractable, both of these conjectures were recently proven to be correct!

The examples of subgroup problems given above together with the concept of the ability to count also prompt the following question.

Problem 1.1. What is the largest possible generalization of the Gaussian elimination algorithm?

Feder and Vardi [53] made a first attempt at answering this by introducing the concept of near-subgroups of a group, and conjectured that they also lead to CSPs that could be solved in polynomial time. Using a result of Aschbacher [4], Feder [52] later succeeded in showing that near-subgroup problems can be solved in polynomial time.

In this case, however, they could have asked for more. Hubie Chen [39] studied the “expressive rate” of a constraint language $\Gamma$, which is defined as the function that takes $n$ to the logarithm of the number of $n$-variable relations which can be defined as solutions sets to CSPs over $\Gamma$. He observed that on a two element domain, this expressive rate always either grows as a polynomial or as an exponential function, and that the cases where it grows polynomially are exactly the cases where the class of relations which can be defined from $\Gamma$ is “polynomially learnable”. The same conjecture occurs in chapter 10 of Víctor Dalmau’s thesis [41], in an algebraic form.

Conjecture 1.3. For any constraint language $\Gamma$, the logarithm of the number of distinct $n$-variable relations which can be defined by primitive positive formulas over $\Gamma$ always either grows as a polynomial or an exponential function. In the case of polynomial growth this class of relations is efficiently learnable and the associated CSP can be solved in polynomial time.
This conjecture was solved via the theory of algebras with “few subpowers”, which classifies CSPs such that the solution sets always have “compact representations”, and gives general procedures for manipulating these compact representations.

In order to approach these questions, the key conceptual ingredient turned out to be a Galois duality from universal algebra, relating a family of relations to the set of operations which “preserve” the relations. This allows us to view CSPs as algebraic structures in disguise, and to use algebraic techniques to study the structure of their solution sets and to design algorithms. However, the algebraic structures we end up studying are much less structured than groups or lattices - they are in a sense the most basic algebraic structures that have any good properties at all.

The new framework was introduced by Jeavons [67], who reinterpreted an instance of a CSP as a homomorphism problem between relational structures.

Definition 1.14. An instance of the general combinatorial problem, or GCP, is a pair of relational structures $\langle A, B \rangle$ having the same signature (a relational structure is a set together with a family of named relations on that set, and the signature of a relational structure is a list of names of relations together with specifications of their arities). The question is whether there exists a homomorphism $A \to B$.

Example 1.9. Suppose that $T$ is a CSP template (in the sense of Feder and Vardi above), interpreted as a relational structure $(D, R_1, ..., R_n)$. To any instance of the CSP, we can associate a relational structure $X = (V, C_1, ..., C_n)$, where $V$ is the set of variables of the instance, and each $C_i$ is a list of those tuples of variables of $V$ which are required to satisfy the constraint $R_i$. Then a homomorphism of relational structures $X \to T$ is the same as an assignment of values in $D$ to each variable in $V$, such that each tuple of variables in each $C_i$ is mapped to an element of $R_i$. In other words, the GCP instance $\langle X, T \rangle$ is equivalent to the instance of CSP($T$) corresponding to $X$.

Jeavons also gives a few ways for other well-known problems (not CSPs) to be realized as instances of his general combinatorial problem.

Example 1.10. If $G$ is a graph and $K_q$ is a clique with $q$ vertices, then the GCP instance $\langle K_q, G \rangle$ is the $q$-CLIQUE problem. Note that in this case, the target of the homomorphism is the main variable, while the source stays fixed (aside from the parameter $q$).

Example 1.11. Let $G = (V, E)$ be a graph on $n$ vertices, and let $C_n = (W, F)$ be a cycle on $n$ vertices. Then the GCP instance $\langle (W, F, \neq), (V, E, \neq) \rangle$ is the problem of determining whether $G$ has a Hamiltonian circuit.

These other sorts of problems, where the target of the homomorphism varies arbitrarily and the source varies according to some parameter can be studied from the point of view of parametrized complexity and fixed parameter tractability. It turns out that hardness and easiness in this alternative setting is determined by the treewidth of the source structures [58]. We won’t discuss this research direction much.

After demonstrating the generality of the framework, Jeavons [67] restricts to studying homomorphism problems with a fixed target structure $T$. He calls this GCP($T$), where $\Gamma$ is the list of relations of $T$, but we will call it CSP($T$) in these notes. Note that this is the same problem as the CSP defined in the sense of Feder and Vardi above, but the instances are now treated as relational structures (which is useful notationally), and the new perspective in terms of homomorphisms gives a hint of a more algebraic approach. For instance, the homomorphism point of view prompts the following definition.

7
Definition 1.15. Two relational structures \( A, B \) with the same signature are homomorphically equivalent if there exist homomorphisms \( A \rightarrow B, B \rightarrow A \).

The homomorphism point of view now makes it obvious that if \( A \) and \( B \) are homomorphically equivalent, then \( \text{CSP}(A) \) and \( \text{CSP}(B) \) are equivalent problems - that is, a “yes” instance of one will always be a “yes” instance of the other. For instance, every bipartite graph \( H \) having at least one edge is homomorphically equivalent to the complete graph \( K_2 \) on two vertices, so if \( H \) is bipartite then the \( H \)-coloring problem is trivial.

Jeavons [67] points out that for a given CSP template, one can introduce new relations without changing the complexity of the CSP so long as these new relations are built out of the old relations in certain ways. Specifically, Jeavons shows that up to logspace reductions, we may as well assume that the collection of relations \( \Gamma \) contains the equality relation, and is closed under the following four operations:

- permutation of inputs,
- adding dummy variables (extra variables which are ignored by the relation),
- existential projection onto a subset of the variables, and
- intersection.

Note that any new relation which can be built out of these four operations can be viewed as the solution set to some instance of \( \text{CSP}(\Gamma) \), projected onto some subset of the variables. We can also think of the new relation as being defined by a primitive positive formula, that is, a formula built out of the existential quantifier \( \exists \), the relations \( R_i \) of \( \Gamma \) (and equality), and conjunctions \( \land \), but which does not involve negation, disjunction, implication, or universal quantification.

Example 1.12. The template we gave for HORN-SAT did not contain all possible Horn clauses - it stopped at the 3-ary Horn clause \( x \land y \implies z \). The 4-ary Horn clause \( x \land y \land z \implies w \) can be represented by the following primitive positive formula:

\[
\exists u \ (x \land y \implies u) \land (u \land z \implies w).
\]

The Horn clauses of higher arity can be represented by primitive positive formulas over HORN-SAT in a similar way.

Definition 1.16. A set of relations \( \Gamma \) on a fixed domain \( D \) is called a relational clone if it contains the equality relation, and is closed under permutations, adding dummy variables, projection, and intersections. Equivalently, a relational clone is a set of relations which is closed under defining new relations via primitive positive formulas.

The connection to algebra comes from the following fundamental result.

Theorem 1.17. There is a Galois duality between relational clones and clones. In particular, a relational clone is completely determined by its set of “polymorphisms”, that is, the set of functions that “preserve” all of the relations of \( \Gamma \).

In order to understand this result we must define clones, polymorphisms, and the concept of a function preserving a relation.
Definition 1.18. A set of functions $D^k \to D, k \in \mathbb{N}$ is called a clone if it contains the projections $\pi_i^k : D^k \to D$ which satisfy $\pi_i^k(x_1, \ldots, x_k) = x_i$ (generally the superscript $k$ is omitted when it is clear), and is closed under composition, the operation which takes a $k$-ary function $f$ and $k$ $l$-ary functions $g_1, \ldots, g_k$ to the function

$$(f \circ (g_1, \ldots, g_k)) : (x_1, \ldots, x_l) \mapsto f(g_1(x_1, \ldots, x_l), \ldots, g_k(x_1, \ldots, x_l)).$$

The reader should play with the above definition in order to convince themselves that every natural method of building new functions from old functions can be described in terms of the composition operation given above together with the projections $\pi_i^k$. For instance, the function $f(x, g(y, x))$ can be built out of $f$ and $g$ as follows:

$$(f \circ (\pi_1, g \circ (\pi_2, \pi_1)))(x, y) = f(x, g(y, x)).$$

Definition 1.19. A $k$-ary function $f$ is said to preserve an $m$-ary relation $R$, written $f \triangleright R$, if for every choice of $k$ $m$-tuples in $R$, applying $f$ componentwise produces a new $m$-tuple which is also in $R$. If we think of elements of $R$ as column vectors, we can write this as

$$\begin{bmatrix} x_{11} \\ \vdots \\ x_{1m} \end{bmatrix}, \ldots, \begin{bmatrix} x_{k1} \\ \vdots \\ x_{km} \end{bmatrix} \in R \implies f \left( \begin{bmatrix} x_{11} \\ \vdots \\ x_{1m} \end{bmatrix}, \ldots, \begin{bmatrix} x_{k1} \\ \vdots \\ x_{km} \end{bmatrix} \right) = \begin{bmatrix} f(x_{11}, \ldots, x_{k1}) \\ \vdots \\ f(x_{1m}, \ldots, x_{km}) \end{bmatrix} \in R.$$ 

A function $f$ is a polymorphism of a relational structure $(D, \Gamma)$ or of a relational clone $\Gamma$ if $f$ preserves $R_i$ for each relation $R_i \in \Gamma$.

The concept of preservation can be understood in two different ways. From the relational point of view, we have $f \triangleright R$ iff $f : D^k \to D$ is a homomorphism of relational structures $(D, R)^k \to (D, R)$, where $(D, R)^k$ is the categorical $k$th power of the relational structure $(D, R)$ (the $k$th power of $(D, R)$ has underlying set $D^k$ and relation $R^k$ given by listing all $m$-tuples of $k$-tuples such that the $m$-tuple of $i$th coordinates is in $R$ for each $i \leq k$). From the algebraic point of view, we have $f \triangleright R$ iff the subset $R \subseteq D^m$ is a subalgebra of the algebraic structure $(D, f)^m$, where $(D, f)^m$ is the categorical $m$th power of the algebraic structure $(D, f)$, where the basic operation is simply $f$ acting componentwise on $D^m$.

The Galois duality between relational clones and clones prompts a shift in one way of thinking about CSPs. Instead of studying a CSP template, one studies an algebraic structure whose operations are the polymorphisms of the original CSP template. Constraints that can be expressed in terms of the original library of relations become subalgebras of powers of this algebraic structure. Instances of a CSP become questions about whether intersections of various subalgebras of a power of the original algebra are empty or not.

Example 1.13. Suppose $\mathbb{A} = (\mathbb{Z}/p, f)$ is the algebraic structure with basic operation $f : (x, y, z) \mapsto x - y + z \pmod p$ for some prime $p$. Then a subalgebra of $\mathbb{A}^n$ - that is, a subset which is closed under $f$ - is exactly the same as an affine linear subspace of $(\mathbb{Z}/p)^n$ (recall that affine linear subspaces are like vector subspaces, but that they might not pass through the origin). Checking whether a collection of affine linear subspaces has a nonempty intersection is equivalent to solving a system of linear equations $\pmod p$.  

9
By using an old result classifying the minimal (nontrivial) clones on the domain \(\{0,1\}\) (under the Galois duality, a minimal clone of functions corresponds to a maximal relational clone), Jeavons [67] was able to give a new and relatively simple proof of Schaefer’s dichotomy theorem [103]. The algebraic structures corresponding to the basic polynomial time solvable problems are as follows.

**Example 1.14.** If \(\Gamma = (\{0\}, \{1\}, \{0,1\}^3 \setminus \{1,1,0\})\) is the template corresponding to HORN-SAT, then the clone of polymorphisms is generated by the function \(\text{min} : \{0,1\}^2 \to \{0,1\}\). This operation is an example of a semilattice operation.

**Example 1.15.** If \(\Gamma = (\leq, \neq)\) is the template corresponding to 2SAT, then the clone of polymorphisms is generated by the majority (or median) function \(\text{maj} : \{0,1\}^3 \to \{0,1\}\).

**Example 1.16.** If \(\Gamma = \{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\}, \neq\) is the template corresponding to XOR-SAT, then the clone of polymorphisms is generated by the affine linear function \((x,y,z) \mapsto x - y + z \pmod 2\) (this function is sometimes referred to as the minority function).

Early results focused on generalizing these basic examples, and developing the algebraic perspective further:

- If all polymorphisms of \(\Gamma\) are unary, then \(\text{CSP}(\Gamma)\) is NP-hard by a gadget reduction from NAE-SAT (if the domain has size 2) or \(k\)-coloring (if the domain has size \(k \geq 3\)).

- Generalized arc-consistency solves any CSP which has an associative, commutative, idempotent polymorphism. These types of operations were called ACI operations at the time, but are now generally referred to as semilattice operations.

- Later, Dalmau and Pearson [45] showed that generalized arc-consistency solves a CSP iff it has a “set” polymorphism, also known as a family of “totally symmetric” polymorphisms, where the output depends only on the set of inputs and not on their order or multiplicity.

- Already in Feder and Vardi’s work [53], it was shown that a CSP has strict width \(l\) iff it has an \(l+1\)-ary “near-unanimity” polymorphism, that is, an operation such that whenever all but one of the inputs are equal, their common value is the output. This fact is closely connected to a result in universal algebra known as the Baker-Pixley theorem [7].

- Bulatov and Dalmau [31] gave an algorithm generalizing Gaussian elimination as well as the algorithm for the general subgroup problem introduced by Feder and Vardi to the case of CSPs with a Mal’cev polymorphism \(m(x,y,z)\), that is, a polymorphism satisfying the identity \(m(x,y,y) = x = m(y,y,x)\) for all \(x,y\). In the case of groups such an operation is given by \((x,y,z) \mapsto xy^{-1}z\), but such operations also exist in quasigroups, making this a very wide generalization.

- One can restrict to the case where \(\Gamma\) contains a unary relation for every singleton subset of the domain. On the algebraic side, this corresponds to restricting to the case of idempotent algebras (that is, algebras where every singleton subset forms a subalgebra).

- At this point multiple authors started to realize that whether a CSP is hard or not doesn’t depend on the particular polymorphisms, but rather on the identities that are satisfied by the polymorphisms. One of the first papers to point this out was a paper by Bulatov and Jeavons [32] which also introduced a notion of polymorphisms for multisorted relations, as well as the use of “tame congruence theory” from universal algebra.
In particular, it was shown that if no finite subset of the identities satisfied by the polymorphisms imply that the polymorphisms can’t be unary, then CSP(Γ) is NP-hard by a gadget reduction from NAE-SAT or k-coloring. It was conjectured that this is an if and only if, that is, if a CSP has a family of polymorphisms that satisfy a nontrivial set of identities, then the CSP can always be solved in polynomial time.

The first big result was a comprehensive generalization of Gaussian elimination, generalizing the algorithm for Mal’cev operations as far as was reasonably possible. The basic idea here is to represent the solution space of all the constraints processed so far by giving a small generating set for that solution space, considered as a subalgebra of a power of the domain. In order for any algorithm along these lines to exist, there must first be a guarantee that every subalgebra of any power of the domain actually has a small generating set.

**Theorem 1.20 (Few Subpowers [20], [66]).** The following are equivalent for an algebraic structure $A$ on a finite domain:

- the number of subalgebras of $A^n$ grows like $2^{O(nk)}$ for some fixed $k$,
- every subalgebra of $A^n$ has a (nice) generating set of size $O(nk)$ for some fixed $k$, called a compact representation,
- $A$ has a $k$-edge term for some $k$, that is, a term satisfying the identity
  \[
  f \left( \begin{array}{cccc}
  y & y & x & x \\
  y & x & y & x \\
  x & x & x & y \\
  \vdots & \vdots & \vdots & \vdots \\
  x & x & x & y \\
  \end{array} \right) = \begin{array}{c}
  x \\
  x \\
  \vdots \\
  x \\
  \end{array},
  \]
  where all but the first column has exactly one $y$ and $k - 1$ $x$s, and
- a Gaussian-elimination-like algorithm solves CSP($A$) in polynomial time (the degree of the polynomial may depend on $k$).

Note that when $k = 2$, a $k$-edge term is the same as a Mal’cev operation. Additionally, if a $k$-edge term ignores its first input, then it is a $k$-ary near-unanimity term. So few subpowers algebras generalize both subgroup CSPs and CSPs with bounded strict width. There is still an important open question connected to few subpowers algebras, related to the following algebraic problem.

**Problem 1.2 (Subpower Membership Problem).** Given a finite subset $S \subseteq A^n$ and an element $x \in A^n$, determine if $x$ is in the subalgebra of $A^n$ generated by $S$.

A result of Kozik [77] shows that for general algebraic structures, this problem can be EXPTIME-complete. A recent result of Shriner [104] has upgraded this hardness result to most situations where algebras do not automatically have few subpowers, such as the case of congruence distributive algebras.

**Conjecture 1.4.** If $A$ has few subpowers, then the subpower membership problem can be solved in polynomial time.
Peter Mayr \[89\] has shown that this conjecture holds for nilpotent Mal’cev algebras of prime power order (and for expansions of such algebras). In a different direction, a recent result of Bulatov, Mayr, and Szendrei \[33\] has proved the conjecture in the special case that the algebra \( A \) is “residually small” (for those with a little universal algebra background, this means that every subdirectly irreducible algebra in the variety it generates has size bounded by some fixed cardinal - in the case of groups, it is equivalent to all Sylow subgroups being abelian). In the same paper, they also show that the subpower membership problem for algebras with few subpowers is always in NP. As far as I know, the above conjecture is still open even for the special case of quasigroups.

The second big result in this story was the classification of CSPs with bounded width, together with a surprising “collapse” of the bounded width hierarchy. The ideas used in the proof of this result - especially the theory of absorbing subalgebras - led to a number of breakthrough results in universal algebra.

**Theorem 1.21 (Bounded Width Algebras \[35\], \[10\], \[15\], \[83\], \[16\], \[79\], \[78\]).** For an idempotent algebra on a finite domain, the following are equivalent:

- \( \text{CSP}(A) \) has bounded width,
- \( \text{CSP}(A) \) can’t simulate any CSP which has the ability to count,
- \( \text{CSP}(A) \) has width \( (2,3) \),
- \( \text{CSP}(A) \) can be solved by a “cycle consistency” algorithm, which checks that every “cycle” of constraints has a valid solution extending each possible value of every variable in the cycle,
- \( A \) has terms \( f, g \) of arity 3 satisfying the identities
  \[
  g(x,x,y) = g(x,y,x) = g(y,x,x) = f(x,x,y) = f(x,y,x) = f(x,y,y),
  \]
- \( \text{CSP}(A) \) can be “robustly” solved by the basic semidefinite programming relaxation (i.e., if an \( \epsilon \)-portion of the constraints are garbled, then the basic SDP can be used to find an assignment that satisfies all but an \( f(\epsilon) \)-portion of the constraints, where \( f(\epsilon) \to 0 \) as \( \epsilon \to 0 \)).

Furthermore, there is a polynomial time algorithm for checking whether a relational structure (which has all unary singleton relations) has bounded width, and to find terms \( f, g \) as in the above theorem if it does. This algorithm leverages the fact that the canonical width-(2,3) Datalog program will correctly solve any instance of any bounded width CSP in polynomial time, and constructs a CSP whose solution corresponds to a pair of polymorphisms satisfying a nice set of identities. A similar algorithm is not known to exist for checking that a CSP has width 1.

The third big result of the subject was a nice classification of the algebras which were conjectured to correspond to CSPs with polynomial time algorithms. This result was proved with the absorption theory that had been developed for the study of bounded width algebras.

**Theorem 1.22 (Cyclic Terms for Taylor Algebras \[14\]).** For an algebraic structure \( A \) on a finite domain, the following are equivalent:

- there is a finite set of identities satisfied by the operations of \( A \) which can’t be satisfied by essentially unary functions,
• for every prime \( p > |A| \), \( A \) has a “cyclic term” \( f \) of arity \( p \), that is, a term which satisfies the identity
  \[
  f(x_1, \ldots, x_{p-1}, x_p) = f(x_2, \ldots, x_p, x_1),
  \]
• \( A \) has a 4-ary term \( t \) which satisfies the identity
  \[
  t(x, x, y, z) = t(y, z, z, x).
  \]

With this in hand, the main conjecture of the subject was finally possible to state simply:
“CSP(\( A \)) is in P iff \( A \) has a cyclic term.”

The fourth big result of the subject concerns a generalization of CSPs to “valued constraints”.

**Definition 1.23.** A *valued constraint* on \( k \) variables is a cost function from \( D^k \) to \(( -\infty, \infty ] \). An instance of a valued CSP (abbreviated VCSP) consists of a sum of valued constraints applied to various tuples of the variables, possibly with nonnegative coefficients. The goal is to minimize the sum of the cost functions. The associated CSP to a VCSP is the problem of finding an assignment that makes all of the costs finite.

The Galois duality between clones and relational clones can be generalized to a duality between VCSP templates and “fractional polymorphisms” - essentially just formal convex combinations of ordinary polymorphisms, with the property that when they are applied to any tuple of elements of \( D^k \), on average they decrease the cost assigned by any cost function from the VCSP template.

The standard example of a valued constraint with an interesting fractional polymorphism is a submodular function, that is, a cost function \( c \) defined on a lattice (or a power of a lattice) which satisfies the inequality
  \[
  \frac{1}{2}c(A) + \frac{1}{2}c(B) \geq \frac{1}{2}c(A \lor B) + \frac{1}{2}c(A \land B).
  \]

It is well known that submodular cost functions can be minimized using a linear programming relaxation.

**Theorem 1.24** (VCSP Dichotomy [80], [76]). If a VCSP has its associated CSP in P and has a cyclic fractional polymorphism, then by using the algorithm for the associated CSP as a black box to get the set of “feasible” values for each variable and applying the basic linear programming relaxation to the restriction of the VCSP to the feasible values we get the minimum cost solution. If the VCSP has no cyclic fractional polymorphisms, then it is NP-hard.

Finally, the biggest result of all was recently proved independently by Bulatov and by Zhuk.

**Theorem 1.25** (CSP Dichotomy [37], [112]). A finite algebra \( A \) has a cyclic term iff CSP(\( A \)) is in P.

A major open problem is whether one can test if CSP(\( \Gamma \)) is in P in time polynomial in the size of the description of the constraints of \( \Gamma \) (given that \( \Gamma \) contains all singleton unary relations). Zhuk tells me that he conjectures it to be NP-hard to test for the existence of cyclic terms. If so, perhaps this could lead to a new form of public key cryptography, where the private key is a cyclic term, and the public key is a CSP template which is preserved by that cyclic term...

The story has not ended with the proof of the main conjecture. There are at least six interesting research directions that are still being actively investigated: qualitative CSPs, counting complexity,
promise problems, quantified CSPs, “hybrid” tractability (combining restrictions on both the source and the target relational structures), and planar CSPs.

Qualitative CSPs come from allowing the domain of the CSP to be infinite. Of course, this immediately leads to problems, for instance, how can one even specify a set of relations on an infinite domain? The idea, capturing good old fashioned AI intuition about “qualitative” reasoning, is to require that the specific values of the solutions are not important, just the qualitative relationships between them. To make this more precise, we require our domain (and the relations on it) to have a very large automorphism group.

**Definition 1.26.** A permutation group $G$ acting on a set $S$ is oligomorphic if $S^n$ has finitely many $G$-orbits for every $n \geq 1$.

A standard example of an oligomorphic group is the group of order-preserving bijections on the rational numbers. Relations invariant under this group give rise to “temporal” CSPs, where the goal is to find some assignment of variables to times satisfying constraints about their relative ordering.

Bodirsky, in his habilitation thesis \[26\] introduced qualitative CSPs and gave a number of classification results. Before beginning a classification, he first chooses an oligomorphic group $G$ acting on a countable set $S$. He then uses results from model theory (specifically, $\omega$-categorical theories and Fraïssé limits) as well as structural Ramsey theory (and the theory of extremely amenable groups) to understand the relations which are invariant under $G$ and their polymorphisms, and for several groups $G$ he succeeds in finding a complete classification of problems into “easy” and “hard”. The main three cases considered by Bodirsky \[26\] are the following:

- the automorphism group of $\langle \mathbb{Q}, < \rangle$, corresponding to temporal CSPs,
- the automorphism group of the random graph, for which he proves “Schaefer’s Theorem for graphs” (such CSPs can be interpreted as problems where the variables correspond to decisions about whether certain pairs of vertices of an unknown graph are connected by an edge or not), and
- the automorphism group of an infinite branching tree structure $\langle L, \mid \rangle$, where $\mid$ is a 3-ary relation where $ab\mid c$ means that the youngest common ancestor of $a, b$ lies below the youngest common ancestor of $b, c$ - the invariant relations correspond to “branching time constraints”, or “phylogeny constraints”, and the associated CSPs could in principle be of interest to biologists.

Recent results on QCSPs indicate that the difficulty of the classification results seems to be related to the orbit growth function of the oligomorphic group $G$, which takes $n$ to the number of orbits of $n$-tuples under $G$ \[27\]. For sufficiently small orbit growth functions, a dichotomy result has been proven (using the finite case as a black-box). The main conjecture in this field is the following somewhat technical statement.

**Conjecture 1.5** \[19\]. Let $A$ be the core of a reduct of a finitely bounded homogeneous structure. Then CSP($A$) is in P iff $A$ has a 6-ary polymorphism $s$ and unary polymorphisms $\alpha, \beta$ satisfying the “pseudo-Siggers” identity:

$$\alpha \circ s(x, y, x, z, y, z) = \beta \circ s(y, x, z, x, z, y).$$

Otherwise, CSP($A$) is NP-complete.
The most recent development in the study of CSPs is the study of promise problems. Promise problems are similar in spirit to approximation algorithms, but much more amenable to an algebraic approach. A promise problem is defined here to be a pair of problems, one more restrictive than the other, where the goal is to give an algorithm which correctly says “yes” if the less restrictive problem has a solution and says “no” if the more restrictive problem has no solution (if neither case holds, any output is allowable).

**Definition 1.27.** If $A, B$ are relational structures such that a homomorphism $A \to B$ exists, then $PCSP(A, B)$ is the following problem. The input is a relational structure $C$ s.t. there exists a homomorphism $C \to A$ (the promise), although this map is not revealed to us. The desired output is a homomorphism from $C$ to $B$.

A typical strategy for proving tractability of $PCSP(A, B)$ is to find a relational structure $X$ such that there exist homomorphisms $A \to X \to B$ and such that $CSP(X)$ is in $P$.

**Example 1.17.** Let $A$ be 1-IN-3 SAT and let $B$ be NAE-SAT (where the 1-IN-3 relation and the NAE relation have the same name in the signature). The identity map on the domain gives a homomorphism $A \to B$ since the 1-IN-3 relation is contained in the NAE relation. Although both problems are NP-complete, the PCSP associated to the pair is tractable: let $X = (\mathbb{Z}, x + y + z = 1)$, note that the inclusion map $A \to X$ is a homomorphism, and that the map $sgn : \mathbb{Z} \to \{0, 1\}$ given by

$$
sgn(x) = \begin{cases} 
0 & x \leq 0 \\
1 & x \geq 1 
\end{cases}
$$

defines a homomorphism $X \to B$. The CSP associated to $X$ is tractable (even though it is defined over an infinite domain), since it boils down to solving a system of linear equations over the integers. It is not possible to find a finite relational structure $X$ with polynomial time CSP that fits between 1-IN-3 SAT and NAE-SAT (see [8] for a proof).

The relevant algebraic object in this context is $Pol(A, B)$, the set of homomorphisms $A^k \to B$. However, one can still write down “minor identities” between the functions in $Pol(A, B)$ such as $f(x, x, y) = g(y, x)$, and compare the set of minor identities obtained to the identities that occur in polymorphism algebras of tractable CSPs. This approach of studying minor identities has been surprisingly useful, and has led to the proposal to call sets of functions such as $Pol(A, B)$ “minions” (a competing proposed name is “clonoid”).

Unlike the situation for CSPs, it is still quite hard to prove hardness results for PCSPs. The following basic problem is still wide open.

**Conjecture 1.6.** For any $k \geq l \geq 3$, the promise problem $PCSP(K_l, K_k)$ is NP-hard (this problem is the problem of $k$-coloring a graph which is promised to be $l$-colorable).

One of the first results in this direction concerned a PCSP called $(2 + \epsilon)$-SAT, where one is given clauses of $2k + 1$ variables and wants to satisfy the associated instance of SAT given the promise that it is possible to find an assignment in which every clause has at least $k$ satisfied literals. The $(2 + \epsilon)$-SAT problem was proven to be NP-hard [10], and this result was slightly generalized and put into the PCSP framework in [28].

Recent results in the study of PCSPs include a result of Barto, Bulín, Opršal, and Krokhin [8] in which they used minion techniques to show that $PCSP(K_d, K_{2d-1})$ is NP-hard for every $d \geq 3$.  

15
reducing from the hypergraph promise problem PCSP(NAE₂, NAEₖ). The hypergraph coloring problem PCSP(NAE₂, NAEₖ) was itself shown to be hard via a reduction from a variant of the PCP theorem [8]. The big result in PCSPs is the following result which connects computational complexity to height 1 identities satisfied by the minion of polymorphisms.

**Theorem 1.28** (Barto, Bulín, Opršal, Krokhin [8]). If there is a “minion homomorphism” from Pol(A₁, B₁) to Pol(A₂, B₂), then PCSP(A₂, B₂) has a logspace reduction to PCSP(A₁, B₁).

**Remark 1.1.** For those who like category theory, an abstract minion is just a covariant functor from the category of (finite) sets to the category of sets, and a minion homomorphism is just a natural transformation of functors. We could say that a “representation” of an abstract minion over A, B is a natural transformation to the functor I ↦ Hom(A¹, B).

PCSP(A, B) ends up being logspace equivalent to the problem of distinguishing between diagrams in the category of sets of size at most N (N any fixed large enough number) which have a nonempty limit (“yes” instances), and diagrams such that the image under the minion Pol(A, B) has an empty limit (“no” instances) (this is the “promise satisfaction of a minor condition” problem of [8]).

## 2 The Inv-Pol Galois connection

We begin by recalling some definitions from the introduction.

**Definition 2.1.** A set of relations Γ on a fixed domain D is called a **relational clone** if it contains the equality relation, and is closed under permutations, adding dummy variables, existential projection, and intersections. Equivalently, a relational clone is a set of relations which is closed under defining new relations via primitive positive formulas.

**Definition 2.2.** A set of functions Dᵏ → D, k ∈ ℕ is called a **clone** if it contains the projections πᵏᵢ : Dᵏ → D which satisfy πᵏᵢ(₁, ..., xₖ) = xᵢ (generally the superscript k is omitted when it is clear), and is closed under composition, the operation which takes a k-ary function f and k l-ary functions g₁, ..., gₖ to the function

(f ◦ (g₁, ..., gₖ)) : (x₁, ..., xₖ) ↦ f(g₁(x₁, ..., xₖ), ..., gₖ(x₁, ..., xₖ)).

**Definition 2.3.** A k-ary function f is said to **preserve** an m-ary relation R, written f ⪰ R, if for every choice of k m-tuples in R, applying f componentwise produces a new m-tuple which is also in R. If we think of elements of R as column vectors, we can write this as

\[
\begin{pmatrix}
  x_{11} \\
  \vdots \\
  x_{1m}
\end{pmatrix}
, \ldots,
\begin{pmatrix}
  x_{k1} \\
  \vdots \\
  x_{km}
\end{pmatrix} \in R \implies f\left(
\begin{pmatrix}
  x_{11} \\
  \vdots \\
  x_{1m}
\end{pmatrix}
, \ldots,
\begin{pmatrix}
  x_{k1} \\
  \vdots \\
  x_{km}
\end{pmatrix}
\right) =
\begin{pmatrix}
  f(x_{11}, ..., x_{k1}) \\
  \vdots \\
  f(x_{1m}, ..., x_{km})
\end{pmatrix} \in R.
\]

A function f is a **polymorphism** of a relational structure (D, Γ) or of a relational clone Γ if f preserves Rᵢ for each relation Rᵢ ∈ Γ.

We can write the condition for f ⪰ R more compactly as M ∈ Rᵏ \implies f(M) ∈ R, where M ∈ Rᵏ means that M is a matrix with k columns, each of which belongs to R, and f(M) is the column vector obtained by applying f to the rows of M.

In order to state the Galois connection, we need a few additional definitions.
**Definition 2.4.** If $\Gamma$ is any set of relations on a domain $D$, then we define $\langle \Gamma \rangle$ to be the relational clone generated by $\Gamma$ (that is, $\langle \Gamma \rangle$ is the smallest relational clone which contains $\Gamma$). Similarly, if $\mathcal{O}$ is any set of operations on $D$, we define $\langle \mathcal{O} \rangle$ to be the clone generated by $\mathcal{O}$. If $\mathcal{A} = (D, \mathcal{O})$ is an algebraic structure, we let $\text{Clo}(\mathcal{A})$ be the clone generated by the basic operations of $\mathcal{A}$, so $\text{Clo}(\mathcal{A}) = \langle \mathcal{O} \rangle$.

**Definition 2.5.** If $\Gamma$ is any set of relations on a domain $D$, then we define $\text{Pol}(\Gamma)$ to be the set of operations on $D$ that preserve every relation of $\Gamma$. If $\mathcal{O}$ is any set of operations on $D$, we define $\text{Inv}(\mathcal{O})$ to be the set of relations which are preserved by every operation in $\mathcal{O}$. If we want to restrict to operations or relations of a specific arity, we use the notations $\text{Pol}_k(\Gamma) = \{ f : D^k \to D \mid \forall R \in \Gamma, f \triangleright R \}$, $\text{Inv}_m(\mathcal{O}) = \{ R \subseteq D^m \mid \forall f \in \mathcal{O}, f \triangleright R \}$.

It is worth thinking about what sort of information about an algebraic structure $(D, \mathcal{O})$ can be found in $\text{Inv}(\mathcal{O})$.

**Example 2.1.** If $\mathcal{A} = (D, \mathcal{O})$ is an algebraic structure, then $\text{Inv}_2(\mathcal{O})$ determines (among other things)

- the lattice of subalgebras of $\mathcal{A}$,
- $\text{Aut}(\mathcal{A})$, the automorphism group of $\mathcal{A}$,
- $\text{End}(\mathcal{A})$, the semigroup of endomorphisms of $\mathcal{A}$,
- $\text{Con}(\mathcal{A})$, the lattice of congruences on $\mathcal{A}$,
- the set of partial orders on $D$ which are compatible with the operations of $\mathcal{A}$, and
- $\text{Inv}_2(\mathcal{B})$ for any subalgebra $\mathcal{B} \subset \mathcal{A}$ or quotient $\mathcal{B} = \mathcal{A}/\sim$.

It is easy to see that for all $\Gamma$, $\text{Pol}(\Gamma)$ will be a clone, and that for all $\mathcal{O}$, $\text{Inv}(\mathcal{O})$ will be a relational clone. As a consequence, we have $\langle \Gamma \rangle \subseteq \text{Inv}(\text{Pol}(\Gamma))$ and $\langle \mathcal{O} \rangle \subseteq \text{Pol}(\text{Inv}(\mathcal{O}))$. The next two results show that these inclusions are actually equalities.

Before diving into the proof, the following concrete example will be useful for understanding the notation. Consider what it means for a ternary function $f$ to preserve the binary relation $\leq$ (functions which preserve $\leq$ are often called monotone). Since $0 \leq 0$, $0 \leq 1$, and $1 \leq 1$, we have

\[
\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \in \leq \implies \begin{bmatrix} f(0,0,1) \\ f(0,1,1) \end{bmatrix} \in \leq,
\]

that is, $f(0,0,1) \leq f(0,1,1)$. It’s convenient to abbreviate the above as follows:

\[
\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \in \leq^3 \implies f \left( \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \right) \in \leq.
\]

**Theorem 2.6.** If $\Gamma$ is a set of relations on a finite domain $D$, then $\text{Inv}(\text{Pol}(\Gamma)) = \langle \Gamma \rangle$. In fact, if a relation $S \subseteq D^m$ is preserved by $\text{Pol}(\Gamma)$ and can be generated by $k$ elements of $D^m$ (using operations of $\text{Pol}(\Gamma)$), then $S$ can be defined by a primitive positive formula over $\Gamma$ which involves at most $|D|^k$ auxiliary variables.
Suppose that $S$ is generated by elements $x_1, \ldots, x_k \in D^m$, and let $X$ be the matrix having the $x_i$s as columns. Then $S = \{f(X) \mid f \in Pol_k(\Gamma)\}$, so as a starting point we will construct a primitive positive formula $\Phi$ that describes $Pol_k(\Gamma)$.

Note that $D^{D_k}$ is naturally interpreted as the set of functions $f : D^k \to D$: if $f \in D^{D_k}$, then the $(a_1, \ldots, a_k)$-coordinate of $f$ is $f(a_1, \ldots, a_k)$. We can now give a positive primitive formula for $Pol_k(\Gamma) \subseteq D^{D_k}$:

$$\Phi(f) := \bigwedge_{R \in \Gamma} \bigwedge_{M \in R^k} f(M) \in R.$$  

If $\Gamma$ is infinite, the outer $\bigwedge$ will be an infinite conjunction. However, since there are only finitely many possible subsets of $D^{D_k}$, some finite subset $\Phi'$ of the inner conjunctions will define the same subset of $D^{D_k}$.

Finally, to define $S$ we use the primitive positive formula

$$S(a) := \exists f \in D^{D_k} \Phi'(f) \land (f(X) = a). \tag*{\Box}$$

**Theorem 2.7.** If $\mathcal{O}$ is a set of operations on a finite domain $D$, then $Pol(Inv(\mathcal{O})) = \langle \mathcal{O} \rangle$.

**Proof.** Suppose that $f \in Pol(Inv(\mathcal{O}))$ is a $k$-ary function. Let $\mathcal{F}(k) \subseteq D^{D_k}$ be the subalgebra of the algebraic structure $(D, \mathcal{O})^{D_k}$ generated by the functions $\pi_i : D^k \to D$, $\pi_i(x_1, \ldots, x_k) = x_i$. Then $\mathcal{F}(k)$, interpreted as a set of functions from $D^k$ to $D$, is exactly the set of $k$-ary functions in $\langle \mathcal{O} \rangle$.

Since $f \in Pol(Inv(\mathcal{O}))$ and $\mathcal{F}(k) \in Inv(\mathcal{O})$, we must have $f \upharpoonright \mathcal{F}(k)$, so in particular we must have $f(\pi_1, \ldots, \pi_k) \in \mathcal{F}(k)$. But $f(\pi_1, \ldots, \pi_k)$ is exactly $f$ thought of as an element of $D^{D_k}$, so this means that $f \in \langle \mathcal{O} \rangle$. \tag*{\Box}

**Corollary 2.8.** There is an order reversing bijection between clones and relational clones, given by the operations $Inv$ and $Pol$.

**Remark 2.1.** The map $\{1, \ldots, k\} \to D^{D_k}$ given by $i \mapsto \pi_i$, where $\pi_i : D^k \to D$ is given by $\pi_i(x_1, \ldots, x_k) = x_i$, shows up in the theory of approximation algorithms as the long code, which is the longest way of encoding $\{1, \ldots, k\}$ over the alphabet $D$ that doesn’t have any redundant coordinates.

**Example 2.2.** In the next section we will prove the following three correspondences between clones and relational clones on the domain $\{0, 1\}$:

- $\langle 2\text{SAT} \rangle = \langle \subseteq, \neq \rangle$ corresponds to $\langle \text{maj} \rangle$ (the majority function on three inputs),
- $\langle \text{HORN-SAT} \rangle = \langle \{0\}, \{1\}, x \land y \Rightarrow z \rangle$ corresponds to $\langle \text{min} \rangle$ (the minimum function on two inputs), and
- $\langle \text{XOR-SAT} \rangle = \langle \{1\}, x + y + z \equiv 0 \pmod{2} \rangle$ corresponds to $\langle x - y + z \pmod{2} \rangle$.

**Definition 2.9.** If $\mathcal{A}, \mathcal{A}'$ are two algebraic structures on the same domain such that every basic operation of $\mathcal{A}'$ is in $\text{Clo}(\mathcal{A})$, then we say that $\mathcal{A}'$ is a reduct of $\mathcal{A}$ and that $\mathcal{A}$ is an expansion of $\mathcal{A}'$. If $\text{Clo}(\mathcal{A}) = \text{Clo}(\mathcal{A}')$, then $\mathcal{A}$ and $\mathcal{A}'$ are called term equivalent.

The lattice of clones on the domain $\{0, 1\}$ has been completely described - it has countably many elements, and is known as Post’s lattice [98] (see also chapter II.3 of [84]). It is known that on a domain of size $\geq 3$, there are uncountably many clones [110], [111]. In particular, we see that most clones and relational clones can’t be generated by finitely many functions or relations.
Definition 2.10. A clone $O$ is said to be finitely generated if there is a finite set $S$ of operations such that $O = \langle S \rangle$. It is said to be finitely related if there is a finite set of relations $\Gamma$ such that $O = \text{Pol}(\Gamma)$.

Example 2.3. The clone on $\{0, 1\}$ generated by the binary implication function $\to$, given by $\to(x, y) = \neg x \lor y$, is finitely generated but not finitely related. One quick way to prove this is to show that for every $n \geq 3$, the $n$-ary threshold function $t^2_n$ defined by

$$t^2_n(x_1, ..., x_n) = \begin{cases} 1 & \sum_i x_i \geq 2 \\ 0 & \sum_i x_i \leq 1 \end{cases}$$

is not in $\langle \to \rangle$, but every way of identifying two coordinates of $t^2_n$ gives a function which is in $\langle \to \rangle$ (exercise: why does this prove that $\langle \to \rangle$ can not be finitely related?). $\text{Inv}(\to)$ is generated by the infinite sequence of relations $R_1, R_2, ...$ given by $R_n = \{0, 1\}^n \setminus \{(0, ..., 0)\}$, and $\langle \to \rangle$ consists of all functions of the form $f(x_1, ..., x_n) \lor x_i$.

In general, it is a very difficult problem to determine if a given finitely generated clone is finitely related, or vice versa.

Remark 2.2. The Galois connection between relational clones and clones on a finite set was originally discovered by Geiger in 1968 [56], and Reinhard Pöschel investigated the general case (where the domain may be infinite) in [97] - in the infinite case, the main difference is that clones must also be taken to be closed in the topology of pointwise convergence. (Jeavons reproved one direction of the connection - that relational clones on a finite domain are determined by their polymorphisms - in [67].)

Remark 2.3. The Galois connection presented here, between operations and relations, can be viewed as being induced by the two-sorted preservation relation $\succ$. In general, whenever one has a two-sorted binary relation $R$ on a pair of sets $A, B$, one can define operations $F, G$ on the power sets of $A, B$ respectively by

$$F(S) = \{b \in B \mid \forall a \in S \ a R b\},$$
$$G(T) = \{a \in A \mid \forall b \in T \ a R b\}.$$ 

The abstract order-theoretic properties of such a pair $F, G$ are

- $F$ and $G$ are antitone: $S \subseteq S' \implies F(S) \supseteq F(S')$, and similarly for $G$, and
- for any $S \in \mathcal{P}(A)$ and $T \in \mathcal{P}(B)$, we have

$$T \subseteq F(S) \iff S \subseteq G(T).$$

Actually the first property listed is redundant, as we have

$$(S \subseteq S') \land (F(S') \subseteq F(S')) \implies S \subseteq S' \subseteq G(F(S')) \implies F(S') \subseteq F(S),$$

and either of $F, G$ is determined by the other together with the second property: $F(S) = \bigcup_{S \subseteq G(T)} T$. Additionally, the second property follows from the first property together with $S \subseteq G(F(S))$ and $T \subseteq F(G(T))$: for one direction, we have

$$T \subseteq F(S) \implies S \subseteq G(F(S)) \subseteq G(T).$$
Any such pair $F,G$ determines the binary relation $R$, since
\[(a,b) \in R \iff b \in F\{a\} \iff a \in G\{b\},\]
and $F,G$ are both determined by the second property and their restrictions to singletons, since
\[b \in F(S) \iff \forall a \in S, \ a \in G\{b\} \iff \forall a \in S, \ b \in F\{a\}.\]

Then one can define closure operators $G \circ F, F \circ G$ on subsets of $A$ and $B$. When we say these are “closure operators”, we mean that the images of these operators form collections of “closed” sets, such that any intersection of closed sets is closed, and for $S \subseteq A$, $G \circ F(S)$ is equal to the smallest closed set which contains $S$. All of these properties are easy to show directly in terms of the binary relation $R$, but they can also be proved order theoretically.

For the order theoretic proof, note that
\[F(S) \subseteq F(S) \implies S \subseteq G \circ F(S),\]
and similarly for $F \circ G$, and so we have
\[F(S) \subseteq F \circ G(F(S)) = F(G \circ F(S)) \subseteq F(S),\]
and we see that a set in $X \subseteq B$ is closed iff it is of the form $F(S)$ for some $S \subseteq A$. For the intersection property, note that
\[S \subseteq G(X) \cap G(Y) \iff X \cup Y \subseteq F(S) \iff S \subseteq G(X \cup Y),\]
and for the characterization of the closure of $S$ we have
\[G \circ F(S) \subseteq G(Y) \iff Y \subseteq F \circ G \circ F(S) = F(S) \iff S \subseteq G(Y).\]

Then $F$ and $G$ will provide a Galois correspondence between the closed subsets of $A$ and the closed subsets of $B$. The nontrivial thing to do is to describe the closure operators explicitly.

In our case, the relation $R$ was given by $\triangleright$, and the sets $A,B$ were the sets of operations and relations on a given domain. Our main difficulty was in proving that the closure operators $G \circ F = \text{Pol} \circ \text{Inv}$ and $F \circ G = \text{Inv} \circ \text{Pol}$ were concretely described by the closure operators $\mathcal{O} \mapsto \langle \mathcal{O} \rangle$ for clones and $\Gamma \mapsto \langle \Gamma \rangle$ for relational clones, respectively. In ordinary Galois theory, the sets $A,B$ are taken to be a field and a group of automorphisms of the field, and the relation $R$ determines whether a given element of the field is fixed by a given automorphism (exercise: find the corresponding closure operations).

## 3 Three basic examples

We start with the correspondence between 2SAT and majority.

**Theorem 3.1.** Suppose that a relation $R \subseteq \{0,1\}^m$ is preserved by the majority function $\text{maj} : \{0,1\}^3 \to \{0,1\}$. Then $R$ is bijunctive, that is, $R$ can be written as a conjunction of binary and unary relations.
Proof. We prove this by induction on $m$. If $m \leq 2$ then there is nothing to prove. Otherwise, for each $i \leq 3$ let $R_i$ be the existential projection of $R$ onto all variables except for the $i$th. We will show that $R$ is equivalent to

$$\Phi(x_1, \ldots, x_m) := R_1(x_2, x_3, \ldots, x_m) \land R_2(x_1, x_3, \ldots, x_m) \land R_3(x_1, x_2, \ldots, x_m),$$

and the result will then follow by the induction hypothesis. It is clear that $R \subseteq \Phi$, so suppose $(x_1, \ldots, x_m) \in \Phi$. Then by the definitions of $R_1, R_2, R_3$, there exist $x'_1, x'_2, x'_3$ such that $(x'_1, x_2, x_3, \ldots, x_m), (x_1, x'_2, x_3, \ldots, x_m), (x_1, x_2, x'_3, \ldots, x_m) \in R$, and applying $\text{maj}$ to these three tuples we see that $(x_1, \ldots, x_m) \in R$ as well. \hfill \qed

**Definition 3.2.** An operation $f : \{0,1\}^k \to \{0,1\}$ is called monotone if it preserves the relation $\leq$. It is called self-dual if it preserves the relation $\neq$.

**Theorem 3.3.** Suppose that a function $f : \{0,1\}^k \to \{0,1\}$ is monotone and self-dual. Then $f \in \langle \text{maj} \rangle$.

**Proof.** We prove this by induction on $k$. It’s easy to check that there are no monotone self-dual functions of arity $\leq 2$ other than the projections, so assume that $k \geq 3$. By the induction hypothesis, any function we can make by identifying two variables of $f$ is in $\langle \text{maj} \rangle$. We claim that we have

$$f(x, y, z, \ldots) = \text{maj}(f(x, y, y, \ldots), f(z, y, z, \ldots), f(x, x, z, \ldots)),$$

where the ... always represent the remaining $k - 3$ variables. To see this, note that the formula is trivially true when $x = y = z$, so we only need to check it when one of the variables is different from the other two. We will check it in the case $(x, y, z) = (0, 1, 0)$, since every other case is analogous (via cyclically permuting $x, y, z$ or swapping 0s and 1s throughout). In this case, since $f$ is monotone we have

$$f(0, 1, 1, \ldots) \geq f(0, 1, 0, \ldots) \geq f(0, 0, 0, \ldots),$$

so the majority of these three values will be $f(0, 1, 0, \ldots) = f(x, y, z, \ldots)$. \hfill \qed

Monotone self-dual functions can be interpreted as voting functions. They also have a combinatorial interpretation in terms of maximal “intersecting families” of sets.

**Definition 3.4.** Let $S$ be a set. A family $\mathcal{F} \subseteq \mathcal{P}(S)$ is called an intersecting family of subsets of $S$ if $A, B \in \mathcal{F}$ implies $A \cap B \neq \emptyset$.

**Proposition 3.5.** An intersecting family of subsets of a set $S$ is maximal (with respect to containment) if and only if for every set $A \subseteq S$ we have either $A \in \mathcal{F}$ or $(S \setminus A) \in \mathcal{F}$. For every $n \geq 1$ there is a bijection between maximal intersecting families $\mathcal{F}$ of subsets of $\{1, \ldots, n\}$ and monotone self-dual boolean functions $f : \{0,1\}^n \to \{0,1\}$.

We can describe a maximal intersecting family of subsets of a set more compactly by describing its collection of minimal elements. We can mutate an intersecting family by taking one of its minimal elements $A$, deleting it, and replacing it with its complement - this is called “switching” the subset $A$ with its complement.

**Definition 3.6.** For every $n$, we define an undirected graph $\mathcal{M}_n$ whose vertices are the maximal intersecting families of subsets of $\{1, \ldots, n\}$, and whose edges are the pairs of families $\mathcal{F}, \mathcal{G}$ such that $|\mathcal{F} \setminus \mathcal{G}| = 1$. 21
Theorem 3.7. Suppose that a relation \( R \subseteq \{0, 1\}^m \) is preserved by the minimum function \( \min : \{0, 1\}^2 \rightarrow \{0, 1\} \). Then \( R \) can be written as a conjunction of Horn clauses.

Proof. Write \( R = \bigwedge_i C_i \) in conjunctive normal form, such that each clause \( C_i \) is minimal. Note that this means that for each literal \( l \in C_i \), there is some assignment to the variables that satisfies \( R \), and has the rest of the literals in \( C_i \) other than \( l \) set to 0.

Suppose, for a contradiction, that some clause \( C_i \) has at least two non-negated variables in it, and assume without loss of generality that \( C_i = x_1 \lor \cdots \lor x_p \lor \bar{x}_{p+1} \lor \cdots \lor \bar{x}_{p+q} \), \( p \geq 2 \). By the minimality of \( C_i \), there are assignments \( a, b \) which satisfy \( R \) and such that \( a_2 = \cdots = a_p = \bar{a}_{p+1} = \cdots = \bar{a}_{p+q} = 0 \) and \( b_1 = \cdots = b_{p-1} = \bar{b}_{p+1} = \cdots = \bar{b}_{p+q} = 0 \). But then \( \min(a, b) \) fails to satisfy \( C_i \), and hence fails to satisfy \( R \). \( \square \)

Theorem 3.8. Suppose that a function \( f : \{0, 1\}^k \rightarrow \{0, 1\} \) preserves the relations \( \{0\}, \{1\} \), and \( x \land y \implies z \). Then there is a nonempty subset \( I \subseteq \{1, \ldots, k\} \) such that \( f(x_1, \ldots, x_k) = \min_{i \in I} x_i \).
Proof. Since \( f \) preserves \( \{0\} \) and \( \{1\} \), we have \( f(0,\ldots,0) = 0 \) and \( f(1,\ldots,1) = 1 \). Since \( \leq \) is in the relational clone generated by \( x \land y \implies z \), \( f \) must be monotone.

For each subset \( I \subseteq \{1,\ldots,k\} \), let \( \chi_I \) be the indicator vector of \( I \). Suppose that \( I,J \) have \( f(\chi_I) = f(\chi_J) = 1 \), then from \( \chi_I \land \chi_J \implies \chi_{I \cap J} \) (coordinatewise) we see that we must have \( f(\chi_{I \cap J}) = 1 \) as well. Thus, there is a unique minimum subset \( I^* \) satisfying \( f(\chi_{I^*}) = 1 \). Since \( f \) is monotone, we have \( f(\chi_J) = 1 \iff J \supseteq I^* \).

Finally, we come to the affine linear case. We leave the proofs of the following two results to the reader.

**Theorem 3.9.** Suppose that a relation \( R \subseteq (\mathbb{Z}/p)^m \) is preserved by the ternary operation \( x - y + z \pmod{p} \). Then \( R \) is an affine linear subspace of \( (\mathbb{Z}/p)^m \) - that is, a vector subspace of \( (\mathbb{Z}/p)^m \) offset by a fixed vector - and \( R \in \langle \{1\}, x + y \equiv z \pmod{p} \rangle \).

**Theorem 3.10.** Suppose that a function \( f : (\mathbb{Z}/p)^k \to \mathbb{Z}/p \) preserves the relations \( \{1\} \) and \( x+y \equiv z \pmod{p} \). Then \( f \) is an affine linear function - that is, a linear function such that the sum of the coefficients is 1 - and \( f \in \langle x - y + z \pmod{p} \rangle \). If \( p \) is odd, we have \( \langle x - y + z \pmod{p} \rangle = \langle \frac{x+y}{2} \pmod{p} \rangle \).

4 Varieties, Birkhoff’s HSP theorem, and the hardness proof

From here on we switch over to the algebraic language. To a relational structure \( A = (D, \Gamma) \) we associate an algebraic structure \( A = (D, O) \) with \( \langle O \rangle = \text{Pol}(\Gamma) \). We let \( \text{CSP}(A) \) be another name for \( \text{CSP}(A) = \text{CSP} (\text{Inv}(A)) \).

**Remark 4.1.** Suppose \( \mathbb{A}, \mathbb{B} \) are two algebraic structures with associated relational structures \( A, B \).

It is tempting to think that a homomorphism \( \mathbb{A} \to \mathbb{B} \) will correspond to a homomorphism \( A \to B \), or vice versa. Unfortunately, this is total nonsense - if the (functional) signatures of \( \mathbb{A} \) and \( \mathbb{B} \) match, the (relational) signatures of \( A \) and \( B \) will likely have nothing to do with each other!

In a similar vein, the automorphism groups \( \text{Aut}(\mathbb{A}) \) and \( \text{Aut}(A) \) have almost nothing to do with each other. A trivial but illuminating example is the case where \( A \) has no functions at all, so that \( \text{Aut}(\mathbb{A}) \) is the full symmetric group - in this case, \( A \) has every possible relation in its signature, including named singleton unary relations for every element of the domain. Thus, if \( \mathbb{A} \) is trivial, then \( A \) is rigid, with \( \text{Aut}(A) = \{1\} \).

We will now use the algebraic language to relate the complexities of CSPs with different domains. This will finally clarify what we meant by one CSP “simulating” another CSP in the introduction (well, there is one more method of simulation that will be introduced in the next section).

**Theorem 4.1.** If \( \mathbb{A} \) is an algebraic structure, and \( \mathbb{B} \) is either

- a subalgebra of \( \mathbb{A} \),
- a power of \( \mathbb{A} \), or
- a quotient of \( \mathbb{A} \),

then there is a logspace reduction from \( \text{CSP}(\mathbb{B}) \) to \( \text{CSP}(\mathbb{A}) \).
Proof. If $B$ is a subalgebra of $A$, we can convert any instance of CSP($B$) into an instance of CSP($A$) by simply adding an extra unary constraint for each variable corresponding to the relation $B \subseteq A$. If $B = A^n$ for some $n$, then we can convert an instance of CSP($B$) to an instance of CSP($A$) by replacing each variable with an $n$-tuple of variables, and using the fact that every subalgebra of $(A^n)^m$ is a subalgebra of $A^{mn}$. 

If $B = A/\sim$ for some congruence $\sim \subseteq A^2$ on $A$, then every relation $R \subseteq B^m$ lifts to a relation $\tilde{R} \subseteq A^m$ by the rule $x \in \tilde{R} \iff x/\sim \in R$.

More generally, if we have several algebras $A_1, A_2, \ldots$ in the same (functional) signature, we can define CSP($\{A_1, A_2, \ldots\}$) to be the problem where each variable comes with a sort - that is, a specific algebra $A_i$ that it lives in - and each relation is multisorted, where a multisorted relation is “allowed” if it cuts out a subalgebra of the relevant product of the $A_i$s. This sort of multisorted relation was considered by Bulatov and Jeavons [32]. In this framework, there is a logspace equivalence between CSP($A_1 \times A_2$) and CSP($\{A_1, A_2\}$).

So we see that we are naturally led to study families of finite algebras (all sharing a signature) which are closed under taking finite products, subalgebras, and quotients. This leads us to the concept of a variety (or pseudovariety, if the family of finite algebras is not finitely generated). Lurking in the background here is a new Galois connection, this time between families of identities and families of algebras.

**Definition 4.2.** A term (in a given functional signature) is either a variable name or a $k$-ary function symbol applied to a $k$-tuple of previously constructed terms. An identity is a formal expression $s \approx t$, where $s$ and $t$ are terms. An algebra $A$ satisfies an identity $s \approx t$, written $A \models s \approx t$, if

$$\forall x_1, \ldots, x_n \in A \ s(x_1, \ldots, x_n) = t(x_1, \ldots, x_n)$$

(here we are assuming that the variables of $s$ and $t$ are drawn from $x_1, \ldots, x_n$).

The $\approx$ notation is confusing at first, since in the context of universal algebra it is viewed as a statement which is stronger than ordinary equality. The idea here is that approximate equality is never considered in universal algebra, so there should be no confusion in repurposing the symbol $\approx$ into an abbreviation for universal quantifiers. For instance, the intended meaning of the expression “$f(x, y) \approx f(y, x)$” is “$\forall x, y \ f(x, y) = f(y, x)$”. An alternate point of view is that $\approx$ refers to the congruence on the absolutely free algebra corresponding to the identities which are satisfied by the algebras we are interested in.

**Definition 4.3.** The variety $V(T)$ determined by a set of identities $T$ is the set of algebras that satisfy all of the identities in $T$. If $A_1, A_2, \ldots$ is a collection of algebras, then $V(A_1, A_2, \ldots)$ is the variety associated to the set of all identities that hold simultaneously in all of the algebras $A_i$.

Birkhoff [21] introduced a convenient notation for manipulating sets (strictly speaking these are classes, not sets) of algebras: if $S$ is a set of algebras, then $PS$ is the set of products of algebras from $S$ (possibly infinite - $P_{fin}$ is the notation if one restricts to finite products), $SS$ is the set of subalgebras of algebras from $S$, and $HS$ is the set of quotients (homomorphic images) of algebras from $S$.

**Theorem 4.4** (Birkhoff’s HSP Theorem [21]). For any collection $S$ of algebras, we have $V(S) = HSP(S)$, that is, an algebra $A$ satisfies every identity which is satisfied in every element of $S$ if and only if it is the homomorphic image of a subalgebra of a product of elements of $S$. Furthermore, if $S$ is a finite collection of finite algebras, then the set of finite algebras in $V(S)$ is equal to $HSP_{fin}(S)$.  

24
Proof. It is easy to check that if $A, B \models s \approx t$, then $A \times B \models s \approx t$, and similarly every subalgebra and quotient of $A$ also satisfies $s \approx t$. Thus $HSP(S) \subseteq \mathcal{V}(S)$.

For the other containment, suppose that $A \in \mathcal{V}(S)$, and suppose that $A$ is generated by a subset $I \subseteq A$. We let $P$ be the product of all the algebras of $S$, and define the “free algebra” $F(I)$ be the subalgebra of $P$ which is generated by the projection functions $\pi_i$ for $i \in I$, given by $\pi_i(x) = x_i$. We claim that there is a surjective homomorphism $h : F(I) \twoheadrightarrow A$ with $h(\pi_i) = i$.

Suppose not. Then there are two terms $s, t$ with $s(\pi_{i_1}, ..., \pi_{i_n}) = t(\pi_{i_1}, ..., \pi_{i_n})$ in $P$, but $s(i_1, ..., i_n) \neq t(i_1, ..., i_n)$ in $A$. But then $s \approx t$ is satisfied by $P$, and hence by every algebra in $S$, and is not satisfied in $A$, contradicting our assumption that $A \in \mathcal{V}(S)$.

For the last claim, note that if $A, S$, and every element of $S$ are finite, then so are $I, P, P^I$, and $P^P^I$.

Birkhoff’s HSP theorem gives one half of the Galois connection between identities and algebras. The other half is a result from model theory, which explains why elementary results in algebra can always be proved by writing down a long string of equalities.

**Theorem 4.5.** If $T$ is a family of identities, then the set of identities which hold in $\mathcal{V}(T)$ is equal to the closure of $T \cup \{x \approx x\}$ under:

- substituting a term for a variable in an identity,
- applying a $k$-ary function to both sides of a $k$-tuple of identities,
- deducing $s \approx t$ from $t \approx s$, and
- deducing $s \approx u$ from $s \approx t$ and $t \approx u$.

*Proof.* Define the free algebra $F_T(x_1, ...)$ on countably many variables by taking the set of all terms on these variables, and then taking the quotient of this term algebra by the congruence generated by the images under all possible substitutions of the identities in $T$. The result will be an algebra satisfying all of the identities of $T$, and one can check directly from the definition of a congruence that the identities that hold in this free algebra are exactly the ones described in the theorem statement.

Using Birkhoff’s theorem, we can give a criterion for NP-completeness.

**Theorem 4.6.** If $CSP(A)$ is not NP-complete, then there is a finite set of identities $s_i \approx t_i$ which are satisfied by $A$, which can’t be satisfied by assigning each function symbol to a projection of the same arity.

*Proof.* If $\mathcal{V}(A)$ contains an algebra $B$ of size at least 2 where each function symbol acts as a projection, then $CSP(B)$ is NP-complete and has a logspace reduction to $CSP(A)$. Such an algebra $B$ will exist if there is a way to assign the function symbols to projections that satisfies every identity satisfied by $A$. To see that we only have to consider finite sets of identities, we apply a compactness argument (each function symbol has only a finite number of projections it can be assigned to, so we can apply König’s Lemma).
Example 4.1. Consider the algebra \( A = \{0, 1\}, \text{min} \), and use the binary function symbol \( s \) to abbreviate \( \text{min} \). Then \( V(A) = SP(A) = V(T_{\text{semi}}) \), where \( T_{\text{semi}} \) is the following set of identities:

\[
s(x, x) \approx x, \quad s(x, y) \approx s(y, x), \quad s(x, s(y, z)) \approx s(s(x, y), z).
\]

The second identity above, \( s(x, y) \approx s(y, x) \), can’t be satisfied by assigning \( s \) to either of the projections \( \pi_1, \pi_2 \).

An algebra in \( V(T_{\text{semi}}) \) is called a semilattice, and can be visualized as a poset where every nonempty finite subset has a greatest lower bound (if we visualize it this way, we often call it a meet semilattice).

Any finite meet semilattice which has a greatest element can be extended to a lattice, since every finite subset will also have a least upper bound (just take the greatest lower bound of the collection of all upper bounds, which is nonempty by the assumption that there is a greatest element).

Alternatively, a semilattice can be thought of as a poset where every nonempty finite subset has a least upper bound, if we are thinking in terms of an operation like \( \text{max} \) - if we are visualizing it in this way, we call it a join semilattice. (I generally prefer to visualize semilattices as join semilattices.)

Example 4.2. Consider the algebra \( A = \{0, 1\}, \text{maj} \), and use the ternary function symbol \( m \) to abbreviate \( \text{maj} \). Then \( V(A) = SP(A) = V(T_{\text{med}}) \), where \( T_{\text{med}} \) is the following set of identities:

\[
m(x, y, z) \approx m(y, z, x) \approx m(x, z, y),
m(x, x, y) \approx x,
m(m(x, y, z), u, v) \approx m(x, m(y, u, v), m(z, u, v)).
\]

The identity \( m(x, y, z) \approx m(y, z, x) \) can’t be satisfied by assigning \( m \) to one of the projections \( \pi_1, \pi_2, \pi_3 \).

An algebra in \( V(T_{\text{med}}) \) is called a median algebra. A finite median algebra corresponds to a median graph, that is, a graph with the property that for every three vertices \( x, y, z \) there exists a unique vertex which lies on a shortest path connecting every pair of \( x, y, z \). Examples of median graphs are paths, trees, planar grids, “squaregraphs”, hypercubes, Hasse diagrams of distributive lattices, and the graph \( M_n \) of maximal intersecting families from the last section. For more about the theory of median algebras, see [102].

One of the most famous examples of median algebras is the case of distributive lattices. It isn’t hard to show that in any distributive lattice, the following identity holds:

\[
(x \land y) \lor (y \land z) \lor (z \land x) \approx (x \lor y) \land (y \lor z) \land (z \lor x),
\]

and in fact this identity is equivalent to the lattice being distributive. The common value of both sides is called the median operation \( m(x, y, z) \) on the lattice - the reader can easily check that it satisfies the identities \( T_{\text{med}} \). In fact, if a median algebra has two elements \( 0, 1 \) with \( m(0, x, 1) = x \) for all \( x \), then it forms a distributive lattice under the operations \( x \land y = m(0, x, y) \) and \( x \lor y = m(x, y, 1) \), and the median operation \( m \) can be recovered from \( \land, \lor \) via the above formula [24].

Example 4.3. The operation \( m(x, y, z) = x - y + z \ (\text{mod} \ p) \) satisfies the identity \( m(x, y, y) \approx x \approx m(y, y, x) \), and this identity can’t be satisfied by assigning \( m \) to one of the projections \( \pi_1, \pi_2, \pi_3 \).

Similarly, if \( p \) is odd, the operation \( m(x, y) = \frac{x+y}{2} \ (\text{mod} \ p) \) satisfies the identity \( m(x, y) \approx m(y, x) \), which can’t be satisfied by projections.
As with clones and relational clones, there are several natural finiteness questions that come up with varieties.

**Definition 4.7.** A variety $\mathcal{V}$ is **finitely generated** if there is a finite list of finite algebras $A_1, \ldots, A_n$ such that $\mathcal{V} = \mathcal{V}(A_1, \ldots, A_n)$. A variety $\mathcal{V}$ is **locally finite** if the free algebra on $n$ generators $F_\mathcal{V}(x_1, \ldots, x_n)$ is finite for every $n$. A variety $\mathcal{V}$ is **finitely based** if there is a finite set of equations $T$ such that $\mathcal{V} = \mathcal{V}(T)$.

A variety $\mathcal{V}$ is locally finite iff for all $A \in \mathcal{V}$ and for all finite subsets $\{a_1, \ldots, a_n\} \subseteq A$, the subalgebra of $A$ generated by $a_1, \ldots, a_n$ is finite. Every finitely generated variety is locally finite (by the proof of the HSP Theorem). In general, determining whether a given finitely generated variety is finitely based, or vice versa, is a very difficult problem. For instance, the famous Burnside problem is the problem of determining whether the variety of groups satisfying the identity $x^n \approx x$ is locally finite.

**Remark 4.2.** Sometimes we want to consider infinite families of finite algebras with a finite functional signature, closed under finite products, subalgebras, and homomorphisms. Such a family of algebras is called a **pseudovariety**. There are two different ways to describe pseudovarieties in terms of identities.

Eilenberg and Schützenbérger [50] show that a pseudovariety is determined by an infinite sequence of identities, such that every algebra in the pseudovariety satisfies all but finitely many of the identities in the sequence.

Reiterman [101] shows that a pseudovariety is determined by identities between “implicit operations”: operations which aren’t defined from terms directly, but which can still be defined on any particular finite algebra in a way that is compatible with homomorphisms. Examples of implicit operations in the language of a unary function $f$ are

$$f^\infty = \lim_{n \to \infty} f^{\circ n!}, \quad f^{\infty-1} = \lim_{n \to \infty} f^{\circ n!-1},$$

where the limits are taken pointwise (note that the functions $f^{\circ n!}$ stabilize once $n$ exceeds the size of the domain). For any function $f$ on a finite domain, $f^\infty$ will always satisfy the identity $f^\infty(f^\infty(x)) \approx f^\infty(x)$, while the pseudovariety of invertible functions on finite sets is cut out by the identities

$$f(f^{\infty-1}(x)) \approx f^{\infty-1}(f(x)) \approx x.$$

For those who like category theory, a $k$-ary implicit operation of a pseudovariety $\mathcal{V}$ with underlying set functor $S : \mathcal{V} \to \text{Set}$ is a natural transformation from $S^k$ to $S$. If a free algebra on $k$ elements exists in $\mathcal{V}$, then a standard argument shows that every $k$-ary implicit operation of $\mathcal{V}$ is actually explicit, that is, a term of $\mathcal{V}$. In general, every finite subset of $\mathcal{V}$ will generate a locally finite subvariety of $\mathcal{V}$, which shows that the restriction of any implicit operation to this subset agrees with some term of $\mathcal{V}$. Reiterman [101] puts a metric structure on the set of implicit operations of a pseudovariety such that the collection of implicit operations becomes the completion of the collection of explicit operations.

### 5 Cores and Idempotent Reducts

In this section we briefly return to the relational point of view, and the concept of homomorphic equivalence, to provide one last algebraic ingredient: the restriction to **idempotent** algebraic operations.
Definition 5.1. Two relational structures $A, B$ with the same signature are *homomorphically equivalent* if there exist homomorphisms $A \rightarrow B, B \rightarrow A$.

The prototypical example of homomorphic equivalence is a (non-surjective) endomorphism from a relational structure to itself, providing a homomorphic equivalence between the original relational structure and the restriction of the relational structure to a proper subset of its domain. On the algebraic side, this manifests as a unary operation which is not invertible. The algebraic implications of such unary operations in the polynomial clone of an algebra are at the heart of the subject called “tame congruence theory”, which was introduced to give the first structure theory for finite algebras \[65\].

Example 5.1. Consider the relational structure $A$ corresponding to the binary implication algebra $A = (\{0, 1\}, \rightarrow)$. This relational structure has as basic relations $R_n = \{0, 1\}^n \setminus \{(0, ..., 0)\} = x_1 \lor \cdots \lor x_n$. The unary algebraic operation $\rightarrow(x, x)$ of $A$ takes every element to 1, and defines an endomorphism of relational structures $A \rightarrow A$ whose image is $\{1\}$. Together with the inclusion relation, we get a homomorphic equivalence between $A$ and the one-element relational structure with domain $\{1\}$ and relations $R_n |_{\{1\}} = \{1\}^n$, whose CSP is clearly trivial.

As the example shows, non-surjective endomorphisms provide trivial ways to simplify CSPs.

Definition 5.2. A relational structure on a finite domain $A$ is called a *core* if every endomorphism of $A$ is also an automorphism of $A$. If $A$ is not a core, then $B$ is called a *core of $A$* if $B$ is a core and $B$ is homomorphically equivalent to $A$.

Remark 5.1. In the infinite case, the definition of a core must be modified: an infinite relational structure is called a core if every endomorphism is an *embedding*, i.e. an injective map that is an isomorphism onto the restriction of the target relational structure to its image. An example of an infinite core is $(\mathbb{Q}, \lt)$. See section 3.6 of \[26\] for more information about cores of infinite structures.

Proposition 5.3. Every relational structure on a finite domain $A$ has a core. Any two cores of $A$ are isomorphic.

Proof. The first statement follows directly from induction on the size of $A$: if $A$ is not a core, then it is homomorphically equivalent to its restriction to some proper subset of itself. For the second statement, note that if $B, B'$ are two cores of $A$ then they are homomorphically equivalent, and composing the maps $B \rightarrow B', B' \rightarrow B$ gives us endomorphisms of $B, B'$ which must both be invertible by the definition of a core. \qed

Note that although restricting our attention to cores seems like a trivial step, we are sweeping the following problem under the rug.

Problem 5.1. Given a finite relational structure $A$ as input, determine whether or not $A$ is a core.

Obviously there is a brute-force approach to checking if $A$ is a core: simply write down every possible endomorphism, and go through them one by one. Since we only have to do this brute force once for a given CSP template, this is not as bad as it sounds, but it is still far from ideal. Unfortunately, as it turns out, a brute force approach is pretty much the best one can do.

Theorem 5.4 (Hell, Nešetřil \[62\]). Determining whether a given undirected graph is a core is \textit{NP}-complete, even if the graph is assumed to be 3-colorable (with a given 3-coloring).
The next main idea comes from “self-reducibility”: often, when solving a CSP, one makes a guess (or deduces) that a certain variable should have a certain value. We would like to be able to express a CSP together with some constraints stating that certain variables have certain values using the language of the original CSP. If this is possible, then an algorithm for deciding whether the CSP has a solution can be directly converted into an algorithm for finding a solution to the CSP.

**Definition 5.5.** A relational structure \( A \) is a rigid core if it has no endomorphisms other than the identity. (In general, a structure is called rigid if it has no automorphisms.)

**Theorem 5.6.** A relational structure \( A \) on a finite domain \( D \) is a rigid core if and only if it has the following property: for every element \( a \in D \), the unary relation \( \{a\} \) is contained in the relational clone generated by the relations of \( A \).

**Proof.** This follows directly from the Inv-Pol Galois connection: \( \{a\} \in \langle A \rangle \) iff \( \{a\} \) is closed under \( \text{Pol}(A) \), and since \( \{a\} \) is generated by a single element, we only need to check that it is closed under \( \text{Pol}_1(A) \), which is exactly the set of endomorphisms of \( A \). We can also give a direct proof, by unraveling the proof of the Inv-Pol connection in this special case, as follows. Define a CSP with a variable \( f_a \) for each \( a \in D \). For every relation \( R \subseteq D^m \) of \( A \) and every tuple \((a_1, \ldots, a_m) \in R\), we impose the constraint \( R(f_{a_1}, \ldots, f_{a_m}) \) on our CSP. Now the solution-set to our CSP exactly corresponds to the set of endomorphisms of \( A \), and if \( A \) is a rigid core then existentially projecting onto the variable \( f_a \) produces the unary relation \( \{a\} \).

So it is very desirable to restrict our attention to rigid cores. Most of the example CSPs from the introduction were rigid cores, with the notable exceptions of \( k \)-coloring and NAE-SAT. The \( k \)-coloring problem is an excellent toy example: the reader may be already be aware of the fact that \( \text{CSP}([1, \ldots, k], \neq) \) (the \( k \)-coloring problem) is logspace equivalent to \( \text{CSP}([1, \ldots, k], \neq, \{0\}, \ldots, \{k\}) \) - the rigid core obtained by adjoining the unary singleton relations to \( k \)-coloring. It is worth examining the proof of that equivalence and understanding how the next result generalizes it.

**Theorem 5.7.** Suppose that \( A = (D, \Gamma) \) is a core on a finite domain \( D \), and let \( A^{\text{rig}} \) be the rigid core obtained by adjoining all singleton unary relations to \( A \). Then \( \text{CSP}(A) \) is equivalent to \( \text{CSP}(A^{\text{rig}}) \) under logspace reductions.

**Proof.** We need to find a way to convert an instance of \( \text{CSP}(A^{\text{rig}}) \) to an instance of \( \text{CSP}(A) \) without changing whether it has a solution. As in the previous result, introduce a set of variables \( f_a \) for each element \( a \in D \), and define a primitive positive formula \( \Phi \) by

\[
\Phi(f) := \bigwedge_{R \in \Gamma} \bigwedge_{(a_1, \ldots, a_m) \in R} R(f_{a_1}, \ldots, f_{a_m}).
\]

Suppose that our instance of \( \text{CSP}(A^{\text{rig}}) \) has the form

\[
\Psi(x) = \exists x_{n+1}, \ldots, x_{n+m} \Psi_0(x) \land \bigwedge_{(i, a) \in E} x_i \in \{a\},
\]

where \( \Psi_0 \) is a primitive positive formula using the relations of \( \Gamma \), and \( E \) is a set describing the additional unary singleton constraints of \( \Psi \). Let \( \Psi' \) be the following formula:

\[
\Psi'(x) := \exists f \exists x_{n+1}, \ldots, x_{n+m} \Phi(f) \land \Psi_0(x) \land \bigwedge_{(i, a) \in E} x_i = f_a.
\]
We claim that the instance $\Psi'$ of $\text{CSP}(A)$ has a solution iff the instance $\Psi$ of $\text{CSP}(A^\text{rig})$ has a solution. Suppose that $f, x$ solves $\Psi'$, then by the construction of $\Phi(f)$ $f$ describes an endomorphism $f : A \to A$, and since $A$ is a core this endomorphism must have an inverse $f^{-1}$. Then $f^{-1}(x)$ satisfies $\Psi_0$ (since $f^{-1}$ is an endomorphism of $A$), and for $(i, a) \in E$ we have $f^{-1}(x_i) = f^{-1}(f_a) = a$, so $f^{-1}(x)$ is a solution to $\Psi(x)$. \hfill \Box

Now we look at what the restriction to rigid cores means on the algebraic side.

**Definition 5.8.** A function $f : D^k \to D$ is idempotent if it satisfies the identity $f(x, x, ..., x) \approx x$. An algebraic structure $A = (D, O)$ is idempotent if every $f \in O$ is idempotent. Equivalently, $A$ is idempotent if every singleton subset of $D$ is a subalgebra of $A$.

**Definition 5.9.** If $A = (D, O)$ is an algebraic structure, then the idempotent reduct $A^{id}$ of $A$ has the same domain, and has as its operations the set of all idempotent functions $f \in O$ (or, alternatively, some smaller generating set of idempotent functions).

**Proposition 5.10.** If $A$ is a core corresponding to the algebraic structure $A$, then the rigid core $A^{rig}$ corresponds to the idempotent reduct $A^{id}$. In particular, every CSP is equivalent up to logspace reductions to $\text{CSP}(A)$ for some idempotent algebra $A$.

**Example 5.2.** If $A = (\mathbb{Z}/p, +, 0, 1)$, then $A^{id}$ has as its operations the set of all affine linear functions on $\mathbb{Z}/p$, and one can take $\{x - y + z \pmod{p}\}$ as a generating set of basic operations (or, if $p$ is odd, one can alternatively take $\{\frac{x+y}{2} \pmod{p}\}$ as a generating set of basic operations).

**Example 5.3.** There is an example of a core structure $A$ which has polymorphisms satisfying a nontrivial system of identities, but such that its rigidification $A^{rig}$ has no such polymorphisms and is therefore NP-complete. This example is due to Ross Willard and can be found in [19].

The underlying set of $A$ is the set of expressions $a_i$ with $a \in \{1, 2, 3\}$ and $i \in \{0, 1\}$. The relations of $A$ are given by

\[ R(a_i, b_j) := (i = j) \land (a \neq b), \]
\[ S(a_i, b_j) := i \neq j. \]

It is easy to check that this structure is a core.

Polymorphisms of $A$ include the unary automorphism $\alpha(a_i) = a_{1-i}$ and the ternary function $s$ given by

\[ s(a_i, b_j, c_k) = \begin{cases} c_k \quad i = j, \\ a_i \quad i \neq j. \end{cases} \]

These polymorphisms satisfy the identity

\[ s(x, x, y) \approx s(y, \alpha(y), x) \approx y, \]

which can’t be satisfied by projections.

Since the unary relation $\{a_i \mid i = 0\}$ is definable in $A^{rig}$, we see that polymorphisms of $A^{rig}$ restrict to idempotent polymorphisms of the triangle $K_3$. We will show that $K_3$ has no nontrivial idempotent polymorphisms: in fact, we’ll show that every polymorphism of $K_3$ is the composition of a projection with an automorphism of $\{1, 2, 3\}$.

To see that all polymorphisms of $K_3$ are essentially unary, suppose that $f : K_3^n \to K_3$ depends nontrivially on its first coordinate, that is, that there are $x, y \in K_3^n$ with $x_i = y_i$ for all $i > 1$
with \( f(x) \neq f(y) \). By composing with automorphisms of \( \{1, 2, 3\} \), we may assume without loss of generality that

\[
f(1, 1, \ldots, 1) = 1, f(2, 1, \ldots, 1) = 2.
\]

Since \( f \) preserves the \( \neq \) relation, we must then have

\[
f(3, 2, \ldots, 2) = f(3, 3, \ldots, 3) = 3.
\]

These imply that

\[
f(\{1, 2\}^n) \subseteq \{1, 2\}, f(\{1, 2\} \times \{1, 3\}^{n-1}) \subseteq \{1, 2\}.
\]

For any \( z_2, \ldots, z_n \), we can find \( x_2, \ldots, x_n \in \{1, 2\} \) and \( y_2, \ldots, y_n \in \{1, 3\} \) with \( x_i, y_i, z_i \) all distinct. Thus we must have

\[
f(3, z_2, \ldots, z_n) = 3
\]

for all \( z_2, \ldots, z_n \), and in particular \( f(3, 1, \ldots, 1) = 3 \). Now we can repeat the argument with 1 or 2 in place of 3 to see that \( f(x_1, \ldots, x_n) = x_1 \) for all \( x_1, \ldots, x_n \), that is, \( f = \pi_1 \).

Alternatively, we could have shown that \( \text{Pol}(K^{(2)}) \) is trivial by instead showing that every relation on \( \{1, 2, 3\} \) is primitively positively definable from the singleton relations together with \( \neq \). We leave this as an exercise for the reader (hint: once you have all ternary relations of the form \( (x = a) \land (y = b) \implies (z = c) \), it’s easy to construct the rest).

### 5.1 Reflections and Height 1 Identities

Barto, Oprešal, and Pinsker [18] find it unsatisfactory to have so many unrelated methods of proving reductions between CSPs (that is, firstly the algebraic HSP operations together with another operation \( E \) for expansions, secondly homomorphic equivalence, and thirdly adding unary singleton relations to cores), and looked for a single framework that could encompass all known techniques for proving reductions. They show that every single method of proving a reduction between CSP(\( A \)) and CSP(\( B \)) introduced so far can be described by combining just two basic cases:

- if \( B \) is a “pp-power” (defined below) of \( A \), then CSP(\( B \)) has a logspace reduction to CSP(\( A \)),
- if \( B \) is homomorphically equivalent to \( A \) then CSP(\( B \)) = CSP(\( A \)).

Furthermore, they show that one can always assume that the pp-power step is taken before the homomorphic equivalence step.

**Definition 5.11.** A **pp-power** of a relational structure \( A \) is a relational structure \( B \) with domain \( A^n \) for some \( n \), such that every relation of \( B \) can be defined by a primitive positive formula using the relations of \( A \) (note that in this case the signatures of \( A \) and \( B \) will generally be different).

For instance, here is how they go about adding a singleton unary relation \( \{a\} \) to a core \( A \). Let \( B \) be the relational structure which has the new unary relation \( \{a\} \). We will define a relational structure \( C \) which will be a pp-power of \( A \) having domain \( A^2 \), and show that \( C \) is homomorphically equivalent to \( B \). Let \( O \) be the orbit of \( a \) under \( \text{Aut}(A) \) - note that \( O \) is in the relational clone defined by \( A \) - and for every \( m \)-ary relation \( R \) of \( A^2 \), make a corresponding relation \( \hat{R} \) of \( C \) by

\[
((x_1, y_1), \ldots, (x_m, y_m)) \in \hat{R} \iff (x_1, \ldots, x_m) \in R \land y_1 = \cdots = y_m \in O.
\]
For the relation \{a\} of \(B\), we make a corresponding relation \(S\) of \(C\) given by
\[(x, y) \in S \iff x = y \in O.\]

To show that \(B\) and \(C\) are homomorphically equivalent, we just need to exhibit a pair of homomorphisms between them. The homomorphism \(B \to C\) is given by \(x \mapsto (x, a)\). To define the homomorphism from \(C\) to \(B\), we need to choose an automorphism \(g_y\) of \(A\) with \(g_y(y) = a\) for every \(y \in O\). Then the homomorphism \(C \to B\) is given by \((x, y) \mapsto g_y(x)\) if \(y \in O\) (and \((x, y)\) maps to an arbitrary element if \(y \not\in O\)).

Barto, Opršal, and Pinsker [18] then characterize what happens on the algebraic side of the picture when one relates two relational structures by a pp-power or a homomorphism equivalence. The new thing here is really the homomorphic equivalence: if \(g : A \to B\) and \(h : B \to A\), then there is a relationship between \(\text{Pol}(A)\) and \(\text{Pol}(B)\) which they call a reflection, which takes a function \(f \in \text{Pol}_k(A)\) to the operation
\[\xi(f) : (x_1, \ldots, x_k) \mapsto g(f(h(x_1), h(x_2), \ldots, h(x_k)))\]
in \(\text{Pol}_k(B)\). Note that \(\xi\) does not respect composition: \(\xi(f_0 \circ (f_1, \ldots, f_k))\) is not in general equal to \(\xi(f_0) \circ (\xi(f_1), \ldots, \xi(f_k))\). However, \(\xi\) does preserve height 1 identities.

**Definition 5.12.** An identity is called a height 1 identity, or a minor identity, if it has the form \(f(x_1, \ldots, x_k) \approx g(y_1, \ldots, y_l)\), where the \(x_i\)s and \(y_j\)s are (not necessarily distinct) variables. A map \(\text{Pol}(A) \to \text{Pol}(B)\) (taking functions to functions) which respects height 1 identities is called a height 1 clone homomorphism or a minor homomorphism.

**Definition 5.13.** If \(\mathbb{A} = (A, \mathcal{O})\) is an algebraic structure and \(B\) is a set, and maps \(g : A \to B\), \(h : B \to A\) are given, then the reflection of \(\mathbb{A}\) induced by \(g, h\) is defined to be the algebraic structure \(\mathbb{B}\) with domain \(B\) and the same signature as \(\mathbb{A}\), with the operation \(g \circ f \circ (h, \ldots, h)\) on \(B\) corresponding to the operation \(f \in \text{Pol}_k(A)\) to the relation

**Proposition 5.14.** \(B\) is homomorphically equivalent to a pp-power of \(A\) iff \(\text{Pol}(B)\) contains a reflection of \(\text{Pol}(A)^n\) for some \(n\) (by \(\text{Pol}(A)^n\) we mean the clone of operations of \(\text{Pol}(A)\) acting on a power of the domain).

**Proof.** We prove the non-obvious direction. Let \(A, B\) be the underlying sets of \(A, B\), and suppose that \(g : A^n \to B\) and \(h : B \to A^n\) induce a reflection \(\xi : \text{Pol}(A)^n \to \text{Pol}(B)\). We will construct a pp-power \(C\) of \(A\) with underlying set \(A^n\) which is homomorphically equivalent to \(B\). For every relation \(R\) of \(B\), let \(\tilde{R}\) be the relation
\[\tilde{R} := \{f(h(r_1), \ldots, h(r_k)) \mid f \in \text{Pol}_k(A), r_1, \ldots, r_k \in R\}.\]
By definition, \(\tilde{R}\) is the closure of \(h(R)\) under \(\text{Pol}(A)\), so \(\tilde{R}\) is defined by a primitive positive formula over \(A\). We use \(\tilde{R}\) as the relation corresponding to \(R\) in \(C\). Finally, we just need to check that \(g : C \to B\) and \(h : B \to C\) are homomorphisms. That \(h\) is a homomorphism follows from \(h(R) \subseteq \tilde{R}\). To check that \(g\) is a homomorphism, note that if \(x = f(h(r_1), \ldots, h(r_k)) \in \tilde{R}\) with \(r_1, \ldots, r_k \in R\), then \(g(x) = \xi(f)(r_1, \ldots, r_k)\) is an element of \(R\) since \(\xi(f) \in \text{Pol}(B)\) by assumption.

**Theorem 5.15** (ERP Theorem [18]). \(\text{Pol}(B)\) contains a reflection of \(\text{Pol}(A)^n\) for some \(n\) iff there is a height 1 clone homomorphism \(\text{Pol}(A) \to \text{Pol}(B)\).
Corollary 5.16. Let $\mathcal{A} = (A, \text{Pol}(A))$ be the algebraic structure corresponding to $A$, and suppose $\xi : \text{Pol}(A) \to \text{Pol}(B)$ is a height 1 clone homomorphism. Let $F$ be the subalgebra of $\mathcal{A}^B$ generated by the functions $\pi_b : A^B \to A$ given by $\pi_b : f \mapsto f(b)$. Note that $F$ is secretly the free algebra over $\mathcal{A}$ on $|B|$ generators.

Define maps $g : \mathcal{A}^B \to B$ and $h : B \to \mathcal{A}^B$ by $h(b) = \pi_b$ and $g(f(\pi_{b_1}, \ldots, \pi_{b_k})) = \xi(f)(b_1, \ldots, b_k)$ for $f \in \text{Pol}_k(A)$, and define $g(x)$ arbitrarily for $x \notin F$. To see that $g$ is well-defined, note that if $f_0(\pi_{b_1}, \ldots, \pi_{b_k}) = f_1(\pi_{c_1}, \ldots, \pi_{c_k})$, then $f_0, f_1$ are related by a height 1 identity in $\mathcal{A}$ which implies that $\xi(f_0)(b_1, \ldots, b_k) = \xi(f_1)(c_1, \ldots, c_k)$. Finally, we see that $g, h$ induce $\xi$ as a reflection from $\mathcal{A}^B : \xi(f)(b_1, \ldots, b_k) = g(f(\pi_{b_1}, \ldots, \pi_{b_k})) = g(f(h(b_1), \ldots, h(b_k)))$. \hfill \Box

As a consequence, we see that the complexity of a CSP only depends on the set of height 1 identities satisfied by its polymorphisms, and that identities involving composition of functions are in a sense superfluous. We also have the following result.

**Corollary 5.16.** Let $A$ be a relational structure with core $B$, and let $B^{\text{rig}}$ be $B$ together with any finite collection of singleton unary relations. Then a system of height 1 identities can be satisfied in $\text{Pol}(A)$ iff it can be satisfied in $\text{Pol}(B^{\text{rig}})$.

## 6 Taylor Algebras

Once we restrict to idempotent algebras, we can start playing games with identities involving nesting functions to simplify our criterion for NP-completeness.

**Definition 6.1.** An algebra $A$ is a *Taylor algebra* if it has an idempotent term $t$ that satisfies a system of identities of the form

$$
t(t \left( \begin{bmatrix} x & ? & \cdots & ? \\ ? & x & \cdots & ? \\ \vdots & \vdots & \ddots & \vdots \\ ? & ? & \cdots & x \end{bmatrix} \right)) \approx t \left( \begin{bmatrix} y & ? & \cdots & ? \\ ? & y & \cdots & ? \\ \vdots & \vdots & \ddots & \vdots \\ ? & ? & \cdots & y \end{bmatrix} \right),$$

where the $?$s are filled in somehow with $x$s and $y$s. Such an operation $t$ is called a *Taylor term*, and a variety with a Taylor term is called a *Taylor variety*.

Note that from the defining identities of any Taylor term $t$, $t$ can’t be any projection (unless the algebra in question has only one element).

**Theorem 6.2** (Taylor [107]). *If an idempotent algebra $A$ satisfies any set of identities that can’t be satisfied by projections, then it has a Taylor term. Equivalently, an idempotent variety is Taylor iff it does not contain a two element algebra having no nontrivial operations.*

Before we prove Taylor’s theorem, we will work through an example.

**Example 6.1.** Let $f$ be an idempotent ternary term satisfying the identity

$$f(f(y, x, z), x, f(z, y, y)) \approx f(x, y, z).$$

Then

$$t(x_1, \ldots, x_9) := f(f(x_1, x_2, x_3), f(x_4, x_5, x_6), f(x_7, x_8, x_9))$$
is a Taylor term, since it satisfies the identities
\[ t(y, x, z, x, x, y, y, y, y, z, z, z) \approx t(x, y, z, x, y, z, x, y, z), \]
and by specializing these identities (substituting \( x = y, y = z, \) or \( z = x \)) we can get a system of Taylor identities for \( t \).

**Definition 6.3.** If \( f : D^k \to D \) and \( g : D^l \to D \), we define the star composition \( f \ast g : D^{kl} \to D \) to be \( f \circ (g, g, \ldots, g) \).

**Proposition 6.4.** If \( f, g \) are idempotent, then \( f, g \in \langle f \ast g \rangle \).

**Proof.**
\[
f(x_1, ..., x_k) \approx f(g(x_1, ..., x_1), ..., g(x_k, ..., x_k)) \quad \text{and} \quad g(x_1, ..., x_l) \approx f(g(x_1, ..., x_1), ..., g(x_1, ..., x_l)).
\]

**Definition 6.5.** An identity is called a height 1 identity, or a minor identity, if it has the form \( f(x_1, ..., x_k) \approx g(y_1, ..., y_l) \), where the \( x_i \)s and \( y_j \)s are (not necessarily distinct) variables.

**Proposition 6.6.** If an idempotent term satisfies a system of height 1 identities which can’t be satisfied by projections, then it is a Taylor term.

**Proof of Taylor’s Theorem.** By a compactness argument, there is a finite set \( \mathcal{T} \) of identities satisfied by a finite set of operations \( f_1, ..., f_n \) of \( \mathbb{A} \) which can’t be satisfied by projections. Let \( s = f_1 \ast \cdots \ast f_n \). Then each \( f_i \in \langle s \rangle \), so we can convert \( \mathcal{T} \) into a collection \( \mathcal{T}' \) of identities in \( s \) which can’t be satisfied by projections either.

The identities of \( \mathcal{T}' \) might involve some amount of nesting of \( s \) within itself, that is, they may not be height 1 identities. Let \( m \) be the greatest nesting depth occurring in \( \mathcal{T}' \), and let \( t = s \ast \cdots \ast s \), with \( m \) copies of \( s \). Let \( \mathcal{T}'' \) be the set of height 1 identities involving \( t \) only which are satisfied by \( \mathbb{A} \). We claim that \( \mathcal{T}'' \) can’t be satisfied by projections.

To see this, note first that for every \( k \leq m \), if we index the variables of \( t \) by \( m \)-tuples \( (i_1, ..., i_m) \) of indices for coordinates of \( s \), and if we let \( x^k \) be the tuple of variables given by
\[
x_{(i_1, ..., i_m)}^k = y_{i_k}
\]
for all \( (i_1, ..., i_m) \), then we have
\[
t(x^1) \approx \cdots \approx t(x^m) \quad (\approx s(y)).
\]
If this system of height 1 identities in \( t \) is satisfied by a projection \( \pi_{(i_1, ..., i_m)} \), then we see that we must have \( i_1 = \cdots = i_m = i \), say, for some index \( i \) of the variables of \( s \). But then there is some identity of \( \mathcal{T}' \) which is incompatible with \( s = \pi_i \), and this identity of \( \mathcal{T}' \) can be modified by replacing variables \( z \) by expressions \( s(z, ..., z) \) repeatedly until it becomes a height 1 identity involving only \( t \), which will then be incompatible with \( t = \pi_{(i, ..., i)} \). \( \square \)

**Corollary 6.7.** If \( \mathbb{A} \) is an idempotent algebra and CSP(\( \mathbb{A} \)) is not NP-complete, then \( \mathbb{A} \) has a Taylor term.
Remark 6.1. A recent result of Olšák simplifies the identities we need to consider even further. Olšák [92] proves that in any Taylor algebra, whether finite or infinite, there is always a 6-ary weak 3-cube term \( t \), that is, a term satisfying the identity

\[
t(x, y, y, y, x, x) \approx t(y, x, y, x, x, x) \approx t(y, y, x, x, x, y).
\]

The weak 3-cube term may be understood as saying that the ternary relation on the free algebra \( F_V(x, y) \) which is generated by the ternary Not-All-Equal relation on \( \{x, y\} \) has a diagonal element.

Olšák’s proof that such a term exists first uses the theory of absorbing subalgebras to produce a 12-ary term which he calls a double loop term, and then simplifies it down to a weak 3-cube term by using an intricate collection of identities that are satisfied by binary idempotent operations.

Remark 6.2. There is a curious connection between systems of two-variable height 1 identities on ternary functions and the problem 1-IN-3 SAT. Suppose that you are given such a system of identities \( T \) on ternary functions \( f_1, ..., f_n \), and that you want to determine whether these identities rule out projections.

Define a set of binary functions \( f^j_i(x, y), j \leq 3 \), by

\[
\begin{align*}
p_i(x, y, z) &= t(y, z, z, x, x, x), \\
q_i(x, y, z) &= t(x, z, y, z, y, z), \\
r_i(x, y, z) &= t(y, x, x, z, y, y), \\
s_i(x, y, z) &= t(x, x, y, z, y, x).
\end{align*}
\]

Then the definitions together with the weak 3-cube identities imply the following system of identities on \( p, q, r, s \):

\[
\begin{align*}
p(x, y, x) &\approx q(y, x, x), \\
p(x, x, y) &\approx r(y, x, x), \\
q(x, y, x) &\approx s(x, y, x), \\
r(x, x, y) &\approx s(x, x, y), \\
s(x, y, y) &\approx q(x, x, y) \approx r(x, y, x).
\end{align*}
\]

It may not be apparent, at a glance, whether or not this system of identities can be satisfied by projections. If we draw the associated 1-IN-3 SAT instance, we find that it has 7 vertices (corresponding to binary terms), 4 occurrences of the 1-IN-3 SAT constraint (for the four ternary terms \( p, q, r, s \)), and one occurrence of the \( \neq \) constraint (coming from the fact that the last identity above relates \( s(x, y, y) \) to \( q(x, x, y) \)):
It is now easy (well, as easy as solving a small instance of 1-IN-3 SAT) to verify that the associated 1-IN-3 SAT instance has no solution, so this system of identities can’t be satisfied by projections.

**Remark 6.3.** Taylor’s original reason for studying Taylor algebras was to try to deeply understand the reason that \( \pi_1 \) of a topological group is always abelian. Taylor \[107\] considers, for any variety \( \mathcal{V} \), the category of topological \( \mathcal{V} \)-objects, that is, topological algebraic structures satisfying the identities of \( \mathcal{V} \). Taylor showed that \( \pi_1 \)s of topological \( \mathcal{V} \)-objects will share a nontrivial property iff \( \mathcal{V} \) has a Taylor term, and that this occurs iff \( \pi_1 \) is always abelian. The fact that a Taylor term must be taken to be idempotent is related to the fact that the fundamental group is really a groupoid (in the sense of category theory), rather than a group, so only the idempotent operations of \( \mathcal{V} \) can constrain its structure (I’m slightly fuzzy on the details).

Aside from the topological details, this can be viewed as an analogue of the Eckmann-Hilton principle \[49\] which states that a unital magma object in the category of unital magmas is necessarily commutative and associative. In fact, the following result holds for Taylor algebras: if \( \mathbb{A} \) is a Taylor algebra, and \( m : \mathbb{A}^2 \to \mathbb{A} \) is a homomorphism such that there exists an element \( e \in \mathbb{A} \) with \( m(e, x) = m(x, e) = x \) for all \( x \), then \( m \) is commutative and associative.

Note that our assumption on \( m \) implies that \( m * m \) satisfies the identities

\[
m(x, y) \approx m * m(x, y, e, e) \approx m * m(x, e, y, e) \approx \cdots \approx m * m(e, e, x, y),
\]

where in each \( m * m \) we always have the \( x \) occurring to the left of the \( y \). Additionally, since \( m * m : \mathbb{A}^4 \to \mathbb{A} \) is a homomorphism, for any \( n \)-ary operation \( t \) of \( \mathbb{A} \) we can evaluate \( (m * m) * t \) on a \( 4 \times n \) matrix of variables in two different ways: we may either start by applying \( t \) to the rows and then apply \( m * m \) to the resulting column vector, or we may first apply \( m * m \) to the columns and then apply \( t \) to the resulting row vector - either way gives the same result.

Using these two observations together with the Taylor identities for an \( n \)-ary Taylor term \( t \), we prove that \( m \) is commutative by writing \( m(x, y) \) as \( (m * m) * t \) applied to a \( 4 \times n \) matrix of \( e \)s, \( x \)s, and \( y \)s where every column has an \( x \) above a \( y \), and manipulate this expression until every \( y \) is above an \( x \). The strategy is to always keep the \( x \)s in the middle two rows and the \( y \)s in the top or bottom, and to move a \( y \) up a column whenever that column is free of \( x \)s. To temporarily move \( x \)s out of the way, we apply the Taylor identities for \( t \) to swap them with \( e \)s, possibly shifting the \( x \)s up and down between the middle two rows to get to a configuration where the Taylor identities will apply. A similar argument with \( m * m * m \) in the place of \( m * m \) can be used to prove associativity. If \( t \) is a 6-ary weak 3-cube term, for instance, then a portion of the proof of the commutativity of
\( m \) goes as follows:

\[
m\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = (m * m) * (m * m) * t\left(\begin{bmatrix} e & e & e & e & e & e \\ x & e & e & x & x & e \\ e & x & x & e & e & e \\ y & y & y & y & y & y \end{bmatrix}\right) = (m * m) * (m * m) * t\left(\begin{bmatrix} e & e & e & e & e & e \\ e & x & x & x & e & e \\ e & x & x & x & e & e \\ y & y & y & y & y & y \end{bmatrix}\right)
\]

\[
= (m * m) * t\left(\begin{bmatrix} y & e & e & e & e & y \\ e & x & x & x & e & e \\ e & x & x & x & e & e \\ e & y & y & y & y & e \end{bmatrix}\right) = (m * m) * t\left(\begin{bmatrix} y & e & e & e & y \\ x & e & x & x & e \\ e & x & x & e & e \\ e & y & y & y & e \end{bmatrix}\right)
\]

\[
= \cdots = (m * m) * t\left(\begin{bmatrix} y & y & y & y & y \\ e & e & x & x & e \\ e & x & e & e & x \\ e & e & e & e & e \end{bmatrix}\right) = m\left(\begin{bmatrix} y \\ x \end{bmatrix}\right),
\]

where we have used the Taylor identity \( t(x, e, e, e, x, x) \approx t(e, e, x, x, e) \) satisfied by a weak 3-cube term to temporarily move the first and last \( x \) out of the way.

### 7 Two simple algorithms (width 1 and bounded strict width)

**Definition 7.1.** A CSP template \( A = (D, \Gamma) \) has **relational width 1** if the relational width \((1, k)\) algorithm below solves it for some \( k \).

**Algorithm 1** Relational width \((1, k)\) algorithm

1: Set \( S_v \leftarrow D \) for each variable \( v \).
2: repeat
3: for all \( v_1, ..., v_k \) do
4: Let \( X \) be the set of solutions to the restriction of the CSP to the variables \( v_1, ..., v_k \) (projecting each constraint onto this subset of variables).
5: Set \( S_{v_i} \leftarrow \pi_j(X \cap (S_{v_1} \times \cdots \times S_{v_k})) \) for each \( i \leq k \).
6: For each constraint \( R \) which involves some \( v_i \), remove all tuples of \( R \) which are incompatible with \( S_{v_i} \).
7: until the sets \( S_v \) stop changing.
8: If any \( S_v = \emptyset \), there is no solution.

Compare this to the generalized arc-consistency algorithm, which is more popular (and more efficient!) in practice.

**Algorithm 2** Generalized arc-consistency algorithm

1: Set \( S_v \leftarrow D \) for each variable \( v \).
2: while some constraint \( R \) on variables \( (v_1, ..., v_m) \) has \( \pi_j(R \cap (S_{v_1} \times \cdots \times S_{v_m})) \neq S_{v_j} \) do
3: Set \( S_{v_j} \leftarrow \pi_j(R \cap (S_{v_1} \times \cdots \times S_{v_m})). \)
4: If any \( S_v = \emptyset \), there is no solution.
Theorem 7.2 (Feder, Vardi [53]). A CSP template \( A \) has relational width 1 iff it is solved by the generalized arc-consistency algorithm.

Sketch. Suppose \( A \) has width \((1, k)\), and let \( B \) be an instance of CSP(\( A \)). By a generalization of the randomized construction of graphs with large girth and large chromatic number, there is a relational structure \( B' \) which has a map to \( B \), has girth larger than \( k \), and which has a map to \( A \) if \( B \) has a map to \( A \) (alternatively, if \( \Gamma \) contains the equality relation, we can cheat by adding long chains of equalities). Since \( B' \) locally looks like a tree, the width \((1, k)\) algorithm and the generalized arc-consistency algorithm give the same results for \( B' \), so if there is no homomorphism from \( B \) to \( A \), then generalized arc-consistency applied to \( B' \) will correctly find that there is no solution. But then generalized arc-consistency applied to \( B \) will also find that there is no solution, since every deduction on \( B' \) can be mimicked on \( B \). \( \square \)

Definition 7.3. A connected relational structure is a tree if every collection of occurrences of relations with arities \( r_1, \ldots, r_k \) involves at least \( 1 + \sum_i (r_i - 1) \) distinct elements. A relational structure \( A \) has tree duality if for every \( B \), there is a map \( B \to A \) iff every tree which maps to \( B \) has a map to \( A \).

Proposition 7.4. \( A \) has width 1 iff it has tree duality.

Proof. If generalized arc-consistency shows that there is no homomorphism \( B \to A \), then we can make a proof tree that shows that some set \( S_v \) eventually becomes empty. Each node of the proof tree corresponds to the fact that some variable \( w \) of \( B \) takes values from a set \( S_w \), and the hyperedges of the proof tree are labeled by relations of \( B \). So the proof tree is actually a relational structure with a map to \( B \), and the same sequence of generalized arc-consistency deductions apply to the proof tree to show that it has no map to \( A \). \( \square \)

Remark 7.1. Essentially the same arguments apply for any width \((l, k)\), with “trees” replaced by \((l, k)\)-trees” (definition left as an exercise to the reader). Note that \((l, k)\)-trees have tree-width \( k - 1 \).

Remark 7.2. Dalmau has shown that any CSP with relational width \((2, 2)\) is also solved by generalized arc-consistency [43]. 2SAT is an example of a CSP with width \((2, 3)\) which is not solved by arc-consistency, so Dalmau’s result is best possible.

Generalized arc-consistency has a close connection with the algebraic concept of a “subdirect product”.

Definition 7.5. A subalgebra \( R \leq A_1 \times \cdots \times A_n \) is called a subdirect product, written \( R \leq_{sd} A_1 \times \cdots \times A_n \), if \( \pi_i(R) = A_i \) for all \( i \).

So an algebraic way of thinking of arc-consistency is that we shrink the domains of the variables until we get to a situation where every relation is a subdirect product. It’s worth noting that as we shrink our domains and relations, the new domains and relations we obtain will always be preserved by any polymorphisms which preserved the original relations, since the new domains and relations can be defined by primitive positive formulas from the original ones.

We now find an algebraic characterization of CSP templates with width 1. The main idea is to consider the “most generic” problem which arc-consistency requires to have a solution, and to ask what such a solution must look like. This most generic problem will have a different variable for each possible nonempty set \( S \subseteq D \), and will have all relations which are consistent with these sets imposed.
Definition 7.6. For $A = (D, \Gamma)$ a relational structure, define $P_\emptyset(A)$ to be the structure with ground set $P(D) \setminus \{\emptyset\}$, and for every $m$-ary relation $R \in \Gamma$ let the corresponding relation $P_\emptyset(R)$ be the set of all $m$-tuples $S_1, \ldots, S_m \in P(D) \setminus \{\emptyset\}$ such that there is some nonempty $X \subseteq R$ with $\pi_i(X) = S_i$ for each $i$.

Note that $P_\emptyset(R)$ can be equivalently defined as the set of $m$-tuples $(S_1, \ldots, S_m)$ such that $\pi_i(R \cap (S_1 \times \cdots \times S_m)) = S_i$ for each $i$.

Definition 7.7. A homomorphism $P_\emptyset(A) \to A$ is called a set polymorphism of $A$.

Definition 7.8. A function $f : D^k \to D$ is called totally symmetric if the value of $f(a_1, \ldots, a_k)$ only depends on $\{a_1, \ldots, a_k\}$. Note that this is stronger than being symmetric, since the multiplicity of the $a_i$s is also ignored.

Theorem 7.9. The following are equivalent:

- $A$ has width 1,
- $A$ has a set polymorphism, and
- $A$ has totally symmetric polymorphisms of every arity.

Proof. If $A$ has width 1, then generalized arc-consistency applied to $P_\emptyset(A)$ shows that there is a homomorphism $f : P_\emptyset(A) \to A$, since at every step the set associated to the variable $S \subseteq D$ will contain $S$ (by induction on the number of steps and the definition of $P_\emptyset(R)$). So suppose that $f$ is a set polymorphism, and for every $k \geq 1$, let $f_k$ be the totally symmetric function

$$f_k(a_1, \ldots, a_k) = f(\{a_1, \ldots, a_k\}).$$

We need to check that $f_k$ is a polymorphism of $A$. Suppose that $x_1, \ldots, x_k \in R$, then if $X = \{x_1, \ldots, x_k\}$, $f_k(x_1, \ldots, x_k)$ has $i$th coordinate equal to $f(\pi_i(X))$. Since $(\pi_1(X), \ldots, \pi_m(X)) \in P_\emptyset(R)$ by the definition of $P_\emptyset(R)$, we see that $f_k(x_1, \ldots, x_k) = (f(\pi_1(X)), \ldots, f(\pi_m(X))) \in R$.

Finally, suppose that $A$ has totally symmetric polymorphisms $f_k$ of every arity, and let $B$ be a (finite) instance such that generalized arc-consistency stops after finding nonempty sets $S_v$ for every variable $v \in B$. Choose $k$ at least as large as the largest number of tuples in any relation that shows up in $B$, and let $f$ be the function on sets of size $\leq k$ associated to $f_k$. We claim that the map $v \mapsto f(S_v)$ defines a homomorphism from $B$ to $A$. To see this, let $(v_1, \ldots, v_m)$ be a tuple with the constraint $R$ imposed, and let $X = R \cap (S_{v_1}, \ldots, S_{v_m}) = \{x_1, \ldots, x_k\}$ (possibly with repeated $x_i$s if $|X| < k$). Then $f_k(x_1, \ldots, x_k) = (f(\pi_1(X)), \ldots, f(\pi_m(X))) = (f(S_{v_1}), \ldots, f(S_{v_m})) \in R$ since $f_k$ is a polymorphism.

Corollary 7.10. A relational structure $A$ has width 1 iff it is homomorphically equivalent to a pp-power of HORN-SAT.

Proof. Let $f$ be a set polymorphism of $A$, and let $f_k$ be the associated totally symmetric polymorphism of arity $k$. We define a height 1 clone homomorphism from $(\min) \to \Pol(A)$ by sending $\min(x_1, \ldots, x_k)$ to $f_k(x_1, \ldots, x_k)$. Now apply the ERP Theorem 5.15 and Proposition 5.14 from the subsection on reflections. 

39
Example 7.1. Suppose that $A$ has a binary polymorphism $s$ which is associative, commutative, and idempotent (such an $s$ is called a semilattice operation). Then we can define $n$-ary polymorphisms $s_n$ inductively by $s_n(x_1, \ldots, x_n) = s(s_{n-1}(x_1, \ldots, x_{n-1}), x_n)$, and $s_n$ will be totally symmetric for every $n$. Thus, every relational structure with a semilattice polymorphism has width 1.

Example 7.2. We give an example of a width 1 algebra which is not a semilattice. Let $f$ be the idempotent set operation on $\{a, b, c\}$ given by

$$f(\{a, b\}) = b, \quad f(\{b, c\}) = c, \quad f(\{c, a\}) = a, \quad f(\{a, b, c\}) = a,$$

and let $f_k$ be the associated totally symmetric polymorphism of arity $k$. We have $f_k \in \langle f_3 \rangle$ for every $k$, and in fact a $k$-ary function $g$ which depends on all its inputs is in $\langle f_3 \rangle$ iff its restriction to every two element subset of $\{a, b, c\}$ is equal to the corresponding restriction of $f_k$ (tricky exercise). The relational clone $\text{Inv}(f_3)$ is generated by the ternary relations $R_{ab}, R_{bc}, R_{ca}$, where $R_{ab}$ is defined by

$$R_{ab}(x, y, z) := (x \in \{a, b\}) \land (x = a \Rightarrow y = z),$$

and $R_{bc}, R_{ca}$ are defined similarly.

Example 7.3. Here we give a more surprising example, of a width 1 clone such that no finitely generated subclone has width 1. Let $f$ be the idempotent set operation on $\{-1, 0, 1\}$ (which we stylize as $\{-, 0, +\}$) given by

$$f(\{0, -\}) = -, \quad f(\{0, +\}) = +, \quad f(\{-, +\}) = f(\{-, 0, +\}) = 0,$$

and let $f_k$ be the associated totally symmetric polymorphism of arity $k$. The clone $O$ generated by the collection of all $f_k$ then has width 1. Every finitely generated subclone of $O$ is contained in $\langle f_k \rangle$ for some $k$. To see that $O \neq \langle f_k \rangle$, consider the $k + 1$-ary relation $R_k$ given by

$$R_k(x_0, \ldots, x_k) := \bigwedge_{i<j} (x_i + x_j \geq 0) \land (x_0, \ldots, x_k) \neq (0, \ldots, 0).$$

Then it is easy to check that $R_k$ is preserved by $f_k$, but is not preserved by $f_{k+1}$. To see that $\langle f_k \rangle$ does not have width 1, define $R_k^{-}$ similarly to $R_k$, but with $x_i + x_j \geq 0$ replaced by $x_i + x_j \leq 0$. Then for $k \geq 2$ the instance

$$R_k(x_0, \ldots, x_k) \land R_k^{-}(x_0, \ldots, x_k)$$

of $\text{CSP}(\text{Inv}(\langle f_k \rangle))$ is arc-consistent (since both $R_k$ and $R_k^{-}$ are subdirect) but has no solution.

The relational clone $\text{Inv}(O)$ corresponding to this example is generated by the unary relation $\{+\}$, the binary relations $x = -y$ and $x \leq y$, and the ternary relation $(x \geq 0) \land (x = 0 \Rightarrow y = z)$. The clone $O$ is an example of a clone which is finitely related but not finitely generated.

Note that one doesn’t need to know what the set polymorphism of $A$ is to apply the arc-consistency algorithm. If $A$ is a rigid core, we can use the self-reducibility of $\text{CSP}(A)$ to find a solution to every solvable instance $B$ of $\text{CSP}(A)$ in polynomial time. By applying this to $\mathcal{P}_0(A)$, we can then find a set polymorphism of $A$ - in time polynomial in the size of $\mathcal{P}_0(A)$, which is sadly exponential in the size of $A$. The following problem is currently open.

Problem 7.1. Given a rigid core $A$, can we determine whether it has width 1 in time polynomial in the size of the description of $A$?
Now we move to the case of bounded strict width. This has a connection to an intriguing paper of Dechter \[46\] which predates the algebraic approach to the CSP. The next definition follows Dechter \[46\].

**Definition 7.11.** A partial assignment of values to variables is *locally consistent* if it satisfies every constraint which only involves those variables. A CSP instance is *strong i-consistent* if every locally consistent partial assignment to less than i variables can always be extended to a locally consistent partial assignment of any containing set of i variables. An instance is *globally consistent* if every locally consistent partial assignment extends to a global solution.

There is a straightforward polynomial time algorithm to enforce strong i-consistency for any fixed i, introducing new constraints of arity < i by intersecting and existentially projecting old constraints until no changes occur. It is desirable to have globally consistent problems, because then a solution may be found greedily. Can we check if a given problem is globally consistent?

**Theorem 7.12** (Dechter \[46\]). If a CSP with domain sizes bounded by n and all constraint arities bounded by m is strong \((n(m - 1) + 1)\)-consistent, then it is globally consistent.

**Proof.** Suppose for contradiction that some locally consistent partial assignment \(a_1, ..., a_k\) to \(v_1, ..., v_k\) can’t be extended to \(v_{k+1}\), \(k \geq n(m - 1) + 1\). Then for every possible value \(a\) of \(v_{k+1}\), there is some constraint \(C_a\) involving at most \(m - 1\) of the variables \(v_1, ..., v_k\) which is inconsistent with this choice of \(a\) and whichever of the \(a_i\)'s are relevant. Thus, there is a collection of at most \(n\) constraints \(C_a\) involving at most \(n(m - 1)\) of the variables from \(v_1, ..., v_k\) together with the variable \(v_{k+1}\), for which a locally consistent partial assignment of all but one of the variables can’t be extended. But this contradicts the assumption of strong \((n(m - 1) + 1)\)-consistency. \(\Box\)

The trouble with applying Dechter’s result is that as we enforce strong consistency, we may need to add constraints of higher and higher arities. To avoid this, we want to find situations in which the newly introduced constraints can always be written as intersections of constraints of low arity.

**Definition 7.13.** A CSP template \(A = (D, \Gamma)\) has *strict width* \(l\) if every strong \((l + 1)\)-consistent instance of CSP\((D, \langle \Gamma \rangle)\) which contains the projections of its relations onto subsets of size at most \(l\) is globally consistent and has its solution-set determined by the collection of relations of arity at most \(l\).

Note that the definition of strict width only makes sense in terms of the whole relational clone generated by \(\Gamma\), a hint that it is properly viewed as an algebraic condition. Algebraically, the relevant result is the Baker-Pixley theorem \[7\].

**Theorem 7.14** (Baker, Pixley \[7\]). The following are equivalent for an algebraic structure \(A\):

- every subalgebra of \(A^n\) is equal to the intersection of its projections onto sets of at most \(l\) coordinates, and
- \(A\) has an \((l + 1)\)-ary near-unanimity term, that is, a term \(t\) satisfying the identities

\[
x \approx t(y, x, ..., x) \approx t(x, y, ..., x) \approx t(x, x, ..., y),
\]

where in each case all but one of the inputs to \(t\) is \(x\).
If \( \mathbb{A} \) is idempotent, then these are both equivalent to every subalgebra of \( \mathbb{A}^{l+1} \) being equal to the intersection of its projections onto sets of \( l \) coordinates.

**Proof.** Note that if \( |\mathbb{A}| \geq 2 \), then either condition implies \( l > 1 \) (consider the equality relation as a subalgebra of \( \mathbb{A}^2 \)). Suppose that the first condition holds, and consider the free algebra on \( l + 1 \) generators \( \mathcal{F}_\mathbb{A}(l+1) \subseteq \mathbb{A}^{l+1} \) which is generated by the projections \( \pi_i : \mathbb{A}^{l+1} \to \mathbb{A} \). Let \( \mathbb{A}^{l+1}_{nu} \) be the set of tuples of elements in \( \mathbb{A}^{l+1} \) which have all but at most one entry equal to each other, and let \( X \subseteq \mathbb{A}^{l+1}_{nu} \) be the projection of \( \mathcal{F}_\mathbb{A}(l+1) \) onto these coordinate tuples.

We claim that \( X \) contains the tuple \( t \) of near-unanimous votes of the entries of the coordinate tuples. By assumption, we just have to check that for every projection \( \pi_{x_1,..,x_l}(X) \) onto at most \( l \) coordinates \( x_1,..,x_l \in \mathbb{A}^l \), there is some element \( f \in \mathcal{F}_\mathbb{A}(l+1) \) with \( \pi_{x_1,..,x_l}(f) = \pi_{x_1,..,x_l}(t) \). But each tuple \( x_i \) has at most one dissenting coordinate, so there must be some coordinate \( j \leq l+1 \) such that each \( (x_i)_j \) is equal to the vote \( t(x_i) \). Thus we can take \( f = \pi_j \) to see that \( \pi_{x_1,..,x_l}(\pi_j) = \pi_{x_1,..,x_l}(t) \).

Now suppose that \( \mathbb{A} \) is idempotent and \( \mathbb{B} \subseteq \mathbb{A}^n \) with \( n > l + 1 \) and \( b \in \mathbb{A}^n \) be such that \( \pi_I(b) \in \pi_I(\mathbb{B}) \) for every \( I \subseteq \{1,..,n\} \) with \( |I| \leq l \), we will show by induction on \( |J| \) that \( \pi_I(b) \in \pi_J(\mathbb{B}) \) for every subset \( J \subseteq \{1,..,n\} \). For the inductive step, if \( |J| \geq l + 1 \) then we may set \( J_1,..,J_{l+1} \) to be subsets of \( J \) formed by deleting different elements of \( J \), and for each \( J_i \) there is some \( b_{J_i} \in \mathbb{B} \) with \( \pi_{J_i}(b_{J_i}) = \pi_{J_i}(b) \) by induction. But then \( b_{J_i} = t(b_{J_1},..,b_{J_{l+1}}) \in \mathbb{B} \) and has \( \pi_{J_i}(b_{J_i}) = \pi_{J_i}(b) \) by the near-unanimity equations.

For the last claim, if \( \mathbb{A} \) is idempotent and \( \mathbb{B} \subseteq \mathbb{A}^n \) with \( n > l + 1 \) and \( b \in \cap_{|I|\leq l} \pi_I(\mathbb{B}) \), then \( \mathbb{B}' = \pi_{\{1,..,n-1\}}(\mathbb{B} \cap (\mathbb{A}^{n-1} \times \{b_n\})) \) is a subalgebra of \( \mathbb{A}^{n-1} \), and we may induct on \( n \) to see that \( \mathbb{B}' = \cap_{|I|\leq l} \pi_I(\mathbb{B}') \), while the assumption on subalgebras of \( \mathbb{A}^{l+1} \) gives \( \pi_I(b) \in \pi_I(\mathbb{B}') \) for every \( I \) with \( |I| = l \).

**Theorem 7.15.** A relational structure \( \mathbb{A} \) has strict width \( l \) if it has an \( (l+1) \)-ary near-unanimity polymorphism.

**Proof.** Let \( \mathbb{A} \) be the associated algebraic structure. For any \( n \) and any \( \mathbb{B} \subseteq \mathbb{A}^n \), the strong \( (l+1) \)-consistent instance formed via the relations \( \mathbb{B} \) and \( \pi_I(\mathbb{B}) \) for all \( I \subseteq \{1,..,n\} \) with \( |I| \leq l \) together with the definition of strict width \( l \) imply that \( \mathbb{B} = \cap_{|I|\leq l} \pi_I(\mathbb{B}) \), so by the Baker-Pixley Theorem \( \mathbb{A} \) has an \((l+1)\)-ary near unanimity term.

For the other direction, suppose that \( t \) is an \((l+1)\)-ary near-unanimity term of \( \mathbb{A} \) and that we have a strong \((l+1)\)-consistent instance of \( \text{CSP}(\mathbb{A}) \), which we may assume by the Baker-Pixley Theorem to only involve relations of arity at most \( l \). Suppose that we have a locally consistent partial solution which assigns the values \( a_1,..,a_k \) to the variables \( v_1,..,v_k \) which we want to extend to the variable \( v_{k+1} \). By strong \((l+1)\)-consistency, we can assume that \( k \geq l + 1 \). By induction on \( k \), we can assume that for each \( i \leq l + 1 \) there is some value \( a_{i+1}^{k+1} \) that we can assign the the variable \( v_{k+1} \) such that if we ignore \( v_i \), we get a locally consistent partial solution.

We claim that assigning the value \( a_{k+1} = t(a_{k+1}^1,..,a_{k+1}^{l+1}) \) to \( v_{k+1} \) gives a locally consistent partial solution. To see this, consider some constraint \( C \) which involves the variable \( v_{k+1} \) and some variables from \( v_1,..,v_k \). For each \( i \leq l + 1 \), by \( l \)-consistency and the fact that \( C \) has arity at most \( l \) we can find a value \( a_i' \) such that \( (a_1,..,a_i',..,a_{k+1}) \) satisfies the constraint \( C \). Applying \( t \) to these \( l + 1 \) tuples, we see that the tuple \( (a_1,..,a_{l+1},..,t(a_{k+1}^1,..,a_{k+1}^{l+1})) \) also satisfies \( C \), by the near-unanimity identities and the fact that \( t \) is a polymorphism of \( C \).
Algorithm 3 Strict width $l$ algorithm

1: Replace each constraint with its projections onto all subsets of at most $l$ variables.
2: repeat
3: for all sets $\{v_1, ..., v_k\}$ of variables with $k \leq l + 1$ do
4: Let $X$ be the set of solutions to the restriction of the CSP to the variables $v_1, ..., v_k$.
5: If $\pi_I(X)$ is not implied by the restriction of the CSP to the variables in $I$ for some $I \subset \{v_1, ..., v_k\}$, add it as a new constraint.
6: until no new constraints are added.
7: Greedily assign values to variables until we find a global solution.

Example 7.4. 2SAT has the ternary polymorphism $\text{maj}$, which is a near-unanimity operation. Therefore 2SAT has strict width 2, a fact which also follows from Dechter’s result above [46].

Example 7.5. Generalizing 2SAT, let $D$ be any domain, and let $d : D^3 \rightarrow D$ be given by

$$d(x, y, z) = \begin{cases} x & \text{if } y \neq z, \\ y & \text{if } y = z. \end{cases}$$

This function $d$ is known as the dual discriminator, and for $|D| \neq 4$ it is the only majority function (up to permuting inputs) on $D$ which preserves the graph of every bijection from $D$ to itself.

A binary relation $R \subseteq D^2$ is preserved by the dual discriminator iff it is a “0/1/all constraint”, that is, a constraint such that when viewed as a bipartite graph on the disjoint union $D \cup D$, every vertex which doesn’t have degree 0 or 1 connects to all vertices on the other side which have positive degree. Typical 0/1/all constraints are displayed below.

For any $a$ in $D$, a generating set of binary relations for $\text{Inv}(d)$ is given by the graphs of a pair of bijections which generate the symmetric group on $|D|$ elements, the unary relation $D \setminus \{a\}$, and the binary relation $x = a \lor y = a$.

Example 7.6. For every $n$, the relational structure $((\{0, 1\}, \{0\}, \{0, 1\}^n \setminus \{(0, ..., 0)\})$ has strict width exactly $n$. A near-unanimity term for it is given by the threshold function

$$t_{2}^{n+1}(x_1, ..., x_{n+1}) = \begin{cases} 1 & \sum_i x_i \geq 2, \\ 0 & \sum_i x_i \leq 1. \end{cases}$$

To see that it doesn’t have strict width less than $n$, note that the relation $\{0, 1\}^n \setminus \{(0, ..., 0)\}$ is not the intersection of its projections onto $n - 1$ coordinates. Note that this template also has width 1 (it is preserved by the semilattice operation max), so the strict width algorithm is far from being the best way to solve it for $n$ large.
Note that the existence of an \((l + 1)\)-ary near-unanimity operation in \(\text{Pol}(A)\) is equivalent to the solvability of the CSP instance \(\Phi\) (of \(A\) together with singleton unary relations) with variables indexed by elements of \(A^{l+1}\) described by the primitive positive formula

\[
\Phi(t) := \bigwedge_{R \in \Gamma} \bigwedge_{M \in R^{l+1}} t(M) \in R \land \bigwedge_{a,b \in A} t(b, a, ..., a) = t(a, b, ..., a) = \cdot = t(a, a, ..., b) \in \{a\}.
\]

This instance may be solved in polynomial time by the strict width \(l\) algorithm, giving us an \((l + 1)\)-ary near-unanimity term \(t\) as output. Note, however, that the number of variables is exponential in \(l\) - what if we just want to know whether the structure \(A\) has bounded strict width, allowing \(l\) to be arbitrarily large?

**Problem 7.2.** Given a relational structure \(A\), determine whether it has bounded strict width.

The good news is that whether the structure is given as a finite relational structure or a finite algebraic structure, the existence of a near unanimity term is at least decidable [87], [9], [113]. The bad news is that the minimal arity of a near-unanimity term may be very large.

**Theorem 7.16 (Zhuk, Barto, Draganov [12], [113]).** Every relational structure \(A\) with \(|A| = n\) and maximal arity of any relation \(m\) which has bounded strict width, has strict width at most \(\frac{1}{2}(2m - 2)^{3n}\). For each \(m \geq 3\) and \(n \geq 2\), there is a relational structure with bounded strict width which has no near-unanimity polymorphism of arity at most \((m - 1)^{2n-2}\), and for \(m = 2, n \geq 3\) there is an example with no near-unanimity polymorphism of arity at most \(2^{2n-3}\).

Luckily, it is possible to determine whether a relational structure has bounded strict width without actually exhibiting a near-unanimity polymorphism. For instance, in [11] a nondeterministic polynomial time algorithm which only tests for the existence of certain chains of ternary polymorphisms of \(A\) is given for deciding whether a given subset of \(A\) is an absorbing subalgebra (defined later). Using the fact that cycle consistency solves CSPs which have bounded width (which we will prove later), this can be converted into a polynomial time algorithm for testing whether \(A\) has bounded strict width.

### 7.1 The Basic LP relaxation of a CSP

Another simple algorithm for solving CSPs, which is closely related to generalized arc-consistency, is the basic LP relaxation. If the domain of each variable \(v\) is \(D_v\), we replace the set of potential values \(D_v\) with its formal convex hull, which we can think of as the set of probability distributions on \(D_v\). We represent the probability distribution corresponding to a variable \(v\) as a tuple of real numbers \(p_{v,a}\), one for each \(a \in D_v\), satisfying

\[
0 \leq p_{v,a} \leq 1, \sum_{a \in D_v} p_{v,a} = 1.
\]

We also replace each constraint with its convex hull. That is, if the constraint \(C\) imposes the relation \(R = R_C\) on the variables \(v_1, ..., v_m\), then we require the existence of a probability distribution \(p_{C,r}\) on the tuples \(r\) of \(R\) such that

\[
0 \leq p_{C,r} \leq 1, \sum_r p_{C,r} = 1,
\]

44
and which is compatible with the probability distributions on the individual variables in the sense that
\[ p_{v_i,a} = \sum_{r_i=a} p_{C,r}. \]

If a problem is known not to be fully satisfiable, we can relax it further by extending the probability distributions over relations \( R \subseteq D_{v_1} \times \cdots \times D_{v_n} \) to probability distributions over all of \( D_{v_1} \times \cdots \times D_{v_m} \), and then try to maximize the sum of the probabilities that tuples which are supposed to be in \( R \) are actually in \( r \):
\[ \frac{1}{\#C} \sum_C \sum_{r \in R_C} p_{C,r}. \]

This system of linear equations and inequalities, with the optimization target above, is known as the basic LP relaxation of a given CSP instance.

**Theorem 7.17** (Kun, O’Donnell, Tamaki, Yoshida, Zhou [SI]). For any relational structure \( A \), the following are equivalent:

- the basic LP relaxation correctly solves every instance of CSP(\( A \)),
- \( A \) has symmetric polymorphisms of every arity.

Furthermore, if \( A \) has width 1 then the basic LP relaxation can be used to robustly solve CSP(\( A \)), that is, if we are given an instance which is \( 1-\epsilon \) satisfiable, then we can find a solution which satisfies a \( 1 - O(1/\log(1/\epsilon)) \) fraction of the constraints.

**Proof.** Suppose first that the basic LP solves CSP(\( A \)), and consider the (by now standard) instance \( \Phi \) that describes the existence of a symmetric polymorphism of arity \( n \):
\[ \Phi(s) := \bigwedge_{R \in \Gamma} \bigwedge M \in R^n s(M) \in R \land \bigwedge_{a_1, \ldots, a_n} \bigwedge_{\sigma \in S_n} s(a_1, \ldots, a_n) = s(a_{\sigma(1)}, \ldots, a_{\sigma(n)}). \]

By the assumption that the basic LP decides CSP(\( A \)), we just need to exhibit a fractional solution to this CSP. This is achieved by taking \( s = \frac{1}{n!} \pi_1 + \cdots + \frac{1}{n!} \pi_n \): as a convex combination of polymorphisms, it satisfies the relaxation of the first collection of constraints, and since it is a symmetric convex combination of its inputs it satisfies the second collection of constraints.

For the other direction, suppose that an instance of the CSP has a fractional solution to its basic LP relaxation, with probability distributions \( p_{v_i,a} \) for each variable/value and \( p_{C,r} \) for each constraint/tuple. We may assume that these probabilities are all rational (since the defining system of linear equations and inequalities had rational coefficients), and that they have a common denominator \( n \). By assumption \( A \) has a symmetric polymorphism \( s \) of arity \( n \), which we can think of as a function from probability distributions with denominator \( n \) over the domain of \( A \) to elements of \( A \).

Applying \( s \) to each \( p_{v_i} \), gives an element \( a_v \in A \), and applying it to each probability distribution \( p_{C} \), gives a tuple \( r_C \) in the associated relation \( R \) (since \( s \) is a polymorphism). Furthermore, the compatibility equations between the distributions \( p_{v_i} \) and \( p_{C} \) that we get when \( v_i \) is the \( i \)th coordinate of the constraint \( C \), together with the symmetry of \( s \), imply that \( a_v = (r_C)_i \) for each \( i \), so \( (a_{v_1}, \ldots, a_{v_m}) = r_C \in R \). Thus the \( a_v \)s form a valid solution to the CSP instance.
Finally, assume that $A$ has width 1, with set polymorphism $f$, and suppose that our original instance was $1 - \epsilon$ satisfiable. Then the basic LP finds a fractional solution with value $\geq 1 - \epsilon$. We will use the polymorphism $f$ to make a randomized rounding scheme. First, we immediately give up on any constraints $C$ that the LP only satisfies with value $\leq 1 - \sqrt{\epsilon}$ - these can form at most a $\sqrt{\epsilon}$ fraction of the constraints by Markov’s inequality. Second, we will choose a threshold $\theta \leq \frac{1}{|A|}$, and for each variable $v$ we assign the value

\[ a_v = f(\{a \in A \mid p_v,a \geq \theta\}). \]

Note that the restriction $\theta \leq \frac{1}{|A|}$ ensures that the sets on the right hand side are nonempty. We will show that if $\theta$ is chosen from a certain probability distribution, then on average we will obtain a good solution to the CSP, and deduce from this that some specific choice of $\theta$ works at least as well. For this we need the following claim.

**Claim.** If $C$ is the constraint $(v_1,\ldots,v_m) \in R$ which is satisfied with value $\geq 1 - \sqrt{\epsilon}$, and if $2\sqrt{\epsilon} \leq \theta \leq \frac{1}{|A|}$ is such that

\[ \theta \notin (p_v,a/(2|R|),p_v,a) \]

for any pair $i \leq m$, $a \in A$, then $(a_{v_1},\ldots,a_{v_m})$ satisfies $C$.

**Proof of Claim.** For each $v$, let $S_v = \{a \mid p_v,a \geq \theta\}$, so $a_v = f(S_v)$. In order to show that $(a_{v_1},\ldots,a_{v_m})$ satisfies $R$, we just need to check that this collection of sets $S_{v_i}$ together with $R$ form a generalized arc-consistent instance. Let $a \in S_{v_i}$ for some $i$, then we have $p_{v_i,a} \geq \theta \geq 2\sqrt{\epsilon}$ by the definition of $S_{v_i}$. From

\[ \sum_{r \in R, r_i = a} p_{C,r} \geq p_{v_i,a} - \sqrt{\epsilon} \geq p_{v_i,a}/2, \]

we see that there must be some $r \in R$ with $r_i = a$ and $p_{C,r} \geq p_{v_i,a}/(2|R|)$. Since $p_{v_i,a} \geq \theta$, by the assumption on $\theta$ we have $p_{v_i,a}/(2|R|) \geq \theta$, so $p_{C,r} \geq \theta$. But then $p_{v_j,r_j} \geq p_{C,r} \geq \theta$ for all $j$, so $r_j \in S_{v_j}$ for all $j$, and we see that $a$ extends to a solution of $R \cap (S_{v_1} \times \cdots \times S_{v_m})$.

To finish the proof, we choose $\theta$ uniformly at random from the set $\left\{ \frac{1}{|A|}, \frac{1}{|A|^T}, \cdots, \frac{1}{|A|^T} \right\}$, where $T$ is twice the maximum number of tuples in any relation $R$ and $b = \lfloor \log(1/2|A|/\sqrt{\epsilon})/\log(T) \rfloor$. Note that $b$ grows like $\log(1/\epsilon)$, that’s the only important thing to keep track of in the mess. Then every constraint of arity $m$ which we hadn’t given up on is satisfied with probability at least $1 - m|A|/b$ (since there are at most $m|A|$ bad choices of $\theta$ where the claim doesn’t apply), and asymptotically that looks like $1 - O(1/\log(1/\epsilon))$. □

**Remark 7.3.** The dependence of the error in $1 - O(1/\log(1/\epsilon))$ on $\epsilon$ in the previous theorem is best possible in the case of HORN-SAT: Guruswami and Zhou [60] show that there are integrality gap instances even for the SDP relaxation, and by a fundamental result of Raghavendra [100] they deduce that under the Unique Games conjecture it is NP-hard to find an assignment satisfying a $1 - o(1/\log(1/\epsilon))$ fraction of the constraints.

**Remark 7.4.** In [81], it is also claimed that the basic LP solves every instance of CSP($A$) if and only if $A$ has width 1. The proof has a subtle error, however. The following counterexample, due to Kun, can be found in [34].

**Example 7.7.** Let $A = \{-1,0,1\}, R_+, R_-$, where $R_+ = \{(a,b,c) \mid a + b + c \geq 1\}$ and $R_-$ =
\{(a, b, c) \mid a + b + c \leq -1\}. Then for every \(h, n\) with \(h < \frac{n}{3}\), the function

\[
s_{h,n}(x_1, \ldots, x_n) = \begin{cases} 1 & \sum_i x_i > h \\ 0 & -h \leq \sum_i x_i \leq h \\ -1 & \sum_i x_i < -h \end{cases}
\]

is a symmetric polymorphism of \(A\). Thus \(\text{CSP}(A)\) is solved by the basic LP relaxation. However, \(A\) has no totally symmetric polymorphism of arity 3, since such a polymorphism would necessarily map the matrices

\[
\begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix} \in R^3_+ \\
\begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & -1 & 1 \end{bmatrix} \in R^3_-
\]

to the same diagonal tuple, so \(A\) does not have width 1.

**Example 7.8.** The previous example can be generalized to a much larger relational structure on \([-1, 0, 1]\) as follows. Set \(s_n = s_{0,n}\), then it isn’t hard to show that \(s_n \in \text{Clo}(s_2)\) for all \(n\) (hint: start by defining \(t_n(x_1, \ldots, x_n) = s_2(x_1, s_{n-1}(x_2, \ldots, x_n))\)), so \(\text{Inv}(s_2)\) also defines a CSP template which is solved by the basic LP relaxation.

<table>
<thead>
<tr>
<th>(s_2)</th>
<th>(- 0)</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-)</td>
<td>(-)</td>
<td>0</td>
</tr>
<tr>
<td>(0)</td>
<td>(-)</td>
<td>+</td>
</tr>
<tr>
<td>(+)</td>
<td>(0)</td>
<td>(+)</td>
</tr>
</tbody>
</table>

\(\text{Inv}(s_2)\) is generated by the relations \(\{1\}\), \(x = -y\), and the set of odd cycle relations, where the \(m\)-th odd cycle relation \(R_m\) is defined by

\[
R_m(x_1, \ldots, x_{2m-1}, y, z) := (x_1 + x_2 \geq 0) \land \cdots \land (x_{2m-1} + x_1 \geq 0) \land (x_1 = \cdots = x_{2m-1} = 0 \implies y = z).
\]

(I found this set of generating relations by a technique I learned from Zhub [114], in which we search for “key” relations \(R\), for which there is some “key tuple” \(x \notin R\) such that the relation \(R\) is maximal among those relations of \(\text{Inv}(s_2)\) which do not contain \(x\). It isn’t hard to show that any key tuple must consist mostly of 0s, and using the negation symmetry we can assume that \(R\) contains all tuples in \(\{0, 1\}^n\) aside from the key tuple. Then we look at the set of pairs of coordinates that can’t simultaneously be set to \(-1\), and prove that the resulting graph can’t be bipartite...)

The clone \(\langle s_2 \rangle\) is not finitely related. To see this, define an operation \(s'_n\) for \(n\) odd by the rule

\[
s'_n(x_1, \ldots, x_n) = \begin{cases} s_{0,n}(x_1, \ldots, x_n) & \text{if some } x_i = 0, \\ s_{1,n}(x_1, \ldots, x_n) & \text{if all } x_i \in \{-1, 1\}. \end{cases}
\]

For every odd \(n = 2m - 1\), the operation \(s'_n \notin \langle s_2 \rangle\) - since it does not preserve the relation \(R_m\) - but the function \(s'_n(x, x, y_3, \ldots, y_n)\) we get by identifying two of its inputs is in \(\langle s_2 \rangle\) (exercise for the reader), so it preserves every relation in \(\text{Inv}(s_2)\) which contains strictly less than \(n\) tuples.

The clone \(\langle s_2 \rangle\) is strictly contained in the width 1 clone from Example 7.3, and corresponds to a strictly larger relational clone with a tractable CSP. Later we will see that this relational clone can be enlarged further, such that the CSP remains solvable by bounded width reasoning.

Currently it is unknown if the following problem is decidable.
Problem 7.3. Given a finite relational structure \( A \), determine if it has symmetric polymorphisms of every arity.

An interesting result in this direction is proved in [38]: an algebraic structure \( A \) has symmetric polymorphisms of all arities iff there is no \( B \in HSP(A) \) which has a pair of automorphisms in \( \text{Aut}(B) \) having no common fixed point (in fact, if \( A \) has no symmetric polymorphism of arity \( n \), we can take \( B \) to be the free algebra on \( n \) variables in the variety generated by \( A \)). If \( HSP(A) \) could be replaced by \( HS(A) \) in their result, then this would imply that it is enough to check for the existence of symmetric polymorphisms of arities up to \( |A| \).

Later we will prove that any Taylor algebra has cyclic polymorphisms of all arities which have no small prime factors, so we might hope that we could use these to help construct symmetric polymorphisms of higher arity. More ingredients are likely needed for such an argument, however: in [38], an example is given of a relational structure which has cyclic polymorphisms of every arity, but which has no symmetric polymorphism of arity 5.

8 Mal’cev algebras

The goal in this section and the next is to generalize group theoretic algorithms (such as the algorithm for solving XOR-SAT) by isolating the special feature of groups which makes them so nice. First we should connect groups to CSPs, by defining the correct analogue of “affine spaces” for general groups.

Proposition 8.1. If \( G \) is a group, then a nonempty subset \( H \subseteq G^n \) is preserved by the ternary operation \((x,y,z) \mapsto xy^{-1}z \) iff \( H \) is a coset of a subgroup of \( G^n \).

Proof. Let \( U \) be the subgroup of \( G^n \) generated by expressions of the form \( y^{-1}z \) for \( y, z \in H \). Then \( H \) is preserved under \((x,y,z) \mapsto xy^{-1}z \) iff \( H \) is closed under the right action of \( U \), so \( H \) is a union of left cosets of \( U \). To see that \( H \) is just a single coset, note that for \( x, y \in H \), we have \( x^{-1}y \in U \) and \( x(x^{-1}y) = y \).

Conversely, if \( H = hU \) for some subgroup \( U \) of \( G^n \), then \( HH^{-1} = hU(hU)^{-1}hU = hUUh^{-1}hU = hUUU = hU = H \).

The idempotent operation \((x,y,z) \mapsto xy^{-1}z \) was isolated by universal algebraists who wanted to understand the underlying reason for the fact that normal subgroups commute: if \( K, N \lhd G \) are normal subgroups of a group \( G \), then \( KN = NK \) and \( KN \) is also a normal subgroup of \( G \). Of course this is easy to verify in the context of groups, but from the point of view of universal algebra it is really saying something interesting about congruences of groups. If \( K, N \) correspond to congruences \( \alpha, \beta \) on \( G \), then we can view this equality as the statement that \( \alpha \circ \beta = \beta \circ \alpha = \alpha \lor \beta \), where composition of binary relations is defined as follows.

Definition 8.2. Let \( R, S \) be binary relations \( R \subseteq A \times B, S \subseteq B \times C \). Then we define their composition \( R \circ S \) to be the subset of \( A \times C \) consisting of pairs \((a, c)\) such that there exists a \( b \in B \) with \( aRb \) and \( bSc \). As a primitive positive formula, we can write this as

\[ R \circ S(a, c) := \exists b \in B \ R(a, b) \land S(b, c). \]

In general, it is not the case that congruences commute. In order to find the smallest congruence containing a pair of congruences in a general algebraic structure, one uses the following fact.
Proposition 8.3. If $\alpha, \beta$ are congruences on an algebraic structure $A$, then their least upper bound $\alpha \vee \beta$ is the transitive closure of $\alpha \circ \beta$, that is,

$$\alpha \vee \beta = \bigcup_{n \geq 0} (\alpha \circ \beta)^{\circ n}.$$ 

If $\alpha, \beta$ do commute, then the above formula simplifies to $\alpha \vee \beta = \alpha \circ \beta$. So it is natural to try to understand the collection of all algebraic structures with commuting congruences. Of course, a structure with no congruences at all has this property - but we want to understand algebraic structures that have a reason for their congruences to commute, so rather than studying algebras in isolation we study varieties with this property.

Definition 8.4. We say that a variety $V$ is congruence permutable if for all $A \in V$ and all $\alpha, \beta \in \text{Con}(A)$ we have $\alpha \circ \beta = \beta \circ \alpha$.

Theorem 8.5. A variety $V$ is congruence permutable iff $V$ has a ternary term $p$ which satisfies the identity

$$p(x,y,y) \approx p(y,y,x) \approx x.$$ 

Proof. Suppose first that $V$ is congruence permutable. Let $F = F_V(x,y,z)$ be the free algebra on three generators in $V$. Define a congruence $\alpha$ on $F$ to be the least congruence with $x/\alpha = y/\alpha$, that is, $\alpha$ is the kernel of the homomorphism $F_V(x,y,z) \to F_V(x,z)$ given by $x,y \mapsto x,z \mapsto z$. Similarly, let $\beta$ be the least congruence on $F$ with $y/\beta = z/\beta$.

Then $(x,z) \in \alpha \circ \beta$, so if $V$ has commuting congruences, then there must be some $p(x,y,z) \in F$ such that $x/\beta = p(x,y,z)/\beta$ and $p(x,y,z)/\alpha = z/\alpha$. But this is equivalent to the pair of identities $x \approx p(x,y,y), p(x,x,z) \approx z$.

Conversely, suppose such a term $p$ exists, and let $A \in V$ and $\alpha, \beta \in \text{Con}(A)$. Then for any $a,b,c$ with $a/\alpha = b/\alpha$ and $b/\beta = c/\beta$ we have

$$p(a,b,c)/\beta = p(a,b,b)/\beta = a/\beta$$

and

$$p(a,b,c)/\alpha = p(a,a,c)/\alpha = c/\alpha,$$

so $(a,c) \in \beta \circ \alpha$. 

Definition 8.6. A ternary term $p$ is called a Mal’cev term if it satisfies the identity $p(x,y,y) \approx p(y,y,x) \approx x$. An algebra with a Mal’cev term is called a Mal’cev algebra, and a variety with a Mal’cev term is called a Mal’cev variety.

One reason universal algebraists like congruence permutability is that it implies that the congruence lattice is modular, a property first isolated by Dedekind in his investigation of the lattice of submodules of a module over a ring.

Definition 8.7. A lattice $L$ is modular if for all $\alpha, \beta, \gamma \in L$, we have

$$\alpha \leq \beta \implies \alpha \vee (\gamma \wedge \beta) = (\alpha \vee \gamma) \wedge \beta.$$ 

Equivalently, a lattice $L$ is modular if it has no five element sublattice isomorphic to the lattice $N_5$ whose Hasse diagram is a pentagon: consider the sublattice generated by $\alpha' = \alpha \vee (\gamma \wedge \beta), \beta' = (\alpha \vee \gamma) \wedge \beta$, and $\gamma$, with top element $\alpha \vee \gamma = \alpha' \vee \gamma$ and bottom element $\gamma \wedge \beta = \gamma \wedge \beta'$, and note that we always have $\alpha' \leq \beta'$.

49
Proposition 8.8. If $\mathbb{A}$ has permuting congruences, then $\text{Con}(\mathbb{A})$ is a modular lattice.

Proof. We just have to check that if $\alpha \leq \beta$, then $\alpha \circ (\gamma \land \beta) \geq (\alpha \circ \gamma) \land \beta$. Suppose that $(x,z) \in (\alpha \circ \gamma) \land \beta$, and choose $y$ such that $(x,y) \in \alpha, (y,z) \in \gamma$. Then $(y,x) \in \alpha \subseteq \beta$ and $(x,z) \in \beta \circ \beta = \beta$, so $(y,z) \in \gamma \land \beta$, so $(x,z) \in \alpha \circ (\gamma \land \beta)$. □

An unexpectedly large example of a Mal’cev variety is the variety of quasigroups.

Definition 8.9. A binary operation on a finite set is called a quasigroup if its multiplication table is a Latin square (i.e. each element appears exactly once in each row and column). The variety of quasigroups has three basic operations ·, /, \, which satisfy the following identities:

\[(a \cdot b)/b \approx a, \quad (a/b) \cdot b \approx a, \quad b\backslash(b \cdot a) \approx a, \quad b \cdot (b \backslash a) \approx a.\]

Note that in the finite case, if · is a quasigroup operation, then the operations /, \ can be defined in terms of · by an iteration argument (for any invertible unary function $f$ on an $n$ element set, $f^{-1} = f^{\circ(n!-1)}$). For infinite quasigroups, they have to be introduced into the language explicitly.

Proposition 8.10. If $\mathbb{A} = (A, \cdot, /, \backslash)$ is a quasigroup, then $p : (x,y,z) \mapsto (x/y) \cdot ((y/y)\backslash z)$ is a Mal’cev term.

Proof. Plugging in $x = y$ we get

\[p(y,y,z) = (y/y) \cdot ((y/y)\backslash z) \approx z,\]

and plugging in $z = y$ we get

\[p(x,y,y) = (x/y) \cdot ((y/y)\backslash y) \approx (x/y) \cdot ((y/y)\backslash((y/y)\backslash y)) \approx (x/y) \cdot y \approx x.\] □

The corresponding property on the CSP side of the picture is something known as the parallelogram property (some authors call this rectangularity, although the definition of rectangularity is often slightly weaker in the case of relations of higher arity).

Definition 8.11. A binary relation $R \subseteq A \times B$ has the parallelogram property if whenever $(a,b),(c,b),(c,d) \in R$, we also have $(a,d) \in R$. A relation of higher arity is said to have the parallelogram property if every way of grouping its coordinates into two groups gives a binary relation with the parallelogram property.

Theorem 8.12. An algebraic structure $\mathbb{A}$ has a Mal’cev term $p$ iff every relation $R \in \text{Inv}(\mathbb{A})$ has the parallelogram property.
Proof. Suppose first that \( A \) has a Mal’cev term \( p \), let \( B, C \in \mathcal{V}(A) \) and let \( R \leq B \times C \) be a subalgebra of their product. Suppose that \((a, b), (c, b), (c, d) \in R\). Then
\[
\begin{bmatrix}
a \\
b \\
c \\
d \\
\end{bmatrix} = p \left( \begin{bmatrix}
a \\
b \\
c \\
d \\
\end{bmatrix}, \begin{bmatrix}
a \\
b \\
c \\
d \\
\end{bmatrix}, \begin{bmatrix}
a \\
b \\
c \\
d \\
\end{bmatrix} \right) \in R,
\]
so \( R \) has the parallelogram property.

Conversely, suppose that every relation in \( \text{Inv}(A) \) has the parallelogram property. Let \( \pi_1, \pi_2 \in A^{k^2} \) be the elements corresponding to the functions \( \pi_i : (a_1, a_2) \to a_i \). Let \( R \leq (A^{k^2})^2 \) be the subalgebra generated by the three pairs \((\pi_1, \pi_1), (\pi_2, \pi_1), (\pi_2, \pi_2)\). Then since \( R \) has the parallelogram property, we must have \((\pi_1, \pi_2) \in R\), so there must be a ternary term \( p \) such that
\[
\begin{bmatrix}
\pi_1 \\
\pi_2 \\
\pi_2 \\
\end{bmatrix} = p \left( \begin{bmatrix}
\pi_1 \\
\pi_1 \\
\pi_2 \\
\end{bmatrix}, \begin{bmatrix}
\pi_2 \\
\pi_1 \\
\pi_2 \\
\end{bmatrix}, \begin{bmatrix}
\pi_2 \\
\pi_2 \\
\pi_2 \\
\end{bmatrix} \right).
\]
But then this \( p \) is a Mal’cev term for \( A \).

**Definition 8.13.** If \( R \leq_{sd} A \times B \) is a subdirect binary relation, then the linking congruence of \( R \) can refer to any of the following three congruences: the congruence \( \ker \pi_1 \lor \ker \pi_2 \) on \( R \), the congruence \( \alpha \) on \( A \) generated by pairs \((a, a') \in A\) such that there exists \((b, b') \in B\) with \((a, b), (a', b') \in R\), or the similar congruence \( \beta \) defined on \( B \). The relation \( R \) is called linked if these congruences are full.

Note that in the above definition, we have \( \alpha = \pi_1(\ker \pi_1 \lor \ker \pi_2), \beta = \pi_2(\ker \pi_1 \lor \ker \pi_2) \), and \( A/\alpha \cong R/(\ker \pi_1 \lor \ker \pi_2) \cong B/\beta \).

A more visual way to understand the linking congruence is to think of the relation \( R \) as a bipartite graph on \( A \sqcup B \), and to define the congruence classes to be the connected components of this graph. In particular, \( R \) is linked iff this bipartite graph is connected.

**Proposition 8.14 (Goursat’s Lemma).** A subdirect binary relation \( R \leq_{sd} A \times B \) has the parallelogram property iff there are congruences \( \alpha, \beta \) on \( A, B \) respectively and an isomorphism \( f : A/\alpha \to B/\beta \) such that, writing \( \pi_\alpha, \pi_\beta \) for the quotient maps, we have \( R = \pi_\alpha^{-1} \circ f^{-1} \circ \pi_\beta \) (treating \( \pi_\alpha, \pi_\beta, f \) as binary relations with inputs on the right and outputs on the left).

**Proof.** Thinking of \( R \) as a bipartite graph on \( A \sqcup B \), we just have to prove that every connected component of \( R \) is a complete bipartite graph. Suppose \( a \in A \) and \( b \in B \) are in the same connected component of \( R \), and let \( a = a_1, b_1, \ldots, a_k, b_k = b \) be a path from \( a \) to \( b \) with \((a_i, b_i) \in R\) and \((a_{i+1}, b_i) \in R\) for each \( i \). We will show that \((a_i, b_i) \in R\) by induction on \( i \):
\[
\begin{bmatrix}
a_1 \\
b_i \\
\end{bmatrix}, \begin{bmatrix}
a_{i+1} \\
b_i \\
\end{bmatrix}, \begin{bmatrix}
a_i \\
b_{i+1} \\
\end{bmatrix} \in R \implies \begin{bmatrix}
a_1 \\
b_{i+1} \\
\end{bmatrix} \in R.
\]

Despite the trivial nature of binary relations with the parallelogram property, higher arity relations can encode more complicated global information.

**Example 8.1.** Consider the affine algebra \( A = (\mathbb{Z}/p, x - y + z) \), and let \( R \leq_{sd} A^n \) be the relation \( x_1 + \cdots + x_n \equiv 0 \) (mod \( p \)). Then if we think of \( R \) as a (subdirect) binary relation on \( A \times A^{n-1} \), it is the graph of the homomorphism \( A^{n-1} \to A \) given by \( (x_2, \ldots, x_n) \mapsto -x_2 - \cdots - x_n \) (mod \( p \)).

More generally, for any \( i \), if we think of \( R \) as a subdirect binary relation on \( A^i \times A^{n-i} \), then the linking congruence gives homomorphisms \( A^i \to A \leftarrow A^{n-i} : (x_1, \ldots, x_i) \mapsto x_1 + \cdots + x_i \) (mod \( p \)) and \( (x_{n-i+1}, \ldots, x_n) \mapsto -x_{n-i+1} - \cdots - x_n \) (mod \( p \)).
Ternary relations on simple Mal’cev algebras have a particularly interesting structure.

**Proposition 8.15.** Let \( \mathbb{A}_1, \mathbb{A}_2, \mathbb{A}_3 \) be simple idempotent Mal’cev algebras, and suppose that \( \mathbb{R} \leq_{sd} \mathbb{A}_1 \times \mathbb{A}_2 \times \mathbb{A}_3 \) has \( \pi_{i,j}(\mathbb{R}) = \mathbb{A}_i \times \mathbb{A}_j \) for each \( i \neq j \) but that \( \mathbb{R} \neq \mathbb{A}_1 \times \mathbb{A}_2 \times \mathbb{A}_3 \). Then for each \( a \in \mathbb{A}_1 \), the relation 
\[ \mathbb{R}_a = \pi_{2,3}(\mathbb{R} \cap \{a\} \times \mathbb{A}_2 \times \mathbb{A}_3) \]
is the graph of an isomorphism between \( \mathbb{A}_2 \) and \( \mathbb{A}_3 \), and for every \( b \in \mathbb{A}_2, c \in \mathbb{A}_3 \) there is a unique \( a \in \mathbb{A}_1 \) such that \( (b, c) \in \mathbb{R}_a \).

**Proof.** Consider \( \mathbb{R} \) as a subdirect relation on \( \mathbb{A}_1 \times (\mathbb{A}_2 \times \mathbb{A}_3) \). Since the linking congruence on \( \mathbb{A}_1 \) is not full (else \( \mathbb{R} \) would be the full relation by the parallelogram property), it must be trivial (since \( \mathbb{A}_1 \) is simple), so \( \mathbb{R} \) is the graph of a homomorphism from \( \mathbb{A}_2 \times \mathbb{A}_3 \) to \( \mathbb{A}_1 \), which proves the last assertion.

Similarly, \( \mathbb{R} \) may be viewed as the graph of a homomorphism from \( \mathbb{A}_1 \times \mathbb{A}_2 \) to \( \mathbb{A}_3 \), so \( \mathbb{R}_a \) is the graph of a surjective homomorphism from \( \mathbb{A}_2 \) to \( \mathbb{A}_3 \) for any \( a \in \mathbb{A}_1 \) (surjective because \( \pi_{1,3}(\mathbb{R}) = \mathbb{A}_1 \times \mathbb{A}_3 \)), and by simplicity of \( \mathbb{A}_2 \) this homomorphism must be an isomorphism.

If we fix an isomorphism \( \mathbb{A}_1 \cong \mathbb{A}_2 \cong \mathbb{A}_3 \cong \mathbb{A} \) coming from the above proposition, then the kernel of the associated homomorphism \( \mathbb{A} \times \mathbb{A} \cong \mathbb{A}_2 \times \mathbb{A}_3 \to \mathbb{A}_1 \) contains the diagonal of \( \mathbb{A} \times \mathbb{A} \) as a congruence class. In this case - that is, the case where \( \mathbb{A} \times \mathbb{A} \) has the diagonal as a congruence class of some congruence - \( \mathbb{A} \) is called an **abelian** algebra.

**Example 8.2.** Let \( \mathbb{A}_n = (\{0, ..., n-1\}, p) \), where \( p \) is the ternary Mal’cev operation defined by
\[ p(x, y, z) = \begin{cases} 
  z & \text{if } x = y, \\
  y & \text{if } x = z, \\
  x & \text{if } x \not\in \{y, z\}.
\end{cases} \]

For \( n \geq 3 \), \( \mathbb{A}_n \) is simple and non-abelian (i.e. the diagonal is not a congruence class of any congruence on \( \mathbb{A}_n^2 \)). \( \text{Inv}(\mathbb{A}_n) \) is generated by a pair of graphs of permutations of \( \{0, ..., n-1\} \) which generate the full symmetric group, the unary relation \( x \neq 0 \), and the ternary relation
\[ (x, y, z \in \{0, 1\}) \land (x + y + z \equiv 0 \pmod{2}) \).

It is a good exercise to prove that the above relations generate \( \text{Inv}(\mathbb{A}_n) \).

**Example 8.3.** Here we describe an example of a three element Mal’cev algebra which is “solvable”, but which is not abelian. Let \( \mathbb{A} = (\{0, 1, *\}, p) \), where \( p \) is the ternary Mal’cev operation defined by
\[ p(x, y, z) = \begin{cases} 
  x & \text{if } y = z, \\
  y & \text{if } x = z, \\
  z & \text{if } x = y, \\
  * & \text{if } \{x, y, z\} = \{0, 1, *\}.
\end{cases} \]

Every two element subset of \( \mathbb{A} \) is a subalgebra isomorphic to the idempotent reduct of \( \mathbb{Z}/2 \), and \( \mathbb{A} \) has a congruence \( \theta \) corresponding to the partition \( \{0, 1\}, \{*\} \) such that \( \mathbb{A}/\theta \) is also isomorphic to the idempotent reduct of \( \mathbb{Z}/2 \).
Along with the obvious relations on \( A \), there is also the ternary relation

\[
(x = y = z = *) \lor (x, y, z \in \{0, 1\} \land x + y + z \equiv 0 \pmod{2}),
\]

whose elements correspond to the columns of the matrix

\[
\begin{bmatrix}
* & 0 & 0 & 1 & 1 \\
* & 0 & 1 & 0 & 1 \\
* & 1 & 1 & 1 & 0
\end{bmatrix}.
\]

That this relation forms a subalgebra of \( A^3 \) is related to the fact that \( \theta \) can be considered to be an “abelian congruence” (in a sense we will define later).

9 Mal’cev algorithm and compact representations

The algorithm for solving CSPs invariant under a Mal’cev operation, due to Bulatov and Dalmau [31], is based on the fact that any Mal’cev constraint has a small generating set. More specifically, we will show that any subset of a relation \( R \) which has the same projection to each factor and contains representatives of all of the “forks” of \( R \) actually generates \( R \).

**Definition 9.1.** If \( R \subseteq A_1 \times \cdots \times A_n \), then we define the *signature* of \( R \), written \( \text{Sig}(R) \), to be the set of triples \((i, a, b)\) with \( i \in \{1, \ldots, n\} \), \( a, b \in A_i \) such that there are some \( t_a, t_b \in R \) with \( \pi_{1, \ldots, i-1}(t_a) = \pi_{1, \ldots, i-1}(t_b) \) and \( \pi_i(t_a) = a, \pi_i(t_b) = b \). In this case we say that the pair \( t_a, t_b \) witnesses the triple \((i, a, b)\).

**Theorem 9.2.** Suppose that a relation \( R \subseteq A_1 \times \cdots \times A_n \) is preserved by a Mal’cev term \( p \), and that \( S \subseteq R \) is a subset with \( \text{Sig}(S) = \text{Sig}(R) \). Then \( R \) is generated by \( S \) (using only \( p \)).

**Proof.** Let \( S \) be the subset of \( R \) generated by \( S \) using \( p \). We will prove by induction on \( i \) that \( \pi_{1, \ldots, i}(S) = \pi_{1, \ldots, i}(R) \).

Suppose that \( t \in R \). By the induction hypothesis, there is some \( t' \in S \) with \( \pi_{1, \ldots, i-1}(t) = \pi_{1, \ldots, i-1}(t') \). Let \( a = \pi_i(t'), b = \pi_i(t) \). Since \( S \subseteq R \), we have \((i, a, b) \in \text{Sig}(R) = \text{Sig}(S)\), so there must be a pair \( t_a, t_b \in S \) witnessing the triple \((i, a, b)\). Define \( t'' \in S \) by

\[
t'' = p(t', t_a, t_b).
\]

Then from \( \pi_{1, \ldots, i-1}(t_a) = \pi_{1, \ldots, i-1}(t_b) \) and the fact that \( p \) is Mal’cev, we have

\[
\pi_{1, \ldots, i-1}(t'') = \pi_{1, \ldots, i-1}(p(t', t_a, t_b)) = \pi_{1, \ldots, i-1}(t') = \pi_{1, \ldots, i-1}(t).
\]

Additionally, from \( \pi_i(t') = \pi_i(t_a) = a \) and the fact that \( p \) is Mal’cev, we have

\[
\pi_i(t'') = p(a, a, b) = b = \pi_i(t),
\]

so \( \pi_{1, \ldots, i}(t'') = \pi_{1, \ldots, i}(t) \). \( \Box \)

**Definition 9.3.** A subset \( S \subseteq R \) is called a *compact representation* of a Mal’cev relation \( R \) if \( \text{Sig}(S) = \text{Sig}(R) \) and \(|S| \leq 2|\text{Sig}(R)|\).
Proposition 9.4. Every Mal'cev relation $\mathbb{R} \subseteq A_1 \times \cdots \times A_n$ has a compact representation $S$. We always have $|S| \leq 2n \cdot \max_i |A_i|^2$.

Now we need some subroutines for manipulating compact representations. The first such procedure is called \texttt{Nonempty}: it takes as input a compact representation $R$ of a relation $\mathbb{R} \subseteq A_1 \times \cdots \times A_n$ and any description of a relation $S \subseteq A_{i_1} \times \cdots \times A_{i_k}$ on a small subset $\{i_1, \ldots, i_k\}$ of the indices, and it tells us whether $\mathbb{R} \cap S \neq \emptyset$. In the case $\mathbb{R} \cap S \neq \emptyset$, \texttt{Nonempty} returns an element of the intersection.

**Algorithm 4** \texttt{Nonempty}(R, i_1, \ldots, i_k, S), $p$ a Mal'cev term, $R$ a compact representation of $\mathbb{R} \subseteq A_1 \times \cdots \times A_n$, $S \subseteq A_{i_1} \times \cdots \times A_{i_k}$.

1: Set $R' \leftarrow R$.
2: \textbf{while} $\pi_{i_1,\ldots,\!i_k}(R')$ is not closed under $p$ and $R' \cap S = \emptyset$ \textbf{do}
3: \quad Pick $t_1, t_2, t_3 \in R'$ with $\pi_{i_1,\ldots,\!i_k}(p(t_1, t_2, t_3)) \notin \pi_{i_1,\ldots,\!i_k}(R')$.
4: \quad Set $R' \leftarrow R' \cup \{p(t_1, t_2, t_3)\}$.
5: \textbf{if} $R' \cap S \neq \emptyset$ \textbf{then}
6: \quad \textbf{return} any element of $R' \cap S$.
7: \quad \textbf{else}
8: \quad \textbf{return} $\emptyset$.

Proposition 9.5. \texttt{Nonempty} correctly determines whether $\mathbb{R} \cap S \neq \emptyset$ in time polynomial in $n$, $|R|$, and $|\pi_{i_1,\ldots,\!i_k}(\mathbb{R})| \leq \prod_{j \leq k} |A_{i_j}|$.

Proof. Since $\mathbb{R}$ is generated by $R$ using $p$, we also have $\pi_{i_1,\ldots,\!i_k}(\mathbb{R})$ generated by $\pi_{i_1,\ldots,\!i_k}(R)$ using $p$. To see the bound on the running time, note that in each iteration of the while loop, the set $\pi_{i_1,\ldots,\!i_k}(R')$ gains a new element, and its size is clearly bounded by $|\pi_{i_1,\ldots,\!i_k}(R)|$. \hfill $\square$

The next subroutine for manipulating compact representations is \texttt{Fix-values}. \texttt{Fix-values} converts a compact representation $R$ of $\mathbb{R} \subseteq A_1 \times \cdots \times A_n$ to a compact representation of $\mathbb{R} \cap (x_1 = a_1) \wedge \cdots \wedge (x_m = a_m)$, for any choice of $m \leq n$ and $a_i \in A_i$ for all $i$. \texttt{Fix-values} is really the core of the algorithm, the other steps are mostly formal (in fact, \texttt{Nonempty} and \texttt{Fix-values} are the only two subroutines which use the Mal'cev term $p$).

Proposition 9.6. \texttt{Fix-values} correctly returns a compact representation of $\mathbb{R}_m = \mathbb{R} \cap (x_1 = a_1) \wedge \cdots \wedge (x_m = a_m)$ in polynomial time.

Proof. We prove by induction on $j$ that $R_j$ is a compact representation of $\mathbb{R}_j$ for each $j \leq m$. Note that for any $(i, a, b) \in \text{Sig}(R_j)$, if $a \neq b$ then we must have $i > j$. For $i \leq j$, we have $(i, a_i, a_i) \in R_j$ iff $R_j \neq \emptyset$ by how we initialize $R_j$.

If $i > j$, then $(i, a, b) \in \text{Sig}(R_j)$ implies $(i, a, b) \in \text{Sig}(R_{j-1})$, witnessed by some pair $t_a, t_b \in R_{j-1}$. Additionally, if $(i, a, b) \in \text{Sig}(R_j)$, then there is certainly some $t \in R_j$ with $\pi_i(t) = a$, so the call to \texttt{Nonempty} inside the while loop will succeed. Then

$$
\pi_{1,\ldots,i-1}(p(t, t_a, t_b)) = \pi_{1,\ldots,i-1}(p(t, t_a, t_a)) = \pi_{1,\ldots,i-1}(t),
$$

54
Algorithm 5 \textbf{Fix-values}(R,a_1,...,a_m), \( p \) a Mal’cev term, \( R \) a compact representation of \( R \leq A_1 \times \cdots \times A_n \).

1: Set \( R_0 \leftarrow R \).
2: for \( j \) from 1 to \( m \) do
3: \hspace{1em} if \( (j,a_j,a_j) \notin \operatorname{Sig}(R_{j-1}) \) then
4: \hspace{2em} return \emptyset.
5: \hspace{1em} else
6: \hspace{2em} Set \( R_j \leftarrow \{t\} \), where \( t \in R_{j-1} \) and the pair \( t,t \) witnesses the triple \( (j,a_j,a_j) \).
7: \hspace{1em} for all \( (i,a,b) \in \operatorname{Sig}(R_{j-1}) \) with \( i > j \) do
8: \hspace{2em} Let \( t_a,t_b \in R_{j-1} \) witness the triple \( (i,a,b) \).
9: \hspace{2em} Let \( t \leftarrow \operatorname{Nonempty}(R_{j-1},j,i,\{(a_j,a)\}) \).
10: \hspace{2em} if \( t \neq \emptyset \) then
11: \hspace{3em} Set \( R_j \leftarrow R_j \cup \{t,p(t,t_a,t_b)\} \).
12: return \( R_m \).

so from \( i > j \) we have \( p(t,t_a,t_b) \in R_j \). From \( \pi_i(t) = a, \pi_i(p(t,t_a,t_b)) = p(a,a,b) = b \), we see that the pair \( t,p(t,t_a,t_b) \) witnesses the triple \( (i,a,b) \).

To see that \textbf{Fix-values} runs in polynomial time, note that every call to \textbf{Nonempty} involves a constraint on two variables. \hfill \(\square\)

Corollary 9.7. Given a compact representation \( R \) of a relation \( R \subseteq A_1 \times \cdots \times A_n \) which is preserved by a given Mal’cev operation \( p \), and given a tuple \( t \in A_1 \times \cdots \times A_n \), we can determine whether \( t \in R \) in polynomial time.

The next subroutine will give a compact representation for the intersection of a relation \( R \) given by a compact representation \( R \) and a relation \( S \) of small arity. In \cite{31} this subroutine was called \textbf{Next-beta}, so we will copy that notation here.

Algorithm 6 \textbf{Next-beta}(R,i_1,...,i_k,S), \( R \) a compact representation of \( R \leq A_1 \times \cdots \times A_n \), \( S \leq A_{i_1} \times \cdots \times A_{i_k} \).

1: Set \( R' \leftarrow \emptyset \).
2: for all \( (i,a,b) \in \operatorname{Sig}(R) \) do
3: \hspace{1em} Set \( t_a \leftarrow \operatorname{Nonempty}(R,i_1,...,i_k,i \times \{a\}) \).
4: \hspace{1em} if \( t_a \neq \emptyset \) then
5: \hspace{2em} Set \( t_b \leftarrow \operatorname{Nonempty}(\operatorname{Fix-values}(R,\pi_1(t_a),...,$$ \pi_{i-1}(t_a)),i_1,...,i_k,i,S \times \{b\}) \).
6: \hspace{2em} if \( t_b \neq \emptyset \) then
7: \hspace{3em} Set \( R' \leftarrow R' \cup \{t_a,t_b\} \).
8: return \( R' \).

Proposition 9.8. \textbf{Next-beta} correctly finds a compact representation of \( R \cap S \) in time polynomial in \( n \), \( |R| \), and \( |\pi_{i_1},...,i_k(R)| \cdot \max_i |A_i| \leq \prod_{j \leq k} |A_{i_j}| \cdot \max_i |A_i| \).

Bulatov and Dalmau \cite{31} then go on to define a subroutine \textbf{Next} which calls \textbf{Next-beta} on larger and larger projections of \( S \), ensuring that \( |\pi_{i_1},...,i_k(R)| \leq |S| \cdot \max_i |A_i| \) every time that \textbf{Next-beta} is called. A better approach, leading to a more powerful algorithm, was found by Maróti \cite{38}.

55
The subroutine \textbf{Intersect} takes two compact representations $R, S$ of relations $\mathbb{R}, \mathbb{S}$ as input and returns a compact representation of $\mathbb{R} \cap \mathbb{S}$ as output.

\begin{algorithm}
\caption{\textbf{Intersect}($R, i_1, \ldots, i_k, S$), $R$ a compact representation of $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, $S$ a compact representation of $\mathbb{S} \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}$}
\begin{enumerate}
\item Let $t_R \in R$ and $t_S \in S$ be any tuples.
\item Set $R' \leftarrow (R \times \{t_S\}) \cup (\{t_R\} \times S) \subseteq \mathbb{A}_1 \times \cdots \mathbb{A}_n \times \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}$.
\item for $j \leq k$ do
\item \hspace{1em} Set $R' \leftarrow \text{Next-beta}(R', i_j, n+j, \mathbb{A}_{i_j})$.
\item return a minimal subset of $\pi_{1,\ldots,n}(R')$ which witnesses every triple $(i,a,b) \in \text{Sig}(\pi_{1,\ldots,n}(R'))$.
\end{enumerate}
\end{algorithm}

\textbf{Theorem 9.9.} Any CSP which is preserved by a Mal’cev operation, where the relations are given by their compact representations, can be solved in time polynomial in the number of variables, the number of relations, and the size of the largest domain. In fact, we can find a compact representation of the solution set in polynomial time.

\textit{Proof.} We start with any compact representation of $\mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, and simply apply the subroutine \textbf{Intersect} repeatedly to find a compact representation of the intersection of all the constraint relations. To see that \textbf{Intersect} works correctly and efficiently, note that $R'$ is initialized as a compact representation of $R \times S$ and ends as a compact representation of $\mathbb{R} \cap \mathbb{S}$ followed by $k$ repeated coordinates. To see that \textbf{Intersect} runs in polynomial time, note that each call of \textbf{Next-beta} involves a relation of arity 2.

\textbf{Corollary 9.10.} For any primitive positive formula $\varphi$ in a collection of relations $\mathbb{R}_i$, if we are given compact representations of each $\mathbb{R}_i$ then we can efficiently find a compact representation of the relation described by $\varphi$.

\textit{Proof.} If we are given a compact representation of a relation and we permute its variables, we can efficiently find a compact representation for the permuted relation by using the \textbf{Intersect} subroutine with $\mathbb{R}$ equal to a full relation. To handle projections, note that we can project onto any initial segment of the variables by just projecting our compact representation and pruning it.

While this might appear to be a fully satisfactory theory, there is still one big question remaining: what happens if instead of having relations described by compact representations, we have relations which are instead described by an arbitrary set of generators? It’s clear that we just need to find a way to compute a compact representation of $\text{Sg}_{\mathbb{A}^n}(S)$ for any small set $S \subseteq \mathbb{A}^n$, and a little thought shows that this can be reduced to the following problem.

\textbf{Problem 9.1.} Let $\mathbb{A}$ be a fixed Mal’cev algebra. Given a subset $S \subseteq \mathbb{A}^n$, and given a tuple $t \in \mathbb{A}^n$, can we determine whether $t \in \text{Sg}_{\mathbb{A}^n}(S)$ in time polynomial in $|S|$ and $n$?

This is a special case of the Subpower Membership Problem $[14.1]$ Even this special case is open (the answer is conjectured to be yes). In the case of groups, the famous Schreier-Sims algorithm gives a positive solution (see [55] for a straightforward exposition).
Remark 9.1. The proof of correctness of the subroutine Nonempty and the algorithm for Fix-values are both directly connected to the proof of Theorem 9.2. The subroutines Next-beta and Intersect use the subroutines Nonempty and Fix-values as black boxes and don’t involve the algebraic structure at all. Thus, in order to generalize the Mal’cev algorithm to more general algebraic structures, the only new ingredient needed is a proof of a generalization of Theorem 9.2.

9.1 Near-subgroups

In this subsection, we will describe the maximal polynomial-time solvable extension $G^*$ of the relational clone $G$ of cosets of subgroups of $\mathbb{G}^n$, where $\mathbb{G}$ is a finite group. The relational clone $G^*$ will turn out to have a Mal’cev polymorphism, so the algorithm for Mal’cev algebras can be used to prove the dichotomy for extensions of $G$.

First, consider the simple case where $G = \mathbb{Z}/n$ is cyclic of order $n$ at least 3. It’s easy to see that if we add the unary relation $\{0,1\}$ to $\mathbb{Z}/n$, then we can simulate 1-IN-3 SAT via the primitive positive formula

$$x + y + z = 1 \land x, y, z \in \{0, 1\}.$$ 

Using an inductive argument with this as the base case, Feder and Vardi \cite{FederVardi} show that if we adjoin any unary relation to $\mathbb{Z}/n$ which isn’t a coset of a subgroup, then we can simulate 1-IN-3 SAT as well.

**Proposition 9.11** (Feder, Vardi \cite{FederVardi}). If we adjoin any unary relation $K$ to the relational structure $G = (\mathbb{Z}/n, \{1\}, x, y = z)$, then the resulting CSP is NP-complete unless $K$ is a coset of a subgroup of $\mathbb{Z}/n$.

**Proof.** Using the binary relation $y = x + i$ for constants $i \in \mathbb{Z}/n$, we see that $K - i$ is in the relational clone generated by $K$ and $G$. Thus we may assume without loss of generality that $0 \in K$, and by possibly restricting to a subgroup we may assume that $\langle K \rangle = \mathbb{Z}/n$. By applying an automorphism of $\mathbb{Z}/n$, we may also assume that $1 \in K$.

We induct on $|K|, n$. If there is an $i \neq 0$ with $i, i + 1 \in K$, then $K \cap (K - i)$ also contains $0, 1$, and will be strictly smaller than $K$ unless $K = K - i$, in which case we may take the quotient by $\langle i \rangle$. Thus we may assume that $i, i + 1$ are not both in $K$ for any $i \neq 0$.

If $K$ contains some $i$ with neither of $i, i - 1$ relatively prime to $n$, then by induction $K \cap \langle i \rangle$ and $(K - 1) \cap (i - 1)$ are subgroups, so $2i, 2i - 1 \in K$ and we so must have $2i - 1 \equiv 0 \pmod{n}$, contradicting the assumption that $i$ has a common factor with $n$.

If $K$ contains $i \neq 1$ with $i$ relatively prime to $n$, then $K \cap (i - K)$ contains $0, i$ but not $1$, and we may apply the induction hypothesis to get a contradiction. Similarly, if $K$ contains $i \neq 0$ with $i - 1$ relatively prime to $n$, then $(K - 1) \cap (i - K)$ contains $0$ and $i - 1$ but not $-1$, and we may apply the induction hypothesis.

Thus the only case to consider is the case $K = \{0, 1\}$, and we have already seen that in this case we can simulate 1-IN-3 SAT (unless $n = 2$, in which case $K = \mathbb{Z}/n$).

Next, consider the case where $G$ is the Klein four-group $(\mathbb{Z}/2)^2$. The only unary relations which aren’t already cosets of subgroups of $(\mathbb{Z}/2)^2$ are the relations with three elements. If we adjoin any three element unary relation to $(\mathbb{Z}/2)^2$, then we can again simulate 1-IN-3 SAT: if we adjoin the relation $K = \{(0,0), (0,1), (1,0)\}$, for instance, then we can use the primitive positive formula

$$\exists t \ (x, y, z \in \{(0,0), (0,1)\} \land x + y + z = (0,1) \land (x, t) \in \{(0,0), (0,0), (0,1), (1,0)\} \land y + t \in K),$$

57
which is satisfied iff exactly one of \(x, y, z\) is \((0, 1)\) and the other two are \((0, 0)\).

Now consider the case where \(G\) is any finite abelian group, and \(K \subseteq G\) is a unary relation which can be added without creating NP-completeness. Then if any \(a, a + b \in K\), we must have \(a + ib \in K\) for all \(i \in \mathbb{Z}\) by the cyclic case. By the Klein four-group case, if we have subgroups \(N \leq M \leq G\) with \(M/N \cong (\mathbb{Z}/2)^2\), then if \(K\) meets any three elements of \(M/N\) it must also meet the fourth.

**Proposition 9.12.** Suppose that \(G\) is an abelian group and that \(K \subseteq G\) has \(0 \in K\), and the property that if \(a, a + b \in K\) then \(a + \langle b \rangle \subseteq K\), and the property that for any subgroups \(N \leq M \leq G\) with \(M/N \cong (\mathbb{Z}/2)^2\), we have \(|(K \cap M)/N| \neq 3\). Then \(K\) must be a subgroup of \(G\).

**Proof.** From the first assumption, for any \(a, b \in K\) we have \(-ia, jb \in K\), so

\[
ia + 2jb = jb - (-ia - jb) \in jb + \langle -ia - jb \rangle \subseteq K,
\]

and similarly \(2ja + ib \in K\) for all \(i, j \in \mathbb{Z}\).

Thus, if we take \(M = \langle a, b \rangle\) and \(N = \langle 2a, 2b \rangle\), we see that either \(|M/N| < 4\) in which case \(a + b \in \langle a, b \rangle \subseteq K\), or \(M/N \cong (\mathbb{Z}/2)^2\) and \(|(K \cap M)/N| \geq 3\). In the latter case, the second assumption implies that there are \(i, j\) such that \((2i+1)a + (2j+1)b \in K\). Then \(a + b = (2i+1)a + (2j+1)b - 2ia - 2jb \in K\) by repeated application of the first assumption. Either way, \(a + b \in K\), so \(K\) is closed under addition.

**Corollary 9.13.** If \(G\) is a finite abelian group and \(G\) the associated relational structure, then for any \(m\)-ary relation \(K\) which is not a coset of a subgroup of \(G^m\), the CSP we get by adding \(K\) to \(G\) is NP-complete.

**Proof.** Apply the previous proposition to the abelian group \(G^m\).

In the case of nonabelian groups, however, we may be able to adjoin interesting new constraints. Note that if we adjoin any constraint, then we automatically adjoin all of its cosets, since for any constant \(b \in G\) the relation \(y = bx\) is a left coset of the diagonal subgroup of \(G^2\). So by the abelian case, the only possibilities for new relations are those described by the following definition.

**Definition 9.14.** A subset \(K \subseteq G\) is a near subgroup of \(G\) if it contains 1, and for any \(b \in K^{-1}\), any \(M \leq G\) and any \(N < M\) with \(M/N\) abelian, the quotient set \((bK \cap M)/N\) is a subgroup of \(M/N\).

**Theorem 9.15** (Aschbacher [4]). The intersection of two near subgroups of a finite group is a near subgroup.

**Corollary 9.16** (Feder [52]). Let \(G\) be a finite group, and let \(G^*\) be the relational structure on the underlying set of \(G\) having as relations all cosets of all near subgroups of \(G^n\). Then \(G^*\) has a Mal’cev polymorphism.

**Proof.** Consider the “free near subgroup generated by two elements”, that is, the smallest near subgroup \(K\) of \(G^{2}\) which contains \(\pi_1, \pi_2\) (a smallest such near subgroup exists since the intersection of all of them is guaranteed to be a near subgroup as well). Let \(N\) be the commutator subgroup of the group generated by \(\pi_1, \pi_2\). Since \(\langle \pi_1, \pi_2 \rangle/N\) is abelian, there must be some \(c \in N\) with \(\pi_1\pi_2c \in K\) by the definition of a near subgroup.

We define a binary polymorphism \(g\) by \(g = \pi_1\pi_2c\), that is, \(g(x, y) = xyc(x, y)\), where \(c \in G^{2}\) is interpreted as a function \(c : G^2 \to G\). Since for all \(x, y\) we know that \(c(x, y)\) is contained in the
commutator subgroup of \( \langle x, y \rangle \), we have \( c(x, 1) = c(1, x) = 1 \) for all \( x \), so \( g(x, 1) = g(1, x) = x \). Now we define a Mal’cev operation \( p \) by

\[
p(x, y, z) = yg(y^{-1}x, y^{-1}z) = xy^{-1}z(x^{-1}, y^{-1}z).
\]

That \( p \) is Mal’cev follows directly from the fact that \( g \) satisfies the identities \( g(1, x) \approx g(x, 1) \approx x \).

To see that \( p \) is really a polymorphism of \( G^* \), let \( X \) be any coset of any near subgroup of \( G^n \), and let \( x, y, z \in X \). Then \( y^{-1}X \) is a near subgroup of \( G^n \). Since \( g = \pi_1 \pi_2 c \in K \), \( g \) preserves every near subgroup of \( G^n \). Thus from \( y^{-1}x, y^{-1}z \in y^{-1}X \) we have \( g(y^{-1}x, y^{-1}z) \in y^{-1}X \), and \( p(x, y, z) = yg(y^{-1}x, y^{-1}z) \in X \), so \( p \) does indeed preserve \( X \).

In order to prove Aschbacher’s Theorem \[9.15\] we first need a more convenient characterization of near-subgroups.

**Definition 9.17.** A subset \( K \) of a finite group \( G \) is a **twisted subgroup** if \( 1 \in K \) and \( x, y \in K \implies xyx \in K \).

**Proposition 9.18.** If \( K \) is a twisted subgroup and \( x \in K \), then \( \langle x \rangle \subseteq K \), so in particular \( K = K^{-1} \). If \( b \in K \), then \( bK \) is also a twisted subgroup.

**Proof.** For the first statement, for any \( x \in K \) we have \( x^k \cdot x^k \cdot x^k \in K \) for all \( k \geq 0 \), so \( \langle x \rangle \subseteq K \). For the second statement, if \( x, y \in bK \) and \( b \in K \), then \( b^{-1}(x \cdot y \cdot x) = (b^{-1}x) \cdot (b^{-1}y \cdot b) \cdot (b^{-1}x) \in K \), so \( xyx \in bK \).

**Proposition 9.19.** A subset \( K \subseteq G \) is a near-subgroup iff it is a twisted subgroup such that for any \( b \in K^{-1} \), any \( M \leq G \) and any \( N < M \) with \( M/N \) isomorphic to the Klein four-group, \(|(bK \cap M)/N| \neq 3 \).

**Proof.** We just need to check that \( K \) being a twisted subgroup is equivalent to \( \langle x \rangle \subseteq bK \) for all \( x, b \) with \( x \in bK \), \( b \in K^{-1} \). The previous proposition proves one direction of the equivalence. For the other direction, if \( x, y \in K \), then \( xy \in yK \) and \( y^{-1} \in \langle y \rangle \subseteq K \), so \( (yx)^2 \in \langle yx \rangle \subseteq yK \), which is equivalent to \( xyx = y^{-1}(yx)^2 \in K \).

**Example 9.1.** An explicit example of a near subgroup which is not a subgroup is given in [53]: Let \( G \) be the group of order \( p^3 \) (\( p \) odd) with elements indexed by triples \( (i, j, k) \subseteq (\mathbb{Z}/p)^3 \) and multiplication given by

\[
(i, j, k)(i', j', k') := (i + i' + jk', j + j', k + k').
\]

Let \( K = \{(\frac{1}{2}jk, j, k) | j, k \in \mathbb{Z}/p\} \). Since \( G \) has odd order, to check that \( K \) is a near subgroup we just need to check that it is a twisted subgroup, i.e. that it contains \((0,0,0)\) and is closed under the binary operation \( x, y \mapsto xyx \). This can be checked by direct calculation: for any \( j, k, j', k' \) we have

\[
(\frac{1}{2}jk, j, k)(\frac{1}{2}j'k', j', k')(\frac{1}{2}jk, j, k) = (\frac{1}{2}(2jk + j'k') + jk + j'k, 2j + j', 2k + k') \in K.
\]

That \( K \) is not a subgroup follows from \((0, 0, 1)(0, 1, 0) = (0, 1, 1) \not\in K \).

That we needed to take \( p \) odd in the above example is no coincidence, as the next proposition shows.
Proposition 9.20. If \( G \) is a 2-group, then any near-subgroup of \( G \) is a subgroup of \( G \).

Proof. We prove this by induction on the order of \( G \). Let \( K \) be a near-subgroup of \( G \), and assume without loss of generality that \( \langle K \rangle = G \).

Let \( z \in Z(G) \) be a nontrivial involution in the center of \( G \), which must exist since every 2-group has a nontrivial center, and every nontrivial element of the center has a power which is a nontrivial involution. By induction we have \( K/\langle z \rangle = G/\langle z \rangle \).

If \( z \in K \) then for any \( g \in G \setminus K \), we have \( gz \in K \), and from \( \langle gz, z \rangle \) abelian and \( gz, z \in K \) we have \( \langle gz, z \rangle \subseteq K \), so in particular \( g = gz \cdot z \in K \). Thus if \( z \in K \) we have \( G = K \).

Thus we may assume that \( z \not\in K \), and in fact that \( Z(G) \cap K = 1 \). Let \( M \) be a maximal subgroup of \( G \) containing \( \langle z \rangle \), then by induction we have \( M \cap K \) a subgroup of \( M \). Since \( (M \cap K)/\langle z \rangle = M/\langle z \rangle \), we have \( M \cong (M \cap K) \times \langle z \rangle \).

Let \( \Phi(M) \) be the Frattini subgroup of \( M \), which for 2-groups is given by \( \Phi(M) = M^2[M,M] \). Then from \( \Phi(\langle z \rangle) = 1 \) we have \( \Phi(M) = \Phi(M \cap K) \), and since \( M \triangleleft G \) we have \( \Phi(M) \triangleleft G \). Thus if \( \Phi(M) \neq 1 \) then by considering parities of the sizes of the orbits of elements of \( \Phi(M) \) under conjugation we see that \( \Phi(M \cap K) = \Phi(M) \) contains a nontrivial element of \( Z(G) \), contradicting \( K \cap Z(G) = 1 \). Thus \( \Phi(M) = 1 \), so \( M \) has exponent 2. Since this holds for every maximal subgroup of \( G \) which contains \( \langle z \rangle \), we see that \( G \) has exponent 2, so \( G \) is abelian.

Next we show that we can reduce to the situation where \( \langle K \rangle \) has an automorphism of order two which sends \( k \) to \( k^{-1} \) for all \( k \in K \).

Definition 9.21. If \( K \) is a twisted subgroup, we define the \( K \)-radical \( \Xi_K \) to be the set of elements of the form \( k_1 \cdots k_n \) with \( k_i \in K \) such that \( k_1^{-1} \cdots k_n^{-1} = 1 \).

Proposition 9.22. If \( K \) is a twisted subgroup and \( \Xi_K \) is the \( K \)-radical, then \( \Xi_K \) is a normal subgroup of \( \langle K \rangle \), and for any \( x \in K \) we have \( x\Xi_K \subseteq K \).

Proof. To see that \( \Xi_K \) is normal in \( \langle K \rangle \), just note that for any \( b \in K \) we have

\[
k_1^{-1} \cdots k_n^{-1} = 1 \iff b^{-1} k_1^{-1} \cdots k_n^{-1} b = 1 \iff bk_1 \cdots k_n b^{-1} \in \Xi_K,
\]

so \( b\Xi_K b^{-1} \subseteq \Xi_K \).

For the second statement, note that \( k_1^{-1} \cdots k_n^{-1} = 1 \iff k_n \cdots k_1 = 1 \), so if \( x \in K \) then we have

\[
x(k_1 \cdots k_n) = (k_n \cdots k_1)x(k_1 \cdots k_n) = k_n(\cdots (k_1 x k_1) \cdots ) k_n \in K.
\]

Proposition 9.23. If \( K \) is a twisted subgroup with \( \Xi_K = 1 \), and if \( \tau \) satisfies \( \tau^2 = 1 \), \( \tau k \tau = k^{-1} \) for \( k \in K \), then \( \tau K \) is preserved under conjugation by elements of \( \langle K, \tau \rangle \).

Proof. If \( x, y \in K \), then \( x^{-1} \tau y x = \tau xy x \in \tau K \), and \( \tau^{-1} \tau y \tau = \tau y^{-1} \in \tau K \).

Proposition 9.24. A twisted subgroup \( K \subseteq G \) is a near-subgroup of \( G \) if and only if the intersection \( bK \cap S \) is a subgroup of \( S \) for every 2-Sylow subgroup \( S \) of \( G \) and every \( b \in K^{-1} \).

Proof. Suppose for contradiction that \( M \leq G \), \( N \triangleleft M \) with \( M/N \) isomorphic to the Klein four-group, and \( b \in K^{-1} \) with \( |(bK \cap M)/N| = 3 \). We may assume without loss of generality that \( M = \langle K \rangle \cap M \) and \( N = \langle K \rangle \cap N \), that \( G = \langle K \rangle \), that \( b = 1 \), and that \( \Xi_K = 1 \). From \( \Xi_K = 1 \), we see that there is an order 2 automorphism \( \tau \) of \( G = \langle K \rangle \) with \( k^\tau = k^{-1} \) for all \( k \in K \), so we work in the semidirect product of \( G \) and \( \langle \tau \rangle \), with \( \tau^2 = 1 \) and \( \tau g \tau = g^\tau \) for \( g \in G \).
Let \( x, y \in K \) be representatives of the nontrivial elements of \( (K \cap M)/N \). We may assume without loss of generality that \( x, y \) have orders equal to powers of 2, since otherwise we may replace them with odd powers of themselves. Let \( S_x, S_y \) be 2-Sylow subgroups of \( M/\langle \tau \rangle \) containing \( \langle x, \tau \rangle, \langle y, \tau \rangle \), respectively, then by the Sylow theorems there is some \( g \in M/\langle \tau \rangle \) with \( g^{-1}S_yg = S_x \). Then \( x, \tau, g^{-1}yg, g^{-1}\tau g \in S_x \), and our strategy is to show that \( x, g^{-1}yg \in K \cap S_x \).

We have \( g^{-1}\tau yg \in \tau K \) by the previous proposition, so \( \tau g^{-1}\tau yg \in K \cap S_x \), and similarly \( \tau g^{-1}tg \in K \cap S_x \). Since \( K \cap S_x \) is assumed to be a subgroup, we have

\[
xg^{-1}yg = x(\tau g^{-1}tg)^{-1}(\tau g^{-1}tg) \in K \cap S_x.
\]

Then since \( M/N \) is abelian and \( \tau y \tau = y^{-1} \equiv N y \), we have \( xg^{-1}yg \equiv_N xy \), contradicting the assumption \( |(K \cap M)/N| = 3 \).

\[\square\]

**Proof of Theorem 10.13.** If \( K, K' \) are near-subgroups of \( G \), then they are both twisted subgroups and so their intersection \( K \cap K' \) is also a twisted subgroup. Now let \( S \) be any 2-group contained in \( G \), then for any \( b \in K \cap K' \) we have \( bK \cap bK' \cap S = (bK \cap S) \cap (bK' \cap S) \) is an intersection of subgroups of \( S \), so it is a subgroup of \( S \), and the previous proposition shows that this implies that \( K \cap K' \) is a near-subgroup of \( G \).

\[\square\]

## 10 Abelian Mal’cev algebras are affine

In this section we will prove that abelian Mal’cev algebras are affine. This is an important step in the proof that problems which do not have the “ability to count” have bounded width. First we will carefully define what an affine algebra is, starting with the more basic concept of a quasi-affine algebra.

**Definition 10.1.** An algebra \( A \) is called **quasi-affine** if there is an abelian group \( G = (G, 0, +, -) \) with underlying set \( G \) containing the underlying set of \( A \), such that the 4-ary relation \( x + y = z + w \) is preserved by all the operations of \( A \).

We want to relate this to the more familiar concept of a module over a ring.

**Definition 10.2.** If \( R \) is a ring and \( M \) is a module over \( R \) with underlying group \( (M, 0, +, -, ) \), then we consider \( M \) to be a universal algebraic object \( (M, 0, +, -, \{\phi_r\}_{r \in R}) \), where for each \( r \in R \) the unary operation \( \phi_r : M \to M \) is given by \( \phi_r : m \mapsto rm \).

In general, a universal algebraic object is called a **module** if it is an expansion of an abelian group by any collection of unary operations that distribute over addition.

The way these concepts are related is a coarser notion than term equivalence, known as **polynomial equivalence** (warning: in some older references, “polynomial equivalence” means term equivalence and “functional equivalence”/“algebraic equivalence” means polynomial equivalence).

**Definition 10.3.** If \( O \) is any set of operations, then the **polynomial clone** generated by \( O \) is the clone generated by \( O \) together with the constant functions (one for each element of the underlying set). Two algebras or clones on the same underlying set are called **polynomially equivalent** if they have the same polynomial clones.

**Proposition 10.4.** An algebra \( A \) is quasi-affine if and only if it is a subalgebra of a reduct of the polynomial clone of a module.
Proof. Let \( A \) be a quasi-affine algebra, and let \((G, 0, +, -)\) be the corresponding group. We may assume without loss of generality that \( 0 \in A \). Suppose that \( f \) is any \( n \)-ary operation of \( A \), and for each \( i \leq n \) let \( \phi_i \) be the unary operation given by

\[
\phi_i(x) = f(0, ..., 0, x, 0, ..., 0) - f(0, ..., 0),
\]

with the \( x \) in the \( i \)th position. Since \( f \) preserves the relation \( x + y = z + w \), we have

\[
\phi_i(x + y) = \phi_i(x) + \phi_i(y),
\]

so each \( \phi_i \) distributes over addition. To finish, we just need to prove that

\[
f(x_1, ..., x_n) = \phi_1(x_1) + \cdots + \phi_n(x_n) + f(0, ..., 0)
\]

for all \( x_1, ..., x_n \in A \), since \( f(0, ..., 0) \) is a constant operation.

We prove this by induction on the number \( k \) of nonzero values among \( x_1, ..., x_n \). The base cases \( k = 0, 1 \) follow from the definition of the \( \phi_i \). For the inductive step, assume without loss of generality that the nonzero values of the \( x_i \)'s are \( x_1, ..., x_{k+1} \). Since \( f \) preserves the relation \( x + y = z + w \), we have

\[
f(x_1, ..., x_{k+1}, 0, ..., 0) + f(0, ..., 0) = f(x_1, ..., x_k, 0, 0, ..., 0) + f(0, ..., 0, x_{k+1}, 0, ..., 0),
\]

so by the inductive hypothesis and the definition of \( \phi_{k+1} \) we have

\[
f(x_1, ..., x_{k+1}, 0, ..., 0) = \phi_1(x_1) + \cdots + \phi_k(x_k) + f(0, ..., 0) + \phi_{k+1}(x_{k+1}).
\]

\[ \square \]

Definition 10.5. An algebra \( A \) is called affine if it is polynomially equivalent to a module.

Proposition 10.6. An algebra is affine iff it is quasi-affine and has a Mal'cev term.

Proof. The hardest step is showing that every affine algebra \( A \) has a Mal'cev term. Since \( A \) is polynomially equivalent to a module, there must be some \( n + 3 \)-ary term \( t \) and some constants \( a_1, ..., a_n \in A \) such that

\[
t(x, y, z, a_1, ..., a_n) = x - y + z
\]

for all \( x, y, z \). Since any affine algebra is quasi-affine, we can write \( t \) in the form

\[
t(x, y, z, u_1, ..., u_n) = x - y + z + \sum_i \phi_i(u_i) + c
\]

for some unary \( \phi_i \) and some constant \( c \). Define \( p(x, y, z) \) by

\[
p(x, y, z) = t(x, t(y, y, x, ..., x), z, x, ..., x).
\]

Then \( p \) is a term of \( A \), and we have

\[
p(x, y, z) = x - (y - y + \sum_i \phi_i(x) + c) + z + \sum_i \phi_i(x) + c = x - y + z,
\]

so \( p \) is Mal'cev.

For the converse, if \( A \) is quasi-affine and has a Mal'cev term \( p \), then \( p(x, y, y) \approx p(y, y, x) \approx x \) imply that \( p(x, 0, 0) = x, p(0, 0, z) = z \), and \( p(y, y, 0) = y + p(0, y, 0) = 0 \), so we must have \( p(x, y, z) = x - y + z \). Thus \( x + z = p(x, 0, z) \) and \( x - y = p(x, y, 0) \) are polynomial operations of \( A \), and therefore for each term \( f \) of \( A \) the unary function \( \phi(x) = f(x, 0, ..., 0) - f(0, ..., 0) \) is a polynomial operation of \( A \) as well. \[ \square \]
It is less trivial to give a universal algebraic definition of what it means to be abelian. We will give several different definitions and prove that they are equivalent to each other, and that they restrict to the right concept in the special case of groups.

**Definition 10.7.** An algebraic structure \( \mathbb{A} \) is called abelian if there is a congruence \( \Theta \) on \( \mathbb{A} \) such that the diagonal \( \Delta_{\mathbb{A}} = \{(a, a) \mid a \in \mathbb{A}\} \) is one of the congruence classes of \( \Theta \).

**Proposition 10.8.** A group is abelian iff it is commutative.

*Proof.* A group \( G \) is abelian iff the diagonal \( \Delta_G \) is a normal subgroup of \( G \times G \). To check that \( \Delta_G \) is normal, we just need to check that it is closed under conjugation by elements of the form \((1, b)\) for all \( b \in G \). Since

\[
(1, b)(a, a)(1, b)^{-1} = (a, bab^{-1}),
\]

the normality of \( \Delta_G \) is equivalent to the identity \( a \approx bab^{-1} \), which is equivalent to \( ab \approx ba \).

Alternatively, we can argue as follows. The group \( G \) is commutative iff the map \( G \to G \) given by \( x \mapsto x^{-1} \) is a homomorphism, and if this occurs then there is a homomorphism \( G \times G \to G \) such that the restriction \( G \times \{1\} \to G \) is the identity, and such that the diagonal maps to \( \{1\} \). Conversely, if the diagonal is a normal subgroup, then every coset intersects \( G \times \{1\} \) and \( \{1\} \times G \) exactly once, so the quotient \( G \times G / \Delta_G \) is isomorphic to \( G \) in two different ways, and composing these isomorphisms we obtain the map \( x \mapsto x^{-1} \), so \( G \) is commutative. \( \square \)

Now we give a second definition of abelian, which is phrased in a way which is closely related to the concept of a “commutator” of congruences in a general algebraic structure.

**Definition 10.9.** We say that an algebraic structure \( \mathbb{A} \) satisfies the term condition if for all terms \( t \in \text{Clo}_{n+1}(\mathbb{A}) \) and all \( u, v \in \mathbb{A} \), \( a_i, b_i \in \mathbb{A} \) for \( i \leq n \), we have

\[
t(u, a_1, \ldots, a_n) = t(u, b_1, \ldots, b_n) \iff t(v, a_1, \ldots, a_n) = t(v, b_1, \ldots, b_n).
\]

**Proposition 10.10.** An algebra \( \mathbb{A} \) is abelian iff it satisfies the term condition.

*Proof.* We think of congruences on \( \mathbb{A}^2 \) as subalgebras of \( \mathbb{A}^{2 \times 2} \), the set of \( 2 \times 2 \) matrices with entries in \( \mathbb{A} \) (here elements of \( \mathbb{A}^2 \) are visualized as column vectors, and an element of \( \mathbb{A}^{2 \times 2} \) is viewed as a row vector of column vectors). To understand the smallest congruence on \( \mathbb{A}^2 \) with \( \Delta_{\mathbb{A}} \) contained in a congruence class, we consider the relation \( \overline{m} \leq \mathbb{A}^{2 \times 2} \) generated by matrices of the form

\[
\begin{bmatrix}
u & v \\ u & v
\end{bmatrix},
\begin{bmatrix}
a & a \\ b & b
\end{bmatrix},
\]

where the first type of matrix corresponds to the fact that any two elements of \( \Delta_{\mathbb{A}} \) are congruent, while the second type of matrix corresponds to the fact that every element of \( \mathbb{A}^2 \) is congruent to itself. Then considering \( \overline{m} \) as a binary relation on \( \mathbb{A}^2 \), the transitive closure of \( \overline{m} \) is a congruence \( \Theta \) on \( \mathbb{A}^2 \), and it is clearly as small as possible given that \( \Delta_{\mathbb{A}} \) is contained in a congruence class of \( \Theta \).

To understand whether \( \Delta_{\mathbb{A}} \) is a congruence class of \( \Theta \), it’s enough to check whether \( \Delta_{\mathbb{A}} \) meets any element of \( \mathbb{A}^2 \setminus \Delta_{\mathbb{A}} \) in \( \overline{m} \). This occurs (that is, \( \mathbb{A} \) is nonabelian) iff there is some term \( t \in \text{Pol}_{m+n}(\mathbb{A}) \) and some \( u_i, v_i \in \mathbb{A} \) for \( i \leq m \), \( a_i, b_i \in \mathbb{A} \) for \( i \leq n \) such that

\[
t(u_1, \ldots, u_m, a_1, \ldots, a_n) = t(u_1, \ldots, u_m, b_1, \ldots, b_n)
\]
but
\[ t(v_1, \ldots, v_m, a_1, \ldots, a_n) \neq t(v_1, \ldots, v_m, b_1, \ldots, b_n). \]
So if \( \mathbb{A} \) is abelian, then it certainly satisfies the term condition (just take \( m = 1 \) in the above). Conversely, if \( \mathbb{A} \) satisfies the term condition, then we will show that the above situation can’t happen by induction on \( m \). We just note that by the induction hypothesis, we have
\[ t(u_1, \ldots, u_m, a_1, \ldots, a_n) = t(u_1, \ldots, u_m, b_1, \ldots, b_n), \]
and then by the term condition applied to a version of \( t \) with variables permuted so that the \( m \)th variable becomes the first, this implies that
\[ t(v_1, \ldots, v_m, a_1, \ldots, a_n) = t(v_1, \ldots, v_m, b_1, \ldots, b_n). \]

\[ \text{Proposition 10.11.} \quad \text{Every quasi-affine algebra satisfies the term condition and is therefore abelian.} \]

\[ \text{Proof.} \quad \text{If} \quad t \quad \text{is an} \quad n+1 \quad \text{-ary term of a quasi-affine algebra, then we can write} \quad t \quad \text{in the form} \]
\[ t(x_0, \ldots, x_n) = \phi_0(x_0) + \cdots + \phi_n(x_n) + c, \]
where the \( \phi_i \) are unary and \( c \) is a constant. Then for any \( u \in \mathbb{A}, a_i, b_i \in \mathbb{A}, \) we have
\[ t(u, a_1, \ldots, a_n) = t(u, b_1, \ldots, b_n) \iff \phi_1(a_1) + \cdots + \phi_n(a_n) = \phi_1(b_1) + \cdots + \phi_n(b_n), \]
and this is a condition which does not depend on the value of \( u \).

\[ \text{Example 10.1.} \quad \text{If a group is commutative, then it is affine, so it satisfies the term condition. Conversely, if a group satisfies the term condition for the binary term} \quad t(x, y) = yxy^{-1}, \quad \text{then the group is commutative, since we have} \quad t(1, 1) = t(1, y) \iff t(x, 1) = t(x, y), \quad \text{that is,} \quad 1 = yy^{-1} \iff x = yxy^{-1}. \]

\[ \text{Example 10.2.} \quad \text{A ring is abelian in the sense of universal algebra iff it is a zero ring, that is, a ring satisfying the identity} \quad xy \approx 0. \quad \text{To see the necessity, we apply the term condition with the term} \quad t(x, y) = xy \quad \text{and the pairs} \quad (u, v) = (0, x) \quad \text{and} \quad (a, b) = (0, y), \quad \text{to see that} \quad 0 \cdot 0 = 0 \cdot y \iff x \cdot 0 = x \cdot y. \]
\[ \text{To see the sufficiency, note that every zero ring is affine.} \]

\[ \text{Example 10.3.} \quad \text{The quasigroup with multiplication table} \]
\[
\begin{array}{c|ccc}
\cdot & 0 & 1 & 2 & 3 \\
\hline
0 & 3 & 2 & 0 & 1 \\
1 & 2 & 3 & 1 & 0 \\
2 & 1 & 0 & 2 & 3 \\
3 & 0 & 1 & 3 & 2 \\
\end{array}
\]
is abelian, but is neither commutative nor associative. In fact it is affine, with underlying group equal to the Klein four-group: the multiplication can be written as \( x \cdot y = x \oplus \phi(y) \oplus 3 \), where \( \phi \) is the transposition \( (2 \ 3) \). This example is from [54].

In terms of congruence lattices, the main important feature of an affine algebra \( \mathbb{A} \) is that \( \text{Con}(\mathbb{A} \times \mathbb{A}) \) contains the following five element sublattice.

---

64
The abstract five element lattice corresponding to this picture is known as the diamond lattice \( M_3 \). The lattice \( M_3 \) has a special role in lattice theory: every modular lattice which isn’t distributive contains a sublattice which is isomorphic to \( M_3 \) (see Proposition A.44 in the appendix).

**Theorem 10.12.** If \( \mathbb{A} \) is an abelian Mal’cev algebra, and if \( \Theta \) is any congruence of \( \mathbb{A}^2 \) which contains the diagonal \( \Delta_\mathbb{A} \) as a congruence class, then the congruences \( \Theta, \ker \pi_1, \ker \pi_2 \) generate a five element sublattice of \( \text{Con}(\mathbb{A}^2) \) isomorphic to \( M_3 \).

**Proof.** In general, we always have \( \ker \pi_1 \vee \ker \pi_2 = 1_{\mathbb{A}^2} \) and \( \ker \pi_1 \wedge \ker \pi_2 = 0_{\mathbb{A}^2} \). Since every element of \( \mathbb{A} \) is congruent under \( \ker \pi_1 \) to an element of the diagonal \( \Delta_\mathbb{A} \), we have \( \ker \pi_1 \vee \Theta = 1_{\mathbb{A}^2} \), and similarly \( \ker \pi_2 \vee \Theta = 1_{\mathbb{A}^2} \).

All that remains is to check that \( \Theta \wedge \ker \pi_1 = \Theta \wedge \ker \pi_2 = 0_{\mathbb{A}^2} \), and this is where we will use the assumption that \( \mathbb{A} \) has a Mal’cev term \( p \). If \( (a, b) \) is congruent to \( (c, d) \) modulo \( \Theta \wedge \ker \pi_1 \), then we must have \( a = c \). Then

\[
\begin{bmatrix} b \\ d \end{bmatrix} = p \begin{bmatrix} b \\ b \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} a \\ d \end{bmatrix} \equiv p \begin{bmatrix} b \\ b \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} b \\ b \end{bmatrix} \in \Delta_\mathbb{A},
\]

so \( (b, d) \in \Delta_\mathbb{A} \), that is, \( b = d \). So from \( (a, b) \equiv \Theta \wedge \ker \pi_1 (c, d) \) we have shown \( (a, b) = (c, d) \), that is, we have \( \Theta \wedge \ker \pi_1 = 0_{\mathbb{A}^2} \).

The idea now is to study the equivalence class geometry on \( \mathbb{A}^2 \), where points are elements of \( \mathbb{A}^2 \), lines correspond to congruence classes of congruences, and two lines are considered parallel if they are both congruence classes of the same congruence. The three congruences \( \ker \pi_1, \Theta, \ker \pi_2 \) on an abelian Mal’cev algebra give us a particularly nice type of combinatorial geometry.

**Definition 10.13.** An S-3-system is a set of points \( S \) together with three parallel classes of lines \( \Theta_1, \Theta_2, \Theta_3 \) on \( S \), which satisfy the following properties:

- for any point \( p \in S \) and any \( i \leq 3 \), there is exactly one line \( l_i \) of \( \Theta_i \) which contains \( p \), and
- if \( l_i, l_j \) are lines of \( \Theta_i, \Theta_j \), respectively, with \( i \neq j \), then their intersection \( l_i \cap l_j \) contains exactly one point \( p \in S \).

Equivalently, an S-3-system is a relational structure \( (S, \Theta_1, \Theta_2, \Theta_3) \) such that:

- each \( \Theta_i \) is an equivalence relation on \( S \),
- for \( i \neq j \) we have \( \Theta_i \wedge \Theta_j = 0_S \), and
- for \( i \neq j \) we have \( \Theta_i \circ \Theta_j = 1_S \).
The assumption $\Theta_i \wedge \Theta_j = 0_S$ says that any pair of non-parallel lines intersect in at most one point, while the assumption $\Theta_i \circ \Theta_j = 1_S$ says that any pair of non-parallel lines intersect in at least one point.

**Corollary 10.14.** If $A$ is an abelian Mal’cev algebra and $\Theta$ is any congruence of $A^2$ with the diagonal as a congruence class, then $(A^2, \ker \pi_1, \ker \pi_2, \Theta)$ is an S-3-system with a Mal’cev polymorphism.

From here on we will classify S-3-systems which have Mal’cev polymorphisms, following Gumm’s approach \[95\]. As a preliminary result, we will show that every S-3-system has a coordinate system which describes the three parallel classes of lines in terms of a loop (recall that a loop is just a quasigroup which has an identity).

**Lemma 10.15.** If $(S, \Theta_1, \Theta_2, \Theta_3)$ is an S-3-system, and $e$ is any point of $S$, then there is a loop $L = (L, \cdot, 1)$ and a bijection $L \times L \to S$ with $(1,1) \mapsto e$, such that for any $x,y,x',y' \in L$ we have

$$
(x,y) \equiv_{\Theta_1} (x',y') \iff x = x',
$$

$$
(x,y) \equiv_{\Theta_2} (x',y') \iff y = y',
$$

$$
(x,y) \equiv_{\Theta_3} (x',y') \iff x \cdot y = x' \cdot y',
$$

where we have implicitly identified $S$ with $L \times L$.

**Proof.** Take $L$ to be the line $l_1$ through $e$ in the parallel class $\Theta_1$, and take $1 = e$. Let $l_2$ be the line through $e$ in the parallel class $\Theta_2$. Then there is a bijection between elements of $l_1$ and elements of $l_2$, taking $x \in l_1$ to $y \in l_2$ when $x,y$ are on a line $l_3$ in the parallel class $\Theta_3$: each $x$ is in a unique such line $l_3$, and each $l_3$ intersects $l_2$ in a unique $y$. Using this bijection, we identify the elements of $l_2$ with $L$ as well.

Now we note that for any point $p \in S$, there is a unique pair of lines $l'_1 \in \Theta_1, l'_2 \in \Theta_2$ with $l'_1 \cap l'_2 = \{p\}$. So we can uniquely identify the point $p$ by describing the point $x \in l_1 \cap l'_2$ and the point $y \in l_2 \cap l'_1$ - this gives us the desired bijection between $L \times L$ and $S$.

Finally, to define the multiplication $\cdot$ on $L$, note that for every $x,y \in L$ there is a point $p \in S$ corresponding to $(x,y)$, and this point $p$ is in a unique line $l_3 \in \Theta_3$. We then define $x \cdot y$ to be the element of $L$ corresponding to the point $l_3 \cap l_1$, or alternatively to the point $l_3 \cap l_2$ (which corresponds to the same element of $L$ by the way we identified points of $l_2$ with points of $l_1$).

The key observation is that the Mal’cev operation is completely determined by the geometry of the configuration.

**Lemma 10.16.** If an S-3-system $S = (S, \Theta_1, \Theta_2, \Theta_3)$ has a Mal’cev polymorphism $p$, then $p$ is completely determined by $S$. In fact, $p(x,y,z)$ can be “geometrically constructed” from the points $x,y,z$. 

66
Proof. First consider the special case where $x, y$ lie on a line $l_1$ and $y, z$ lie on a different line $l_2$. Suppose that $l_1 \in \Theta_1$ and $l_2 \in \Theta_2$. Then $p(x, y, z) \equiv_{\Theta_1} p(y, y, z) = z$ and $p(x, y, z) \equiv_{\Theta_2} p(x, y, y) = x$, so if we draw the line $l_2' \in \Theta_2$ through $x$ and the line $l_1' \in \Theta_1$ through $z$, we see that $p(x, y, z)$ is the intersection point $l_1' \cap l_2'$.

Next consider the special case where $x, y, z$ lie on a line $l_1$, and suppose $l_1 \in \Theta_1$. Draw the line $l_2 \in \Theta_2$ through $y$ and the line $l_3 \in \Theta_3$ through $x$, and let $y' \in l_2 \cap l_3$ be their point of intersection. Draw the line $l_1'$ through $y'$ parallel to $l_1$, draw the line $l_2'$ through $y'$ parallel to $l_2$, and let $z' \in l_1' \cap l_2'$ be their point of intersection. Finally, draw the line $l_3'$ through $z'$ parallel to the line $l_3$, and let $p$ be the intersection point of $l_1$ and $l_3'$.

We claim that $p = p(x, y, z)$. To see this, note that $x \equiv_{\Theta_3} y'$, so $p(x, y, z) \equiv_{\Theta_3} p(y', y, z)$, and $p(y', y, z) = z'$ by the first case we considered. Thus $p(x, y, z) \equiv_{\Theta_3} z'$, i.e. $p(x, y, z) \in l_3'$, and since $x \equiv_{\Theta_1} y \equiv_{\Theta_1} z$, we have $p(x, y, z) \equiv_{\Theta_1} p(x, x, x) = x$, i.e. $p(x, y, z) \in l_1$. Thus $p(x, y, z) \in l_1 \cap l_3'$, so $p(x, y, z) = p$. (Alternatively, we could have used $p(x, y, z) \equiv_{\Theta_2} p(x, y', z') = p$, by the first case.)

For the general case, we can pick any lines $l_1 \in \Theta_1, l_2 \in \Theta_2$, set $x_1, y_1, z_1$ to be the projections of $x, y, z$ onto $l_1$ via lines in $\Theta_2$ and define $x_2, y_2, z_2 \in l_2$ similarly, and note that $p(x, y, z) \equiv_{\Theta_2} p(x_1, y_1, z_1)$ and $p(x, y, z) \equiv_{\Theta_1} p(x_2, y_2, z_2)$, and we can construct $p(x_1, y_1, z_1), p(x_2, y_2, z_2)$ using the second case considered.

**Corollary 10.17.** If $p$ is a Mal’cev polymorphism of an $S$-$\beta$-system, then $p(x, y, z) \approx p(z, y, x)$.

**Proof.** The term $p(z, y, x)$ is also a Mal’cev polymorphism, so by the Lemma it must be identical to $p(x, y, z)$. □
Corollary 10.18. If $p$ is a Mal’cev polymorphism of an $S$-3-system $(S, \Theta_1, \Theta_2, \Theta_3)$, then the graph $\Gamma_p$ of $p$, considered as a 4-ary relation on $S$, is primitively positively definable from $\Theta_1, \Theta_2, \Theta_3$.

Corollary 10.18 can also be interpreted as saying that the map $p : S^3 \to S$ is a homomorphism of the algebraic structure $S$ consisting of all polymorphisms of the relational structure $S$. In particular, $p$ “commutes with itself”, that is, the two ways of computing $p \ast p$ on a $3 \times 3$ grid of variables (columns first or rows first) agree with each other. We can summarize this fact by saying that the Mal’cev operation $p$ is central.

Definition 10.19. An $n$-ary term $t$ of an algebraic structure $A$ is called central if the map $t : A^n \to A$ is a homomorphism.

Now we relate the Mal’cev polymorphism to the coordinate loop $L$. First we will show that $L$ is associative.

Lemma 10.20. If $S = (S, \Theta_1, \Theta_2, \Theta_3)$ is an $S$-3-system with a Mal’cev polymorphism $p$, and if $L$ is a coordinate loop of $S$, then $L$ satisfies

$$(x_1 \cdot y_1 = x_2 \cdot y_2) \land (x_1 \cdot y_3 = x_2 \cdot y_4) \land (x_3 \cdot y_1 = x_4 \cdot y_2) \implies (x_3 \cdot y_3 = x_4 \cdot y_4).$$

In particular, $L$ is associative, that is, $L$ is a group.

Proof. For those who prefer a purely algebraic proof, this follows from

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = p \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_1 \\ y_3 \end{bmatrix}, \begin{bmatrix} x_3 \\ y_1 \end{bmatrix} \equiv_{\Theta_3} p \begin{bmatrix} x_2 \\ y_2 \\ y_4 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_4 \end{bmatrix}, \begin{bmatrix} x_4 \\ y_4 \end{bmatrix} = \begin{bmatrix} x_4 \\ y_4 \end{bmatrix}.$$

To see that this implies the associativity of $L$, let $x, y, z$ be any elements of $L$, and plug in $(x_1, x_2, x_3, x_4) = (1, y, x, x \cdot y), (y_1, y_2, y_3, y_4) = (y, 1, y \cdot z, z)$. Then we get

$$(1 \cdot y = y \cdot 1) \land (1 \cdot (y \cdot z) = y \cdot z) \land (x \cdot y = (x \cdot y) \cdot 1) \implies (x \cdot (y \cdot z) = (x \cdot y) \cdot z).$$

For a geometric way to visualize the proof, note that the stated property of $L$ corresponds to the existence of the dashed line in the following picture.
If we set \( a = (x_1, y_3) \), etc. as in the picture, then the existence of the dashed line follows from the fact that \( p \) preserves the congruence \( \Theta_3 \) and the fact that \( p(a, b, c) \) completes the parallelogram through \( a, b, c \) and \( p(a', b', c') \) completes the parallelogram through \( a', b', c' \).

Lemma 10.21. If \( S = (S, \Theta_1, \Theta_2, \Theta_3) \) is an \( S \)-3-system with a Mal’cev polymorphism \( p \), and if \( L \) is a coordinate group of \( S \), then for \( x = (x_1, x_2), y = (y_1, y_2), z = (z_1, z_2) \in S \), \( p(x, y, z) \) is given by

\[
p\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}\right) = \begin{bmatrix} x_1 \cdot y_1^{-1} \cdot z_1 \\ x_2 \cdot y_2^{-1} \cdot z_2 \end{bmatrix}.
\]

Proof. It’s enough to consider the case where \( x, y, z \) are along the line \( l_1 \in \Theta_1 \) with \( \Theta_1 \)-coordinate 1. Consider the diagram

![Diagram](image)

which we used to construct \( p(x, y, z) \). Then from \( (1, x) \equiv_{\Theta_3} (u, y) \) we have \( 1 \cdot x = u \cdot y \), and from \( (1, p) \equiv_{\Theta_3} (u, z) \) we have \( 1 \cdot p = u \cdot z \). Solving for \( u \) we get \( u = xy^{-1} \), and solving for \( p \) we get \( p = xy^{-1}z \).

Corollary 10.22. If \( S = (S, \Theta_1, \Theta_2, \Theta_3) \) is an \( S \)-3-system with a Mal’cev polymorphism \( p \), and if \( L \) is a coordinate group of \( S \), then \( L \) is commutative.

Proof. From \( p(x, y, z) \approx p(z, y, x) \) we get \( xy^{-1}z \approx yz^{-1}x \) in \( L \), and plugging in \( y = 1 \) gives \( xz \approx zx \), so \( L \) is commutative.

Putting all of this together, we have the main result of this section.

Theorem 10.23. Any abelian Mal’cev algebra \( \mathbb{A} \) is affine.

Proof. By Theorem 10.12 and its corollary, \( S = (\mathbb{A}^2, \ker \pi_1, \ker \pi_2, \Theta) \) is an \( S \)-3-system with Mal’cev polymorphism \( p \), where \( p \) is the Mal’cev term of \( \mathbb{A} \) and \( \Theta \) is a congruence on \( \mathbb{A}^2 \) with the diagonal as a congruence class. By Lemma 10.15 there is a loop structure \( L \) on the underlying set of \( \mathbb{A} \) which describes \( S \). By Lemma 10.20 and its corollary, \( L \) is an abelian group and \( p \) is given by \( p(x, y, z) = x - y + z \) (writing the abelian group operation additively).

By Corollary 10.18, the relation \( x - y + z = p \) is primitively positively definable from \( \ker \pi_1, \ker \pi_2, \Theta \), so the relation \( x + z = y + p \) is preserved by all operations of \( \mathbb{A} \), that is, \( \mathbb{A} \) is quasi-affine. Since \( \mathbb{A} \) was assumed to be Mal’cev, this means that \( \mathbb{A} \) is affine.

We have proved the hardest part of the Fundamental Theorem of Abelian Algebras. For the sake of completeness, we include the rest of it.

Theorem 10.24 (Fundamental Theorem of Abelian Algebras). For an algebraic structure \( \mathbb{A} \), the following are equivalent:

(1) \( \mathbb{A} \) is affine,
(2) \( A \) is abelian and has a Mal’cev polynomial,

(3) \( A \) has a central Mal’cev polynomial.

Proof. That (1) implies (2) and (3) is clear. For (2) implies (1) and (3), note that any polynomial of \( A \) preserves every congruence of \( A^2 \), so the polynomial clone of \( A \) is also abelian and we may apply the previous theorem. For (3) \( \implies \) (1), we just need to show that any Mal’cev operation \( p \) which commutes with itself comes from an abelian group, since then the fact that \( p(x, y, z) = x - y + z \) is central will imply that \( A \) is quasi-affine.

So suppose that \( p \) is a Mal’cev operation which commutes with itself, and pick any element to call 0 in \( A \). We define addition and negation on \( A \) by

\[
    x + y := p(x, 0, y), \quad -x := p(0, x, 0).
\]

That 0 is an identity element for + follows from the Mal’cev identities \( p(x, 0, 0) = p(0, 0, x) = x \).

To see that + is associative, we evaluate the expression

\[
    p * p \left( \begin{bmatrix} x & 0 & y \\ 0 & 0 & 0 \\ 0 & 0 & z \end{bmatrix} \right)
\]

in two ways: evaluating it by rows first, we get \((x + y) + z\), and evaluating it by columns first, we get \(x + (y + z)\).

To see that – computes the inverse, we evaluate the expression

\[
    p * p \left( \begin{bmatrix} x & 0 & 0 \\ 0 & 0 & x \\ 0 & 0 & 0 \end{bmatrix} \right)
\]

in two ways: by rows we get \(p(x, x, 0) = 0\), and by columns we get \(x + (-x)\). A similar argument shows that \((-x) + x = 0\).

For commutativity of +, we evaluate the expression

\[
    p * p \left( \begin{bmatrix} y & 0 & x \\ y & y & x \\ x & y & y \end{bmatrix} \right)
\]

in two ways: by rows we get \(p(y + x, x, x) = y + x\), and by columns we get \(p(x, 0, y) = x + y\).

Finally, to express \( p \) in terms of the group operations +, –, we evaluate the expression

\[
    p * p \left( \begin{bmatrix} x & y & z \\ 0 & y & 0 \\ -y & 0 & 0 \end{bmatrix} \right)
\]

in two ways: by rows we get \(p(p(x, y, z), -y, -y) = p(x, y, z)\), and by columns we get \(p(x - y, 0, z) = x - y + z\). \(\square\)

The method of visualizing algebraic arguments via the geometry of equivalence classes was extended to congruence modular varieties by Gumm in his book “Geometrical methods in congruence...
modular algebras” [59], where he used it to show that any abelian algebra in a congruence modular variety is affine. This was extended further by Hobby and McKenzie [65], who used Tame Congruence Theory to show that any finite abelian algebra in a Taylor variety is affine (in the infinite case, Kearnes and Szendrei [74] show that any abelian Taylor algebra is quasi-affine - the example \((\mathbb{R}, +, \cdot)\) shows that an additional assumption is needed for it to be affine). Later we will go over a simpler proof of the fact that finite abelian Taylor algebras are affine, from [17].

**Remark 10.1.** If we leave the context of Taylor varieties, we can no longer expect abelian algebras to be affine, since they could fail to have any interesting operations at all. But we can still ask whether abelian algebras are quasi-affine. The following problem is open.

**Problem 10.1.** Under what conditions are abelian algebras quasi-affine? Is it true that every idempotent abelian algebra is quasi-affine?

It is known that if we drop idempotence, then some extra condition is needed: Quackenbush [99] gives an example of an infinite, non-idempotent algebra which is abelian but not quasi-affine. Quackenbush’s example is a slight modification of the completely free algebra on 8 elements with a single binary operation, where the modification is that \(x_1 \cdot x_2 = x_5 \cdot x_6, x_3 \cdot x_4 = x_7 \cdot x_8, x_1 \cdot x_4 = x_5 \cdot x_8, \) but \(x_3 \cdot x_2 \not= x_7 \cdot x_6.\)

Kearnes [73] has shown that any *simple* idempotent abelian algebra is quasi-affine - in fact, he shows that any simple idempotent algebra which has a skew congruence (that is, a congruence on some power \(A^n\) which is not the kernel of some projection) either has an absorbing element (that is, an element \(a\) such that every term \(t\) which depends on its first variable has \(t(a,...) = a\)) or is a subalgebra of a simple reduct of a module.

There are a few other contexts in which it is known that abelian implies quasi-affine. In [71], Kearnes shows that any abelian algebra with a central binary polynomial which is cancellative is quasi-affine, and in [105] this is extended to the result that any abelian algebra with a commutative cancellative polynomial is quasi-affine. In [68], it is shown that abelian quandles are quasi-affine.

### 10.1 Commutators

In this subsection we define an extension of the commutator from group theory to a commutator on congruences of general algebraic structures. The purpose of the commutator is to detect situations where the operations of an algebraic structure behave linearly. The theory of the commutator works best in congruence modular varieties, but it still has some use in general Taylor varieties, although slight differences in the technical details of the definition become important outside the world of congruence modular varieties. The commutator we will be discussing is called the *term condition* commutator.

**Definition 10.25.** If \(\alpha, \beta, \delta \in \text{Con}(A)\), we say that \(\alpha\) centralizes \(\beta\) modulo \(\delta\), written \(C(\alpha, \beta; \delta)\) (or \(C(\alpha, \beta)\) if \(\delta = 0_A\)), if for every \(n + 1\)-ary term \(t \in \text{Clo}_{n+1}(A)\), for any \((u, v) \in \alpha\), and for any \((a_1, b_1), ..., (a_n, b_n) \in \beta\), we have

\[
t(u, a_1, ..., a_n) \equiv_\delta t(u, b_1, ..., b_n) \iff t(v, a_1, ..., a_n) \equiv_\delta t(v, b_1, ..., b_n).
\]

The smallest \(\delta\) which satisfies \(C(\alpha, \beta; \delta)\) is called the *commutator* of \(\alpha, \beta\), and is written as \([\alpha, \beta]\). If \(\theta \leq \alpha, \beta\), then we also define the *relative commutator* \([\alpha, \beta]_\theta\) to be the least \(\delta \geq \theta\) which satisfies \(C(\alpha, \beta; \delta)\).
As with the criterion for abelianness, the term condition implies a seemingly stronger version where more variables change at once.

**Proposition 10.26.** If $\alpha$ centralizes $\beta$ modulo $\delta$, then for every $m + n$-ary term $t \in \text{Clo}_{m+n}(\mathbb{A})$, for any $(u_1, v_1), \ldots, (u_m, v_m) \in \alpha$, and for any $(a_1, b_1), \ldots, (a_n, b_n) \in \beta$, we have

$$t(u_1, \ldots, u_m, a_1, \ldots, a_n) \equiv_\delta t(u_1, \ldots, u_m, b_1, \ldots, b_n) \iff t(v_1, \ldots, v_m, a_1, \ldots, a_n) \equiv_\delta t(v_1, \ldots, v_m, b_1, \ldots, b_n).$$

Before we go on, let’s check that this matches the usual commutator from group theory.

**Proposition 10.27.** If $\mathbb{M}, \mathbb{N}$ are normal subgroups of a group $G$, $[\mathbb{M}, \mathbb{N}]$ is the (normal) subgroup generated by commutators $[m, n] = mnm^{-1}n^{-1}$ for $m \in \mathbb{M}$, $n \in \mathbb{N}$, and $\theta_{\mathbb{M}}, \theta_{\mathbb{N}}, \theta_{[\mathbb{M}, \mathbb{N}]}$ are the associated congruences, then $\theta_{[\mathbb{M}, \mathbb{N}]} = [\theta_{\mathbb{M}}, \theta_{\mathbb{N}}]$.

**Proof.** We will show that $\theta_{\mathbb{M}}$ centralizes $\theta_{\mathbb{N}}$ iff every element of $\mathbb{M}$ commutes with every element of $\mathbb{N}$ - this will finish the proof, since $[\mathbb{M}, \mathbb{N}]$ is the smallest normal subgroup $\mathbb{K}$ of $G$ such that every element of $\mathbb{M}/\mathbb{K}$ commutes with every element of $\mathbb{N}/\mathbb{K}$ in $G/\mathbb{K}$.

First suppose that $\theta_{\mathbb{M}}$ centralizes $\theta_{\mathbb{N}}$. Let $t$ be the binary term $t(x, y) = yx$, then for any $m \in \mathbb{M}$, $n \in \mathbb{N}$, by the term condition applied to $(1, m) \in \theta_{\mathbb{M}}$, $(1, n) \in \theta_{\mathbb{N}}$, we have

$$1 = nn^{-1} \iff m = nmn^{-1},$$

so $m$ and $n$ commute.

Now suppose that every element of $\mathbb{M}$ commutes with every element of $\mathbb{N}$, and consider an arbitrary $n + 1$-ary term $t \in \text{Clo}_{n+1}(G)$ and any $(u, v) \in \theta_{\mathbb{M}}$, $(a_1, b_1), \ldots, (a_n, b_n) \in \theta_{\mathbb{N}}$ with

$$t(u, a_1, \ldots, a_n) = t(u, b_1, \ldots, b_n).$$

Thinking of $t(u, a_1, y_1, \ldots, a_n, y_n)$ as a function of $x, y_1, \ldots, y_n$ with parameters $u, a_1, \ldots, a_n$, we may rearrange it into the form

$$t(u, a_1, y_1, \ldots, a_n, y_n) = t'(x, y_1, \ldots, y_n)t(u, a_1, \ldots, a_n)$$

for some $t'$ in the clone generated by the group operations together with the unary conjugation operations $\phi_c : x \mapsto cxc^{-1}$, so we may rewrite our assumption as

$$t'(1, 1, \ldots, 1) = t'(1, a_1^{-1}b_1, \ldots, a_n^{-1}b_n).$$

To show that

$$t(v, a_1, \ldots, a_n) = t(v, b_1, \ldots, b_n),$$

we just need to show that

$$t'(u^{-1}v, 1, \ldots, 1) = t'(u^{-1}v, a_1^{-1}b_1, \ldots, a_n^{-1}b_n),$$

which follows from the assumed equality together with the fact that for each $c, d \in G$ and each $i$, $\phi_c(u^{-1}v) \in \mathbb{M}$ commutes with $\phi_d(a_i^{-1}b_i) \in \mathbb{N}$. \qed

**Example 10.4.** In the case of rings, the term condition commutator applied to a pair of ideals $I, J$ gives $[I, J] = IJ +JI$. Note that this is a bit different from what we might have expected (it has nothing to do with the Lie bracket), but it makes more sense when we remember that we only consider a ring to be abelian if it is a zero ring.
Example 10.5. In a majority algebra, the commutator is given by \([\alpha, \beta] = \alpha \land \beta\). To see this, suppose \((a, b) \in \alpha \land \beta\), and apply the term condition to the majority operation to see that
\[
m(\begin{bmatrix} a \\ a \\ a \end{bmatrix}, a, a) = m(\begin{bmatrix} a \\ a \\ a \end{bmatrix}, a, b) \implies m(\begin{bmatrix} b \\ a, a \end{bmatrix}, a, b),
\]
so \((a, b) \in [\alpha, \beta]\). A similar argument shows that the commutator is given by intersection in any variety with a near-unanimity term.

Example 10.6. In a semilattice, the commutator is given by \([\alpha, \beta] = \alpha \land \beta\). Let \(s\) be the semilattice operation, and let \(s_3\) be the term given by \(s_3(x, y, z) = s(x, s(y, z))\). Then for \((a, b) \in \alpha \land \beta\), we have
\[
s_3(\begin{bmatrix} a, a \end{bmatrix}, a, b) = s_3(\begin{bmatrix} a, a \end{bmatrix}, b, b) \implies s_3(\begin{bmatrix} b, a \end{bmatrix}, a, b) [\alpha, \beta] s_3(\begin{bmatrix} b, a \end{bmatrix}, b, b),
\]
so \(s(a, b) [\alpha, \beta] b\), and similarly \(s(a, b) [\alpha, \beta] a\), so \((a, b) \in [\alpha, \beta]\).

Sometimes it is helpful to visualize the term condition via \(2 \times 2\) matrices.

Definition 10.28. For \(\alpha, \beta \in \text{Con}(A)\), we define the algebra \(M(\alpha, \beta) \leq A^{2 \times 2}\) to be the subalgebra of \(2 \times 2\) matrices which is generated by the matrices of the form
\[
\begin{bmatrix} u & u \\ v & v \end{bmatrix}\text{ with } (u, v) \in \alpha, \quad \begin{bmatrix} a & b \\ a & b \end{bmatrix}\text{ with } (a, b) \in \beta.
\]

Proposition 10.29. If \(\alpha, \beta, \delta \in \text{Con}(A)\), then \(\alpha\) centralizes \(\beta\) modulo \(\delta\) iff for all
\[
\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in M(\alpha, \beta)
\]
we have
\[
a \equiv_{\delta} b \iff c \equiv_{\delta} d.
\]

The usual picture which is drawn to represent the term condition for \(C(\alpha, \beta; \delta)\) is this:

```
\begin{tikzpicture}
    \node (a) at (0,0) {t(\pi, \pi)};
    \node (b) at (3,0) {t(\pi, \pi)};
    \node (c) at (0,3) {t(\pi, \pi)};
    \node (d) at (3,3) {t(\pi, \pi)};
    \draw (a) -- (b) node [midway] {\alpha};
    \draw (c) -- (d) node [midway] {\beta};
    \draw (a) -- (c) node [midway, above] {\delta};
    \draw (b) -- (d) node [midway, above] {\delta};
    \draw (a) -- (d) node [midway, left] {\beta};
    \draw (b) -- (c) node [midway, left] {\alpha};
\end{tikzpicture}
```

where the positioning of the four corners matches with the way we have laid out the \(2 \times 2\) matrices in \(M(\alpha, \beta)\). A mnemonic for remembering where the \(\delta\) edges go is that in the term condition \(C(\alpha, \beta; \delta)\), “\(\delta\) is next to \(\beta\”.

We now prove a few elementary properties of the commutator which hold in general, which are given as exercises in Hobby and McKenzie’s book [65].

Proposition 10.30. For \(\alpha, \beta, \delta \in \text{Con}(A)\), we have

(a) if \(C(\alpha, \beta; \delta_i)\) for \(i \in I\), then \(C(\alpha, \beta; \bigwedge_{i \in I} \delta_i)\), so \([\alpha, \beta]\) and \([\alpha, \beta]_{\theta}\) are well-defined,

(b) if \((\alpha \lor (\beta \land \delta)) \land \beta \leq \delta\) then \(C(\alpha, \beta; \delta)\) holds, so \([\alpha, \beta] \leq \alpha \land \beta\),
(c) if \( \alpha' \leq \alpha, \beta' \leq \beta \), then \( \text{C}(\alpha, \beta; \delta) \Rightarrow \text{C}(\alpha', \beta'; \delta) \), so \([\alpha', \beta'] \leq [\alpha, \beta]\),

(d) for any \( \gamma \) we have \( \text{C}(\alpha, \beta; \delta) \Rightarrow \text{C}(\alpha \land \gamma, \beta; \delta \land \gamma) \),

(e) if \( \text{C}(\alpha_i, \beta; \delta) \) holds for all \( i \in I \) then \( \text{C}(\bigvee_{i \in I} \alpha_i, \beta; \delta) \) holds,

(f) if \( \theta \leq \alpha, \beta, \delta \) then \( \text{C}(\alpha, \beta; \delta) \) holds iff \( \text{C}(\alpha/\theta, \beta/\theta; \delta/\theta) \) holds in \( \mathbb{A}/\theta \), so \([\alpha/\theta, \beta/\theta] = [\alpha, \beta]_\theta/\theta\),

(g) if \( \mathbb{B} \leq \mathbb{A} \) then \( \text{C}(\alpha, \beta; \delta) \Rightarrow \text{C}(\alpha|_{\mathbb{B}}, \beta|_{\mathbb{B}}; \delta|_{\mathbb{B}}) \), so \([\alpha|_{\mathbb{B}}, \beta|_{\mathbb{B}}] \leq [\alpha, \beta]|_{\mathbb{B}}\),

(h) if \([\alpha, \alpha] = 0_\mathbb{A}\), then any congruence class of \( \alpha \) which is also a subalgebra of \( \mathbb{A} \) is an abelian subalgebra.

**Proof.** Parts (a), (c), (d), (f), (g), (h) follow immediately from the definitions. For (b), note that for any \( \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathcal{M}(\alpha, \beta) \) with \( a \equiv_{\delta} b \), we have \( c \equiv_{\delta} a \equiv_{\beta \land \delta} b \equiv_{\alpha} d \) and \( c \equiv_{\beta} d \), so \((c, d) \in (\alpha \circ (\beta \land \delta) \circ \alpha) \land \beta\), which is a subset of \( \delta \) by assumption.

For (e), we string together several instances of the term condition: if \((u, v) \in \bigvee \alpha_i, (a, b_i) \in \beta\), and \( t(u, \overline{a}) \equiv_{\delta} t(u, \overline{b}) \), then if we let \( u = u_0, u_1, ..., u_n = v \) be a sequence of elements of \( \mathbb{A} \) with \((u_i, u_{i+1}) \in \alpha_{j_i}\) for some \( j_i \in I \), then by the term condition \( \text{C}(\alpha_{j_i}, \beta; \delta) \) we have

\[
 t(u_i, \overline{a}) \equiv_{\delta} t(u_i, \overline{b}) \Rightarrow t(u_{i+1}, \overline{a}) \equiv_{\delta} t(u_{i+1}, \overline{b}),
\]

so by inding on \( i \) we get \( t(v, \overline{a}) \equiv_{\delta} t(v, \overline{b}) \). \(\)

**Corollary 10.31.** If an idempotent algebra \( \mathbb{A} \) has any congruences \( \alpha, \beta \in \text{Con}(\mathbb{A}) \) with \([\alpha, \beta] \neq [\alpha \land \beta] \), then some subalgebra of some quotient of \( \mathbb{A} \) is a nontrivial abelian algebra.

**Proof.** Let \( \delta = \alpha \land \beta \), then from \( \delta \leq \alpha, \beta \) we have \([\delta, \delta] \leq [\alpha, \beta] < \alpha \land \beta = \delta\). Thus \( \delta' = \delta/[\delta, \delta] \) is a nontrivial congruence on \( \mathbb{A}/[\delta, \delta] \) with 
\[
[\delta', \delta'] = [\delta, \delta]/[\delta, \delta] = 0_{\mathbb{A}/[\delta, \delta]},
\]

so there is some nontrivial congruence class \( \mathbb{B} \subseteq \delta' \) if \( \delta \subseteq \mathbb{B} \) and \( \mathbb{B} \) is an abelian subalgebra of \( \mathbb{A}/[\delta, \delta] \). \(\)

**Proposition 10.32.** If \([\alpha, \beta] = \alpha \land \beta \) for all \( \alpha, \beta \in \text{Con}(\mathbb{A}) \), then \( \text{Con}(\mathbb{A}) \) satisfies the meet semidistributive law:

\[
\alpha \land \beta = \alpha \land \gamma \Rightarrow \alpha \land (\beta \lor \gamma) = \alpha \land \beta.
\]

**Proof.** If \( \alpha, \beta, \gamma \in \text{Con}(\mathbb{A}) \) satisfy \( \alpha \land \beta = \alpha \land \gamma \), then \( \text{C}(\beta, \alpha \land \beta) \) and \( \text{C}(\gamma, \alpha \land \beta) \) hold, so \( \text{C}(\beta \lor \gamma, \alpha \land \beta) \) holds, so \( \alpha \land (\beta \lor \gamma) = [\beta \lor \gamma, \alpha] \leq \alpha \land \beta \). \(\)

**Definition 10.33.** An algebra \( \mathbb{A} \) is congruence meet semidistributive, written \( \text{SD}(\land) \) for short, if for all \( \alpha, \beta, \gamma \in \text{Con}(\mathbb{A}) \) with \( \alpha \land \beta = \alpha \land \gamma \), we have \( \alpha \land (\beta \lor \gamma) = \alpha \land \beta \). A variety \( \mathcal{V} \) is \( \text{SD}(\land) \) if every algebra \( \mathbb{A} \in \mathcal{V} \) is \( \text{SD}(\land) \).

The next corollary is the key to classifying CSPs which do not have the “ability to count” - as we will see later, a finite idempotent algebra generates an \( \text{SD}(\land) \) variety if and only if the associated CSP has bounded width.

**Corollary 10.34.** If an idempotent variety does not contain any nontrivial abelian algebras, then it is congruence meet semidistributive. Conversely, a congruence meet semidistributive variety does not contain any nontrivial affine algebra.
Proof. For the converse statement, note that if $A$ is affine, then $\text{Con}(A^2)$ contains a copy of the diamond lattice $M_3$, and $M_3$ doesn’t satisfy the meet semidistributive law.

Now we consider some definitions which are useful in the case where the commutator is not trivial (i.e., not given by $[\alpha, \beta] = \alpha \wedge \beta$).

**Definition 10.35.** Suppose that $\alpha \leq \beta \in \text{Con}(A)$. We say that $\beta$ is **abelian** over $\alpha$ if the term condition $C(\beta, \beta; \alpha)$ holds. We say that $\beta$ is **solvable** over $\alpha$ if there is a chain of congruences $\alpha = \alpha_0 \leq \cdots \leq \alpha_n = \beta$ such that $\alpha_{i+1}$ is abelian over $\alpha_i$ for each $i$.

A congruence $\alpha$ is called abelian if it is abelian over $0$ (equivalently $[\alpha, \alpha] = 0$), and similarly $\alpha$ is called solvable if $\alpha$ is solvable over $0$. An algebra $A$ is called solvable if $1$ is solvable.

The **center** of an algebra $A$ is defined to be the largest $\zeta$ such that $C(\zeta, 1_A)$ holds (equivalently, the largest $\zeta$ with $[\zeta, 1_A] = 0$). For $\beta$ a congruence, we define the **centralizer** of $\beta$, written $\langle 0 : \beta \rangle$, to be the largest $\alpha$ such that $[\alpha, \beta] = 0$, and more generally for any $\delta$ we define $\langle \delta : \beta \rangle$ to be the largest $\alpha$ such that $C(\alpha, \beta; \delta)$ holds.

**Proposition 10.36.** For congruences on $A$, we have the following:

(a) for any $\beta, \delta$ there exists a largest $\alpha$ such that $C(\alpha, \beta; \delta)$ holds, so $\langle \delta : \beta \rangle$ (and, in particular, the center of $A$) is well-defined,

(b) if $\gamma$ is solvable over $\beta$ and $\beta$ is solvable over $\alpha$, then $\gamma$ is solvable over $\alpha$,

(c) if $\beta$ is solvable (abelian) over $\alpha$, then $\beta \wedge \gamma$ is solvable (abelian) over $\alpha \wedge \gamma$ for any $\gamma$,

(d) if $\theta \leq \alpha \leq \beta$, then $\beta$ is solvable (abelian) over $\alpha$ iff $\beta/\theta$ is solvable (abelian) over $\alpha/\theta$,

(e) $A/\theta$ is solvable (abelian) iff $1_A$ is solvable (abelian) over $\theta$.

Proof. Part (a) follows from Proposition 10.30(e), part (b) is obvious, part(c) follows from Proposition 10.30(d), part (d) follows from Proposition 10.30(f), and part (e) is part (d) specialized to the case $\beta = 1_A, \alpha = \theta$.

To go further, we need to make an additional assumption on our variety, such as congruence modularity. The interested reader can find the (surprisingly deep) theory of commutators in congruence modular varieties in Appendix A.

A weaker assumption which is still good enough to prove most of the basic properties of commutators is the existence of a ternary term known as a **difference term**, generalizing the Gumm difference term found in congruence modular varieties, which acts like a Mal’cev term on abelian algebras.

**Definition 10.37.** A ternary term $p$ is called a **difference term** for a variety, if it satisfies the identity $p(y, y, x) \approx x$, and for every $(x, y) \in \theta$ for $\theta$ a congruence, we always have $p(x, y, y) \equiv_{[\theta, \theta]} x$.

**Example 10.7.** Any SD($\wedge$) variety has a difference term: just take $p(x, y, z) = z$. That this works relies on the fact that $[\alpha, \beta] = \alpha \wedge \beta$ in SD(\wedge) varieties, which we haven’t proved - this can be found in [74].

One property of a difference term is that it forces several alternative commutators to match with the term condition commutator, and one of these commutators is clearly symmetric.
**Definition 10.38.** For any $n \geq 1$, we define the $n$-cycle commutator $[\alpha, \beta]_n$ to be the least congruence $\delta$ such that for any cycle of $n$ matrices

\[
\begin{bmatrix}
  a_1 & b_1 \\
  c_1 & d_1 \\
  a_2 & b_2 \\
  c_2 & d_2 \\
  \vdots & \vdots \\
  a_n & b_n \\
  c_n & d_n 
\end{bmatrix} \in M(\alpha, \beta)
\]

such that $b_i \equiv \delta a_{i+1}$ for all $i < n$, $b_n \equiv \delta a_1$, and $d_i \equiv \delta c_{i+1}$ for all $i < n$, we have additionally that $d_n \equiv \delta c_1$.

If $A$ is affine, then it is easy to check that $[1_A, 1_A]_n = 0_A$ for every $n$. Note that for $n = 1$, we have $[\alpha, \beta]_1 = [\alpha, \beta]$. Additionally, since we can take the $n$th matrix in the cycle to have a pair of equal columns, we have $[\alpha, \beta]_i \leq [\alpha, \beta]_{i+1}$ for all $i$.

Quackenbush’s famous example of an abelian algebra which is not quasi-affine from [99] is an example of an algebra where $[1_A, 1_A]_1 = 0_A$ but $[1_A, 1_A]_2 \neq 0_A$.

For $n = 2$, the 2-cycle commutator is clearly symmetric: $[\alpha, \beta]_2 = [\beta, \alpha]_2$. Since it is defined via two matrices in $M(\alpha, \beta)$, and since each matrix comes from some term, the commutator $[\alpha, \beta]_2$ is also called the two term commutator. The two term condition is illustrated in the following diagram.

If we have a difference term, then all of the $n$-cycle commutators turn out to be equal.

**Theorem 10.39 (Lipparini [85]).** In a variety with a difference term, we have $[\alpha, \beta]_n = [\alpha, \beta]$ for all $n$. In particular, we have $[\alpha, \beta] = [\beta, \alpha]$.

**Proof.** Suppose that $p$ is a difference term. We will show that $[\alpha, \beta]$ satisfies the $n$-cycle term condition by induction on $n$. Suppose that matrices $\begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \in M(\alpha, \beta)$ for $i \leq n$ are as in the definition of the $n$-cycle condition for $\delta = [\alpha, \beta]$. Applying the difference term, we have

\[
p\left(\begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix}, \begin{bmatrix} b_1 & b_1 \\ d_1 & d_1 \end{bmatrix}, \begin{bmatrix} a_1 & a_1 \\ c_1 & c_1 \end{bmatrix}\right) = \begin{bmatrix} p(a_i, b_1, a_1) & p(b_i, b_1, a_1) \\ p(c_i, d_1, c_1) & p(d_i, d_1, c_1) \end{bmatrix} \in M(\alpha, \beta)
\]

for $2 \leq i \leq n - 1$, and

\[
p\left(\begin{bmatrix} a_n & b_n \\ c_n & d_n \end{bmatrix}, \begin{bmatrix} b_1 & a_1 \\ d_1 & c_1 \end{bmatrix}, \begin{bmatrix} a_1 & a_1 \\ c_1 & c_1 \end{bmatrix}\right) = \begin{bmatrix} p(a_n, b_1, a_1) & p(b_n, a_1, a_1) \\ p(c_n, d_1, c_1) & p(d_n, c_1, c_1) \end{bmatrix} \in M(\alpha, \beta).
\]
The reader can check that these form a system of matrices as in the definition of the $n - 1$-cycle condition for $\delta = [\alpha, \beta]$, so by the inductive hypothesis we have

$$p(c_2, d_1, c_1) \equiv [\alpha, \beta] p(d_n, c_1, c_1).$$

From $c_2 \equiv [\alpha, \beta] d_1$ and the fact that $p$ is a difference term, the left hand side is congruent to $c_1$ modulo $[\alpha, \beta]$. From the fact that $c_1 \equiv \delta d_n$ and $(c_1, d_n) \in \alpha \circ [\alpha, \beta] \circ \alpha = \alpha$, we have $(c_1, d_n) \in [\alpha \wedge \beta, \alpha \wedge \beta]$, so from the fact that $p$ is a difference term we have $p(d_n, c_1, c_1) \equiv [\alpha \wedge \beta, \alpha \wedge \beta] d_n$, and from $[\alpha \wedge \beta, \alpha \wedge \beta] \leq [\alpha, \beta]$ we get $c_1 \equiv [\alpha, \beta] d_n$.

In fact, substantially more is true in varieties with a difference term. Kearnes [72] shows that almost all properties of the commutator which hold in congruence modular varieties generalize to varieties with a difference term, other than $[\alpha_1 \lor \alpha_2, \beta] = [\alpha_1, \beta] \lor [\alpha_2, \beta]$. This property must be weakened, but it is at least true that if $[\alpha_1, \beta] = [\alpha_2, \beta]$ then $[\alpha_1 \lor \alpha_2, \beta] = [\alpha_1, \beta]$ in varieties with difference terms.

11 Generalized Majority-Minority operations (motivating Few Subpowers)

The Few Subpowers algorithm was heavily influenced by Dalmau’s paper on generalized majority-minority operations [2]. Dalmau’s motivation was that in both near-unanimity algebras and Mal’cev algebras, every subalgebra of $A^n$ has a nice generating set: in the Mal’cev case, we can use a compact representation, while in the near-unanimity case, if the arity is $l + 1$, we can use any set of elements which has the same projection onto every subset of the coordinates of size at most $l$. The goal was to unify these two cases.

Definition 11.1. An operation $\varphi$ is a generalized majority-minority operation (abbreviated as gmm operation) if for each pair $a, b$ we either have

$$\varphi(x, y, \ldots, y) = \varphi(y, x, \ldots, y) = \cdots = \varphi(y, y, \ldots, x) = y$$

for all $x, y \in \{a, b\}$,

or

$$\varphi(x, y, \ldots, y) = \varphi(y, y, \ldots, x) = x$$

for all $x, y \in \{a, b\}$.

In the second case we say that $a, b$ is a minority pair for $\varphi$.

Definition 11.2. If $R \subseteq A_1 \times \cdots \times A_n$, then we define the signature of $R$, written $\text{Sig}(R)$, to be the set of triples $(i, a, b)$ with $i \in \{1, \ldots, n\}$, $a, b$ a minority pair in $A_i$, such that there are some $t_a, t_b \in R$ with $\pi_{1,\ldots,i-1}(t_a) = \pi_{1,\ldots,i-1}(t_b)$ and $\pi_i(t_a) = a, \pi_i(t_b) = b$. In this case we say that the pair $t_a, t_b$ witnesses the triple $(i, a, b)$.

Theorem 11.3. If $\mathbb{R} \leq A_1 \times \cdots \times A_n$ is preserved by an $l + 1$-ary gmm operation $\varphi$ and $S \subseteq \mathbb{R}$ has $\text{Sig}(S) = \text{Sig}(\mathbb{R})$ and $\pi_I(S) = \pi_I(\mathbb{R})$ for all $I \subseteq \{1, \ldots, n\}$ with $|I| \leq l$, then $\mathbb{R}$ is generated by $S$ (using only $\varphi$).

Proof. We prove this by induction on the arity $n$ of $\mathbb{R}$. Suppose that $a = (a_1, \ldots, a_n) \in \mathbb{R}$, by the induction hypothesis there is some $b_n$ with $(a_1, \ldots, a_{n-1}, b_n)$ in the subalgebra generated by $S$. We have two cases, based on whether $a_n, b_n$ is a majority pair or a minority pair.
Case 1: \(a_n, b_n\) is a majority pair. In this case we show that for every \(I \subseteq \{1, ..., n\}\), we have \(\pi_I a\) in the subalgebra generated by \(\pi_I S\), by induction on \(|I|\). We already know it for \(|I| \leq l\) and for \(n \not\in I\). Suppose \(I = \{i_1, ..., i_m\}\) with \(i_1 < \cdots < i_m = n\) and \(m \geq l + 1\). By the inductive hypothesis, there are elements \(b_1, ..., \) such that 
\[
(b_{i_1}, a_{i_2}, ..., a_{i_m}), (a_{i_1}, b_{i_2}, ..., a_{i_m}), ..., (a_{i_1}, a_{i_2}, ..., b_{i_m}) \in \text{Sg}_\varphi(S).
\]
If some \(b_i = a_i\) then we are done. If some pair \(a_i, b_i\) is minor then - assuming WLOG that \(a_{i_1}, b_{i_1}\) is minority - we have 
\[
\varphi\left(\begin{bmatrix} b_{i_1} & \cdots & b_{i_1} & a_{i_1} \\ a_{i_2} & \cdots & a_{i_2} & a_{i_2} \\ \vdots & \ddots & \vdots & \vdots \\ a_{i_n} & \cdots & a_{i_n} & b_{i_n} \end{bmatrix}\right) = \begin{bmatrix} a_{i_1} \\ a_{i_2} \\ \vdots \\ a_{i_n} \end{bmatrix} \in \text{Sg}_\varphi(\pi_I S),
\]
where all but the last column of the displayed matrix are equal. Otherwise, if all pairs \(a_i, b_i\) are majority, then we have 
\[
\varphi\left(\begin{bmatrix} b_{i_1} & a_{i_1} & \cdots & a_{i_1} \\ a_{i_2} & b_{i_2} & \cdots & a_{i_2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_n} & a_{i_n} & \cdots & b_{i_n} \end{bmatrix}\right) = \begin{bmatrix} a_{i_1} \\ a_{i_2} \\ \vdots \\ a_{i_n} \end{bmatrix} \in \text{Sg}_\varphi(\pi_I S),
\]
where all of the columns of the displayed matrix are distinct, which is possible because \(m \geq l + 1\).

Case 2: \(a_n, b_n\) is a minority pair. In this case, by the assumption \(\text{Sig}(S) = \text{Sig}(\mathbb{R})\), there are \(c, d \in S\) witnessing the triple \((n, a_n, b_n)\). Set \(b = (a_1, ..., a_{n-1}, b_n)\), then we claim that 
\[
a = \varphi(b, b, ..., b, \varphi(b, d, ..., d, c)).
\]
First consider the last coordinate: since \(a_n, b_n\) is a minority pair and \(c_n = a_n, d_n = b_n\), we have 
\[
\varphi(b_1, ..., b_n, \varphi(b_1, d_1, ..., d_n, c_n)) = \varphi(b_n, ..., b_n, \varphi(b_n, ..., b_n, a_n)) = a_n,
\]
so the last coordinates agree. For \(i < n\), we have \(a_i = b_i\) and \(c_i = d_i\), so 
\[
\varphi(b_i, ..., b_i, \varphi(b_i, d_i, ..., d_i, c_i)) = \varphi(a_i, ..., a_i, \varphi(a_i, c_i, ..., c_i, c_i)) = a_i,
\]
where the last equality holds regardless of whether \(a_i, c_i\) is a majority pair or a minority pair. \(\Box\)

Definition 11.4. A subset \(S \subseteq \mathbb{R}\) is called a compact representation of a relation \(\mathbb{R}\) preserved by an \(l + 1\)-ary gmm operation if \(\text{Sig}(S) = \text{Sig}(\mathbb{R})\), \(\pi_I(S) = \pi_I(\mathbb{R})\) for every \(I\) with \(|I| \leq l\), and \(|S| \leq 2|\text{Sig}(\mathbb{R})| + \sum_{|I| \leq l} |\pi_I(\mathbb{R})|\).

In order to manipulate compact representations of relations, we again define subroutines \text{Nonempty}, \text{Fix-values}, \text{Next-beta}, and \text{Intersect}:

- \text{Nonempty}(R, i_1, ..., i_k, S) takes \(R\) a compact representation of \(\mathbb{R} \leq A_1 \times \cdots \times A_n\), \(S \subseteq A_{i_1} \times \cdots \times A_{i_k}\), computes the subalgebra generated by \(\pi_{i_1, ..., i_k}(R)\) under \(\varphi\), and if this intersects with \(S\), then it returns an element of \(\mathbb{R}\) which maps to an element of the intersection,
• **Fix-values**\((R, a_1, \ldots, a_m)\) takes \(R\) a compact representation of \(\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n\) and returns a compact representation of the relation \(x \in R \land (x_1 = a_1) \land \cdots \land (x_m = a_m)\) by inductively fixing one coordinate \(x_i\) to \(a_i\) at a time, and for each new coordinate that is fixed we compute a new compact representation by computing projections onto at most \(l\) coordinates using **Nonempty** and computing witnesses for triples in the signature using the proof of Case 2 of **Theorem 11.3**.

• **Next-beta**\((R, i_1, \ldots, i_k, S)\) takes \(R\) a compact representation of \(\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n\), \(S \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}\), and returns a compact representation of \(\mathbb{R} \cap S\) by computing all projections onto at most \(l\) coordinates using **Nonempty** and computing witnesses for triples in the signature using **Fix-values** and **Nonempty**, and

• **Intersect**\((R, i_1, \ldots, i_k, S)\) takes \(R\) a compact representation of \(\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n\), \(S \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}\), and computes a compact representation for \(\mathbb{R} \cap S\) by first making a compact representation of \(\mathbb{R} \times S\) and then repeatedly calling **Next-beta** to intersect this with the equality relation on the pair of coordinates \(i_j, n + j\).

The only subroutine which has changed substantially from the Mal’cev case is the **Fix-values** subroutine.

### Algorithm 8 Fix-values\((R, a_1, \ldots, a_m), \varphi\) an \(l + 1\)-ary gmm term, \(R\) a compact representation of \(\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n\).

1. Set \(R_0 \leftarrow R\).
2. for \(j\) from 1 to \(m\) do
3.  Let \(R_j \leftarrow \emptyset\).
4.  for all \(I = \{i_1, \ldots\} \subseteq \{1, \ldots, n\}\) with \(|I| \leq l\) and \((b_{i_1}, \ldots) \in \pi_I(R_{j-1})\) do
5.    Set \(R_j \leftarrow R_j \cup \text{Nonempty}(R_{j-1}, j, i_1, \ldots, i_{|I|}, \{(a_j, b_{i_1}, \ldots, b_{i_{|I|}})\})\).
6.  for all \((i, a, b) \in \text{Sig}(R_{j-1})\) with \(i > j\) and \(a, b\) a minority pair do
7.    Let \(t_a, t_b \in R_{j-1}\) witness the triple \((i, a, b)\).
8.    Let \(t \leftarrow \text{Nonempty}(R_{j-1}, j, i, \{(a_j, a)\})\).
9.    if \(t \neq \emptyset\) then
10.   Set \(R_j \leftarrow R_j \cup \{t, \varphi(t, t, \ldots, t, \varphi(t, t_a, \ldots, t_a, t_b))\}\).
11. return \(R_m\).

Reviewing what we’ve done, we have a procedure for converting proofs that compact representations generate relations into algorithms for computing compact representations of intersections for relations. The most critical step of the algorithm is the step of the **Fix-values** subroutine in which we convert a pair that witnesses a triple \((i, a, b)\) in \(R_{j-1}\) to a pair that witnesses a triple \((i, a, b)\) in \(R_j\).

Before we go on, we can use this algorithm to settle the dichotomy conjecture for constraint languages which contain “swap” relations \(\{(a, b), (b, a)\}\) for every pair of elements \(a, b\).

**Theorem 11.5.** Suppose that \(A = (A, \Gamma)\) is a relational structure where \(\Gamma\) is a set of relations which contains the swap relation \(S_{ab} = \{(a, b), (b, a)\}\) for every pair \(a, b \in A\). Then either \(\text{CSP}(\Gamma)\) is \(\text{NP}-\text{complete}\), or \(A\) has a ternary generalized majority-minority polymorphism. In the second case, \(\text{CSP}(\Gamma)\) can be solved in polynomial time by Dalmau’s algorithm.
Proof. Note that $\Gamma$ is automatically core, since any unary polymorphism of $S_{ab}$ must send $a, b$ to distinct values in $\{a, b\}$. Thus if $\text{CSP}(\Gamma)$ is not NP-complete, then it must have a Taylor polymorphism $t$.

First we will show that this implies that for all $a, b \in A$ there is a ternary polymorphism $f_{ab}$ such that the restriction of $f_{ab}$ to $\{a, b\}$ is either the majority operation or the minority operation. Since $\pi_1(S_{ab}) = \{a, b\}$, the set $\{a, b\}$ is closed under $t$. Let $t' \in \text{Clo}(t)$ have minimal arity such that the restriction of $t'$ to $\{a, b\}$ is not a projection. An elementary combinatorial argument known as Świerczkowski’s Lemma [106] shows that if $t'$ has arity at least four, then there is some way of identifying two variables of $t'$ to get a term $t''$ of smaller arity such that the restriction of $t''$ to $\{a, b\}$ is also not a projection. Thus the arity of $t'$ is at most three. The arity of $t'$ can’t be one or two since $t'$ is idempotent and preserves $S_{ab}$.

Since every way of identifying two variables of $t'|_{\{a,b\}}$ gives a projection, up to reordering the variables of $t'$ there are just three cases. In two of these cases, $t'$ already restricts to a majority or minority operation on $\{a, b\}$. In the remaining case, after reordering the variables we may assume that $t'(x, y, z) = t'(y, y, x) = t'(x, y, x) = x$ for $x, y \in \{a, b\}$, and taking $f_{ab}(x, y, z) = t'(x, t'(x, y, z), z)$ gives a function $f_{ab}$ which restricts to a majority operation on $\{a, b\}$.

Now we choose any ordering of the collection of pairs $\{a, b\}$, with the $i$th pair given by $\{a_i, b_i\}$. We inductively define functions $f_i \in \text{Clo}(t)$ by $f_0 = \pi_1$, and for $i \geq 0$ we set

$$f_{i+1}(x, y, z) = f_{a_i b_i}(f_i(x, y, z), f_i(y, z, x), f_i(z, x, y)).$$

We claim that the final function $f_n$ (with $n = \binom{|A|}{2}$) is a generalized majority-minority polymorphism of $A$. Since each $f_{ab}$ is idempotent, it’s enough to check that the restriction of $f_{i+1}$ to $\{a_i, b_i\}$ is either a pure majority or pure minority function.

From the fact that $f_i$ preserves the unary relation $\pi_1(S_{a_i b_i}) = \{a_i, b_i\}$ and the fact that the restriction of $f_{a_i b_i}$ to $\{a_i, b_i\}$ is invariant under cyclically permuting its input variables, we see that $f_{i+1}$ also restricts to a cyclic term on $\{a_i, b_i\}$. Since $f_{i+1}$ preserves $S_{ab}$, it must therefore either restrict to the pure majority or pure minority function on $\{a_i, b_i\}$. \hfill \Box

There are two examples of generalized majority-minority algebras on a three element domain which do not come from majority or Mal’cev operations, and correspond to maximal tractable constraint languages.

Example 11.1. The first example is $A_1 = (\{a, b, c\}, \varphi_1)$, where $\varphi_1$ is a ternary gmm such that $\{a, x\}$ is a pure minority subalgebra of $A_1$ for all $x$, $\{b, c\}$ is a majority subalgebra of $A_1$, and the equivalence relation corresponding to the partition $\{a\}, \{b, c\}$ is a congruence $\alpha$ on $A_1$ such that the quotient $A_1/\alpha$ is a pure minority algebra. Explicitly, $\varphi_1$ is the symmetric idempotent function of its inputs which is given by

$$\varphi_1(a, a, x) = x, \varphi_1(a, x, x) = a, \varphi_1(b, b, c) = b, \varphi_1(b, c, c) = c, \varphi_1(a, b, c) = a.$$  

The corresponding relational clone is generated by the partial order $\{(a, a), (b, b), (b, c), (c, c)\}$, the order two automorphism $\{(a, a), (b, c), (c, b)\}$, and the affine ternary relation $\{(a, a, b), (a, b, a), (b, a, a), (b, b, b)\}$.

Example 11.2. The second example is $A_2 = (\{a, b, c\}, \varphi_2)$, where $\varphi_2$ is a ternary gmm such that $\{a, x\}$ is a majority subalgebra of $A_2$ for all $x$, $\{b, c\}$ is a pure minority subalgebra of $A_2$, the equivalence relation corresponding to the partition $\{a\}, \{b, c\}$ is a congruence $\alpha$ on $A_2$ such that
the quotient \( A_2/\alpha \) is a majority algebra, and the permutation \((b \ c)\) is an automorphism of \( A_2 \). Explicitly, \( \varphi_2 \) is the cyclically symmetric idempotent function of its inputs which is given by

\[
\varphi_2(a, a, x) = a, \varphi_2(a, x, x) = x, \varphi_2(b, b, c) = c, \varphi_2(b, c, c) = b, \varphi_2(a, b, c) = b, \varphi_2(a, c, b) = c.
\]

The corresponding relational clone is generated by the binary relations \{ \((a, b)\), \((b, a)\) \}, \{ \((a, a)\), \((a, b)\), \((b, b)\) \}, \{ \((a, a)\), \((b, b)\), \((b, c)\), \((c, b)\), \((c, c)\) \} and the ternary relation \{ \((a, a, a)\), \((b, b, b)\), \((b, c, c)\), \((c, b, c)\), \((c, c, b)\) \}.

The reader might notice that generalized majority-minority operations are not defined in terms of satisfying a system of identities. So we should be able to immediately generalize Dalmau’s result to the variety of algebras generated by algebras with a gmm operation, by finding the identities which are satisfied by a gmm operation that were critical to the correctness of the algorithm. How did we apply the operation \( \varphi \), throughout the algorithm \textbf{Fix-values} and the proof of Theorem 11.3?

The first thing to note is that we often set almost all of the entries of \( \varphi \) to the same value. So define auxiliary binary and ternary terms \( p, d \) by

\[
d(x, y) = \varphi(x, y, ..., y, y),
\]

\[
p(x, y, z) = d(\varphi(x, y, ..., y, z), z).
\]

The important property of \( d \) is that we have \( d(a, b) = a \) when \( a, b \) are a minority pair. For \( p \), the important property is that when \( a, b \) are a minority pair, then we have \( p(a, b, b) = a \), and in every case we always have

\[
p(y, y, z) = z.
\]

We can express the fact that \( p(a, b, b) = a \) when \( a, b \) are a minority pair by the equation

\[
p(x, y, y) = d(x, y),
\]

which also holds for majority pairs.

Where did we actually use the function \( \varphi \)? It is only called directly in the subroutine \textbf{Nonempty}. It is crucial that it is actually used there, because the full function \( \varphi \) was necessary for Case 1 of Theorem 11.3. The proof of that case does not immediately appear to generalize, as there was substantial casework within it, based on whether there was a minority pair \( a_i, b_i \) or not. However, clever use of the function \( d(x, y) \) can mimic the casework that appeared there. For each \( a_i, b_i \), the expression \( d(a_i, b_i) \) has the nice property that \( a_i, d(a_i, b_i) \) automatically forms a majority pair (or an equal pair, which we can think of as a degenerate case of a majority pair). So if we define a function \( s(x_0, x_1, ..., x_l) \) by

\[
s(x_0, x_1, ..., x_l) = \varphi(x_0, d(x_0, x_1), ..., d(x_0, x_l)),
\]

then we find that

\[
s(y, x, x, ..., x) = \varphi(y, d(y, x), ..., d(y, x)) = d(y, x),
\]

\[
s(x, y, x, ..., x) = \varphi(x, d(x, y), x, ..., x) = x,
\]

\[
s(x, x, y, ..., x) = \varphi(x, x, d(x, y), ..., x) = x,
\]

\[
\vdots
\]

\[
s(x, x, x, ..., y) = \varphi(x, x, x, ..., d(x, y)) = x.
\]
This function $s$ lets us generalize Case 1 of Theorem 11.3, the case where $d(a_n, b_n) = b_n$, while the function $p$ was necessary to generalize Case 2. To unify them, we should slightly modify our construction of $s$ to create the following term $e$:

$$e(u, v, x_1, ..., x_l) = \varphi(v, d(u, x_1), ..., d(u, x_{l-1}), d(x_1, x_l)).$$

Then $s$ is related to $e$ by

$$s(x_0, x_1, ..., x_l) = e(x_1, x_0, x_1, ..., x_l)$$

if all but one of the $x_i$ are equal.

$p$ is related to $e$ by

$$p(x, y, z) = e(y, x, z, ..., z)$$

if $x = y$ or $y = z$,

and $e$ satisfies the identities

$$e(y, y, x, x, ..., x) = \varphi(y, d(y, x), ..., d(y, x), x) = x,$n
$$e(y, x, y, x, ..., x) = \varphi(x, y, d(y, x), ..., d(y, x)) = x,$n
$$e(x, x, y, x, ..., x) = \varphi(x, x, d(x, y), ..., d(x, y)) = x,$n

$$...$$

$$e(x, x, x, x, ..., y) = \varphi(x, x, x, ..., d(x, y)) = x.$$

Can we use this system of identities to prove an analogue of Theorem 11.3? Yes! The trick is to plug things back into $e$, to make the following term $t$:

$$t(u, v, w, x_1, ..., x_l) = e(p(v, u, x_1), s(w, x_1, ..., x_l), x_1, ..., x_l).$$

Now if we have a tuple $a = (a_1, ..., a_n)$ which we want to prove is in the subalgebra generated by $S$, and if this subalgebra already contains $\langle a_1, ..., a_{i-1}, b_i, a_{i+1}, ..., a_n \rangle$ for each $i$, as well as a pair $\langle c_1, ..., c_{n-1}, a_n \rangle, \langle c_1, ..., c_{n-1}, b_n \rangle$ which witnesses the triple $(n, a_n, b_n)$, then we have

$$t\left(\begin{bmatrix} c_1 & c_1 & a_1 & b_1 & a_1 & \cdots \\ c_2 & c_2 & a_2 & b_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ a_n & b_n & b_n & a_n & a_n & \cdots \end{bmatrix}\right) = e\left(\begin{bmatrix} b_1 & a_1 & b_1 & a_1 & \cdots \\ a_2 & a_2 & a_2 & b_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ d_n & d_n & a_n & a_n & a_n & \cdots \end{bmatrix}\right) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix},$$

where $d_n = d(b_n, a_n)$.

While playing these sorts of games with identities may yield more and more general examples of algebraic structures where relations have compact representations, we are not being very systematic here. So perhaps we should work backwards: what absolutely needs to be true for something like compact representations to exist?

**Proposition 11.6.** If every subpower $R \leq A^n$ has a compact representation $S$ consisting of at most $p(n)$ tuples, then the number of different subalgebras of $A^n$ is at most $|A^n|^{p(n)} = |A|^{n p(n)}$.

**Corollary 11.7.** No analogue of compact representations can exist for subpowers of a nontrivial semilattice.
Proof. It’s enough to consider the case $A = (\{0,1\}, \text{max})$, since every semilattice contains a subalgebra isomorphic to it. The number of subpowers of $A^n$ is at least the number of subsets on $\{0,1\}^n$ which are generated by subsets $S \subseteq \{x \in \{0,1\}^n, \sum_i x_i = n/2\}$ (suppose $n$ is even). Any two distinct subsets $S, S'$ of the set of tuples with weight $n/2$ will generate different subalgebras of $A^n$, so the number of subalgebras of $A^n$ is at least

$$2^{\binom{n}{n/2}} \geq 2^{2n/n},$$

and $2^{n/n}$ clearly grows faster than any polynomial.

What makes the semilattice case so different from the Mal’cev case and the near-unanimity case? The main difference is that the identities satisfied by a semilattice do not allow us to get back to $x$ once we start combining it with other values, while the identities for Mal’cev and near-unanimity terms all have $x$s on the right hand sides.

So we should start by trying to prove that having few subpowers implies that there are terms satisfying a nontrivial system of identities which have $x$s on the right hand sides of each identity, such as the system of identities satisfied by the term $e$ constructed earlier. The trick, as we will see, is to apply the existence of compact representations to the case of a power of the free algebra on two generators, considered as a subalgebra of $(A^{A^2})^n$.

### 12 Algebras with Few Subpowers

First we define an invariant of an algebraic structure and the variety it generates, which is slightly more well-behaved than the function that takes $n$ to the number of subalgebras of $A^n$.

**Definition 12.1.** If $A$ is an algebraic structure and $a_1, \ldots, a_k \in A$, we say that $a_1, \ldots, a_k$ are independent if no $a_i$ is in the subalgebra generated by the rest of the $a_j$s. For every $n$, we define $i_A(n)$ to be the size of the largest independent set in $A^n$.

**Proposition 12.2.** If $A$ is a finite algebra, then any subalgebra of $A^n$ can be generated by at most $i_A(n)$ elements, so the number of subalgebras of $A^n$ is bounded above by $|A^n|^{i_A(n)} = 2^{n \log(|A|) i_A(n)}$. The number of subalgebras of $A^n$ is also bounded below by $2^{i_A(n)}$.

**Proof.** Since $A$ is finite, every subalgebra of $A^n$ has a minimal generating set, and this minimal generating set is necessarily independent. The upper bound on the number of subalgebras follows from counting the number of possible minimal generating sets.

For the lower bound on the number of subalgebras, suppose that $a_1, \ldots, a_k$ are independent in $A^n$. Then every subset $S$ of $\{a_1, \ldots, a_k\}$ generates a distinct subalgebra of $A^n$, since $Sg_{A^n}(S) \cap \{a_1, \ldots, a_k\} = S$ by the definition of independence. Thus $A^n$ has at least $2^k$ distinct subalgebras.

**Proposition 12.3.** If $B \in HSP(A)$ is also finite, then $i_B(n) \leq i_A(cn)$ for some constant $c$ depending only on $B$.

**Proof.** If $A, B$ are both finite, then there is some finite number $c$ such that $B \in HS(A^c)$, that is, there is a subalgebra $C \leq A^c$ and a surjective homomorphism $f : C \to B$. Then every independent set in $B^n$ lifts to an independent set in $(A^c)^n = A^{cn}$ by choosing any section of $f$ and applying it coordinate-wise.
We will apply the above result to the free algebra on two generators $F(\mathbb{A})(x,y) \leq \mathbb{A}^2$ to prove that if an algebra has few subpowers, then it has a cube term. Since cube terms have exponentially high arity, it’s necessary to develop some notation to define them properly.

**Definition 12.4.** For every subset $S \subseteq \{1,\ldots,k\}$, we define the $k$-dimensional column vector $v^S$ by

$$
v^S_i = \begin{cases} y & i \in S, \\ x & i \notin S. \end{cases}
$$

A $k$-cube term is a term $t$ with variables indexed by nonempty subsets of $\{1,\ldots,k\}$, such that if we fix an enumeration $S_1,\ldots,S_{2^k-1}$ of these subsets, we have the identity

$$t(v^{S_1},\ldots,v^{S_{2^k-1}}) \approx v^\emptyset.$$

For instance, if $k = 3$ then (with one possible choice of variable ordering) a 3-cube term is a 7-ary term $t$ satisfying the identity

$$t\left(\begin{bmatrix} y & y & x & y & x & x & x \\ y & y & x & y & x & x & y \\ y & x & y & x & x & y \end{bmatrix}\right) \approx \begin{bmatrix} x \\ x \\ x \end{bmatrix}.$$

Note that a Mal’cev term is the same as a 2-cube term (up to reordering variables).

**Theorem 12.5 (Few subpowers implies cube term [20]).** Let $F = F(\mathbb{A})(x,y) \leq \mathbb{A}^2$ be the free algebra on two generators in the variety generated by $\mathbb{A}$.

- If $i_\mathbb{A}(k) < 2^k$ for any $k$, then $\mathbb{A}$ has a $k$-cube term.
- If $i_\mathbb{A}(m) < \binom{n}{k}$ for any $m,k$, then $\mathbb{A}$ has a $k$-cube term.

In particular, if $i_\mathbb{A}(n) = o(n^k)$ then $\mathbb{A}$ has a $k$-cube term, and if $i_\mathbb{A}(n) = 2^{o(n)}$ then there exists some $k$ such that $\mathbb{A}$ has a $k$-cube term.

**Proof.** For the first statement, if $i_\mathbb{A}(k) < 2^k$, then the vectors $v^S$ for $S \subseteq \{1,\ldots,k\}$ can’t be independent, so some $v^S$ is in the subalgebra generated by the others. By applying an automorphism of $\mathbb{F}^k$ which swaps $xs$ and $ys$ in the coordinates belonging to $S$, we may assume without loss of generality that $S = \emptyset$. From $v^\emptyset \in S_{\mathbb{A}}\{v^S \mid S \neq \emptyset\}$, we see that there is a term $t$ such that $t(v^{S_1},\ldots) = v^\emptyset$, and since $F$ is the free algebra on two generators, this implies the $k$-cube term identities.

For the second statement, consider the set of vectors $v^S$ with $S \in \binom{\{1,\ldots,m\}}{k}$. By assumption, these are not independent, so some $v^S$ is in the subalgebra generated by the others. Then if we project onto the coordinates of $S$ and use the fact that for $S \neq T$ with $|S| = |T|$ we never have $S \subseteq T$, we get the situation of the previous paragraph inside $F^S \cong \mathbb{F}^k$. 

Next, we upgrade the $k$-cube term by repeatedly plugging it into itself to produce simpler terms, finally arriving at the $k$-edge term.

**Definition 12.6.** If $\Delta \subseteq \mathcal{P}(\{1,\ldots,k\}) \setminus \{\emptyset\}$, then we say that $t$ is a $\Delta$-cube term if it has variables indexed by elements of $\Delta$ and satisfies the identity $t(v^{S_1},\ldots) = v^\emptyset$, where $S_1,\ldots$ is an enumeration of the elements of $\Delta$.

If we set $\Delta^e = \{\{1,2\},\{1\},\{2\},\ldots,\{k\}\}$, then a $\Delta^e$-cube term is called a $k$-edge term.
A $k$-edge term is simple enough that we can write out the identities it satisfies explicitly: a $k+1$-ary term $e$ is a $k$-edge term iff it satisfies
\[
e \left( \begin{array}{cccccc}
y & y & x & x & \cdots & x \\
y & x & y & x & \cdots & x \\
x & x & x & y & \cdots & x \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x & x & x & x & \cdots & y \end{array} \right) \approx \left( \begin{array}{c}
x \\
x \\
x \\
\vdots \\
x \end{array} \right).
\]

**Theorem 12.7** (Cube term implies edge term [20]). If $A$ has a $k$-cube term, then it also has a $k$-edge term.

**Proof.** Since it is hard to deal with terms having exponentially many variables, we will do the last step of the proof first, and show that if $A$ has a $\Delta^*$-cube term $t^*$ then it has a $k$-edge term, where $\Delta^* = \{\{1,2\}, \ldots, \{1,k\}, \{1\}, \{2\}, \ldots, \{k\}\}$ only has $2k-1$ elements. The $\Delta^*$-cube term identities for $t^*$ state that
\[
t^* \left( \begin{array}{ccccccc}
y & y & \cdots & y & x & x & \cdots & x \\
y & x & \cdots & x & y & x & \cdots & x \\
x & y & \cdots & x & x & y & \cdots & x \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\
x & x & \cdots & y & x & x & \cdots & y \end{array} \right) \approx \left( \begin{array}{c}
x \\
x \\
x \\
\vdots \\
x \end{array} \right).
\]

In order to show that there is a $k$-edge term, we just need to show that $v^\emptyset$ can be generated from $\{v^S | S \in \Delta^e\}$ using the $\Delta^*$-cube term $t^*$.

Let $a = t^*(x, \ldots, x, y, x, \ldots, x)$, where the only $y$ occurs at the index corresponding to $\{1\}$ (this is the middle index if we order the variables of $t$ as in the displayed identities above). First we will use $t$ to generate vectors $v^{S,a}$ for $S \in \Delta^*$ which look just like the vectors $v^S$, except $y$s in the first coordinate are replaced by $a$s. If $S \in \Delta^*$ and $1 \notin S$, then $S$ is already in $\Delta^e$ and $v^{S,a} = v^S$, so we don’t have to worry about these. If $S = \{1\}$, then we use
\[
t^* \left( \begin{array}{ccccccc}
x & x & \cdots & x & y & x & x & \cdots & x \\
y & x & \cdots & x & y & x & x & \cdots & x \\
x & y & \cdots & x & x & y & x & \cdots & x \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
x & x & \cdots & y & x & x & x & \cdots & y \end{array} \right) = \left( \begin{array}{c}
a \\
x \\
\vdots \\
x \end{array} \right),
\]

and note that every column of the matrix on the left hand side is $v^S$ for some $S \in \Delta^e$. If $S = \{1, 2\}$, then we use
\[
t^* \left( \begin{array}{ccccccc}
x & x & \cdots & x & y & x & x & \cdots & x \\
y & y & \cdots & y & y & y & y & \cdots & y \\
x & x & \cdots & x & x & x & x & \cdots & x \\
\vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
x & x & \cdots & x & x & x & x & \cdots & x \end{array} \right) = \left( \begin{array}{c}
a \\
y \\
\vdots \\
x \end{array} \right),
\]

85
again noting that every column corresponds to an element of $\Delta^f$. Finally, if $S = \{1, i\}$, say $S = \{1, 3\}$ without loss of generality, then we use

$$
t^* \begin{pmatrix}
x & x & \cdots & x & y & x & x & \cdots & x \\
y & y & \cdots & y & y & x & x & \cdots & x \\
x & x & \cdots & x & x & y & y & \cdots & y \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x & x & \cdots & x & x & x & x & \cdots & x
\end{pmatrix} = \begin{pmatrix} a \\
x \\
y \\
\vdots \\
x
\end{pmatrix},
$$

where every row other than the first three (or other than the first, second, and $i$th in the general case) is all $x$s, and again every column belongs to $\Delta^f$.

Now that we’ve constructed the $v^{S,a}$s for all $S \in \Delta^*$, we use $t^*$ to put them all together:

$$
t^* \begin{pmatrix}
a & a & \cdots & a & a & x & x & \cdots & x \\
y & x & \cdots & x & x & y & x & \cdots & x \\
x & y & \cdots & x & x & y & x & \cdots & x \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x & x & \cdots & y & x & x & x & \cdots & y
\end{pmatrix} = \begin{pmatrix} x \\
x \\
x \\
\vdots \\
x
\end{pmatrix}.
$$

Thus if $A$ has a $\Delta^*$-cube term, then it has a $k$-edge term. Explicitly, the construction we just worked through corresponds to the formula

$$e(x_0, x_1, \ldots, x_k) = t^* (t^* (x_2, \ldots, x_2, x_0, x_2, \ldots, x_2), t^* (x_2, \ldots, x_2, x_0, x_3, \ldots, x_3), \ldots, t^* (x_2, \ldots, x_2, x_k, x_2, \ldots, x_k), t^* (x_2, \ldots, x_k, x_1, x_2, \ldots, x_k), x_2, \ldots, x_k).$$

Now that we have the general idea down, we work through the inductive argument needed to prove that if we have a $k$-cube term, then we have a $\Delta^*$-cube term. Let $\Delta^{\ell^*} = \Delta^* \cup P(\{1, \ldots, \ell\}) \setminus \emptyset$. Note that a $k$-cube term is the same as a $\Delta^{k^*}$-cube term, and a $\Delta^*$-cube term is the same as a $\Delta^{0^*}$-cube term.

**Claim:** If $A$ has a $\Delta^{\ell^*}$-cube term $t^\ell$, then it also has a $\Delta^{(\ell-1)^*}$-cube term.

**Proof of Claim:** We argue as before, this time taking $a = t^\ell (x, \ldots, x, y, x, \ldots, x)$, where the lone $y$ occurs in the index corresponding to $\{\ell\}$. For $S \in \Delta^{\ell^*}$, we let $v^{S,a}$ be the vector similar to $v^S$, but with any $y$ in the $\ell$th coordinate replaced with an $a$. We just need to generate each $v^{S,a}$ for $S \in \Delta^{\ell^*}$ using the vectors coming from $\Delta^{(\ell-1)^*}$. Again, if $\ell \notin S$ then $v^{S,a} = v^S$ and $S \in \Delta^{(\ell-1)^*}$ already.

If $S = \{\ell\}$, then we plug in the matrix $M$ to $t^\ell$ which looks just like the matrix which gives the defining identities for $t^\ell$, but has the $\ell$th row replaced by the sequence of $x$s and $y$s we used to define $a$. Explicitly, $M$ is given by

$$
\begin{array}{c|cc}
M_{i,T} & T \neq \{\ell\} & T = \{\ell\} \\
i \neq \ell & v^T_i & x \\
i = \ell & x & y.
\end{array}
$$

Then $t^\ell (M) = v^{(\ell,a)}$, and the $T$th column of $M$ is $v^{T \setminus \{\ell\}}$ if $T \neq \{\ell\}$ and is $v^{(\ell)}$ if $T = \{\ell\}$.
If $\ell \in S$ but $S \neq \{\ell\}$, then we plug in a matrix $M^S$ such that each of its columns is equal to one of $v^{S\setminus\{\ell\}}, v^{\{1\}}, v^{\{1,\ell\}}$: if $\ell \notin T$, then the $T$th column of $M^S$ is $v^{S\setminus\{\ell\}}$, if $\ell \in T$ but $T \neq \{\ell\}$ then the $T$th column is $v^{\{1\}}$, and if $T = \{\ell\}$ then the $T$th column is $v^{\{1,\ell\}}$. Explicitly, $M^S$ is given by

$$
\begin{array}{c|ccc}
  M^S_{i,T} & \ell \notin T & \ell \in T \neq \{\ell\} & T = \{\ell\} \\
  \hline
  i = 1 \in S & y & y & y \\
  i = 1 \notin S & x & y & y \\
  i \neq 1, \ell, i \notin S & y & x & x \\
  i \neq 1, \ell, i \in S & x & x & x \\
  i = \ell & x & x & y.
\end{array}
$$

These choices ensure that $t^\ell(M^S) = v^{S,a}$.

To finish, we apply $t^\ell$ to the set of vectors $v^{S,a}$ for $S \in \Delta^\ell$, and see that the defining identities for $t^\ell$ imply that the resulting vector is $v^0$. Thus there is a $\Delta^{(\ell-1)*}$-cube term $t^{\ell-1}$ which can in principle be written explicitly by plugging in variables to the star composition $t^\ell * t^\ell$.

From a $k$-edge term $e$, we can now construct terms $s, p$ that act like near-unanimity and Mal’cev terms which have been “glued together” by a binary term $d$. I’ve rearranged the variables of these terms from the notation used in [20], for the sake of readability and for consistency with the notation used in Appendix A.

**Theorem 12.8** (Edge terms imply terms $s, p, d$ [20]). If $e$ is a $k$-edge term on a finite algebra $A$, then there are terms $s, p, d \in \text{Clo}(e)$ with $s$ $k$-ary which satisfy the system of identities

\[
\begin{align*}
  s(y, x, x, ..., x) &\approx d(y, x), \\
  s(x, y, x, ..., x) &\approx x, \\
  \vdots \\
  s(x, x, x, ..., y) &\approx x, \\
  p(y, y, x) &\approx x, \\
  p(x, y, y) &\approx d(x, y), \\
  d(d(x, y), y) &\approx d(x, y).
\end{align*}
\]

Furthermore, these terms can be computed from $e$ in time $O(|A|^k)$. If $A$ is infinite, then we can find terms $s, p, d \in \text{Clo}(e)$ satisfying all but the last displayed identity.

**Proof.** If we ignore the last identity involving $d$, we can find terms $s_1, p_1, d_1$ satisfying the other identities as follows:

\[
\begin{align*}
  s_1(x_1, x_2, ..., x_k) &= e(x_2, x_1, x_2, ..., x_k), \\
  p_1(x, y, z) &= e(y, x, z, ..., z), \\
  d_1(x, y) &= e(y, x, y, ..., y).
\end{align*}
\]

We can get the last identity by an iteration argument. For each $i$, we set

\[
\begin{align*}
  s_{i+1}(x_1, x_2, ..., x_k) &= s_1(s_i(x_1, x_2, ..., x_k), x_2, ..., x_k), \\
  p_{i+1}(x, y, z) &= p_1(d_i(x, y), y, z), \\
  d_{i+1}(x, y) &= d_1(d_i(x, y), y).
\end{align*}
\]

87
Then for each $i$, the terms $s_i, p_i, d_i$ satisfy the desired identities aside from the last one. Since $\mathbb{A}$ is finite, we can take $i = |\mathbb{A}|!$ to find that

$$d_{|\mathbb{A}|!}(d_{|\mathbb{A}|!}(x, y), y) = d_{|\mathbb{A}|!}(x, y)$$

for all $x, y \in \mathbb{A}$.

To compute $s_{|\mathbb{A}|!}$ efficiently from $e$, first we compute $s_1$, and then for each choice of $a_2, \ldots, a_k \in \mathbb{A}$ we find the induced unary polynomial $f_{a_2, \ldots, a_k} : x_1 \mapsto s_1(x_1, a_2, \ldots, a_k)$. To finish, we note that for every unary function $f : \mathbb{A} \to \mathbb{A}$ we can compute $f^\infty := \lim_{n \to \infty} f^{\circ n!}$ in time $O(|\mathbb{A}|)$ using a clever algorithm which we leave as an exercise to the reader. \hfill $\square$

**Definition 12.9.** If $s, p, d$ are terms as in the Theorem 12.8, then we say that $a, b \in \mathbb{A}$ are a **minority pair** if $d(b, a) = b$. If $R \subseteq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, then we say that $(i, a, b)$ is a **minority index** of $R$ which is **witnessed** by a pair $t_a, t_b \in R$ if:

1. $a, b$ are a minority pair, i.e. $d(b, a) = b$,
2. the pair $t_a, t_b$ agree up to coordinate $i$: $\pi_{1, \ldots, i-1}(t_a) = \pi_{1, \ldots, i-1}(t_b)$, and
3. we have $\pi_i(t_a) = a, \pi_i(t_b) = b$.

We define the **signature** of $R$, written $\text{Sig}(R)$, to be the set of minority indices which are witnessed by pairs in $R$.

**Definition 12.10.** If $\mathbb{R} \subseteq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ and the $\mathbb{A}_i$ are in a variety with a $k$-edge term, then we say that a set $S \subseteq \mathbb{R}$ is a **compact representation** of $\mathbb{R}$ if:

1. $\text{Sig}(S) = \text{Sig}(\mathbb{R})$,
2. for every $I \subseteq \{1, \ldots, n\}$ with $|I| \leq k - 1$ we have $\pi_I(S) = \pi_I(\mathbb{R})$, and
3. $|S| \leq 2|\text{Sig}(\mathbb{R})| + \sum_{I \subseteq \{1, \ldots, n\}; |I| \leq k-1} |\pi_I(\mathbb{R})|$.

**Theorem 12.11** (Subpowers with edge terms are generated by compact representations [20]). If $\mathbb{R} \subseteq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ and the $\mathbb{A}_i$ are finite algebras in a variety with a $k$-edge term $e$, then for any compact representation $S$ of $\mathbb{R}$, we have $\mathbb{R} = \text{Sg}_e(S)$.

**Proof.** Let $s, p, d$ be terms as in Theorem 12.8. We induct on $n$. Suppose $a = (a_1, \ldots, a_n) \in \mathbb{R}$, then by the induction hypothesis there is $b_n \in \mathbb{A}_n$ with $(a_1, \ldots, a_{n-1}, b_n) \in \text{Sg}_e(S)$. Then if we let $d_n = d(b_n, a_n)$ then we see that $a_n, d_n$ is a minority pair and $(a_1, \ldots, a_n, d_n) \in \mathbb{R}$, so $(n, a_n, d_n) \in \text{Sig}(\mathbb{R})$, from the definition of a compact representation we see that there must be some $c_1, \ldots, c_{n-1}$ such that

$$(c_1, \ldots, c_{n-1}, a_n), (c_1, \ldots, c_{n-1}, d_n) \in S.$$ 

We show by an inner induction on subsets $I \subseteq \{1, \ldots, n\}$ that for each $I$, we have $\pi_I(a) \in \pi_I(\text{Sg}_e(S))$. If $|I| \leq k - 1$ this follows from the definition of a compact representation, while if $n \not\in I$ then this follows from the outer inductive hypothesis. For the sake of notation we will assume that $I = \{1, \ldots, n\}$. Then by the inductive hypothesis, there are $b_1, \ldots, b_{n-1}$ such that for each $i$, we have

$$(a_1, \ldots, a_{i-1}, b_i, a_{i+1}, \ldots, a_n) \in \text{Sg}_e(S).$$
Then we have
\[
\begin{bmatrix}
a_1 & b_1 & a_1 & \cdots \\
a_2 & a_2 & b_2 & \cdots \\
\vdots & \vdots & \vdots & \ddots \\
b_n & a_n & a_n & \cdots \\
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
d_n \\
\end{bmatrix}
= \begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
d_n \\
\end{bmatrix}
\in S_{ge}(S).
\]
Additionally, we have
\[
\begin{bmatrix}
c_1 & c_1 & b_1 \\
c_2 & c_2 & a_2 \\
\vdots & \vdots & \vdots \\
d_n & a_n & a_n \\
\end{bmatrix}
\begin{bmatrix}
b_1 \\
a_2 \\
\vdots \\
d_n \\
\end{bmatrix}
= \begin{bmatrix}
b_1 \\
a_2 \\
\vdots \\
d_n \\
\end{bmatrix}
\in S_{ge}(S).
\]
Now we can apply the \(k\)-edge term \(e\) to see that
\[
\begin{bmatrix}
b_1 & a_1 & b_1 & a_1 & \cdots \\
a_2 & a_2 & b_2 & a_2 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots \\
d_n & d_n & a_n & a_n & \cdots \\
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
d_n \\
\end{bmatrix}
= \begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
d_n \\
\end{bmatrix}
\in S_{ge}(S).
\]

\[\Box\]

**Corollary 12.12.** For a fixed finite algebra \(\mathbb{A}\):

- \(\mathbb{A}\) has a \(k\)-edge term but no \(k-1\)-edge term iff \(i_{\mathbb{A}}(n) = \Theta(n^{k-1})\), and
- \(\mathbb{A}\) has no \(k\)-edge term for any \(k\) iff \(i_{\mathbb{A}}(n) = 2^{\Theta(n)}\).

**Proof.** We only need to check that if \(\mathbb{A}\) has a \(k\)-edge term, then \(i_{\mathbb{A}}(n) = O(n^{k-1})\). Suppose that \(a_1,\ldots,a_m \in \mathbb{A}^n\) are independent, and consider the relations \(R_i = Sg_{\mathbb{A}^n}\{a_1,\ldots,a_i\}\). We can easily find a sequence of compact representations \(S_1,\ldots,S_m\) of \(R_1,\ldots,R_m\) with \(S_i \subseteq S_{i+1}\) for each \(i\). From the independence of the \(a_i\)s, we have \(R_i \neq R_{i+1}\) for all \(i\), so by induction we see that \(|S_i| \geq i\) for all \(i\). Then from the fact that \(S_m\) is a compact representation, we have

\[
m \leq |S_m| \leq 2n|A|^2 + \sum_{I \subseteq \{1,\ldots,n\}, |I| \leq k-1} |A|^{k-1} = O(n^{k-1}).
\]

We can now generalize Dalmau’s generalized majority-minority algorithm to an algorithm for computing compact representations of intersections of two relations which are both described by compact representations. The only changes we need to make are to use the edge term \(e\) in the Nonempty subroutine in the place of the gmm term \(\varphi\), and to modify the Fix-values subroutine to use the ternary term \(p\) from Theorem 12.8.

That the modified Fix-values subroutine works follows from the following Proposition.

**Proposition 12.13.** If the pair of tuples \(t_a, t_b\) witness the minority index \((i,a,b)\), then for any \(t\) with \(\pi_i(t) = a\) the pair of tuples \(t,p(t_b,t_a,t)\) also witnesses the minority index \((i,a,b)\).

**Proof.** From the identity \(p(y,y,x) \approx x\) we have
\[
\pi_{<i}(p(t_b,t_a,t)) = \pi_{<i}(p(t_a,t_a,t)) = \pi_{<i}(t),
\]
and since \((a,b)\) is a minority pair, we have
\[
\pi_i(p(t_b,t_a,t)) = p(b,a,a) = d(b,a) = b.
\]
\[\Box\]
Algorithm 9 Fix-values\((R, a_1, ..., a_m), p, d\) terms as in Theorem 12.8. \(R\) a compact representation of \(\mathbb{R} \leq A_1 \times \cdots \times A_n\).

1: Set \(R_0 \leftarrow R\).
2: for \(j\) from 1 to \(m\) do
3: Let \(R_j \leftarrow \emptyset\).
4: for all \(I = \{i_1, \ldots\} \subseteq \{1, \ldots, n\}\) with \(|I| < k\) and \((b_{i_1}, \ldots) \in \pi_I(R_{j-1})\) do
5: Set \(R_j \leftarrow R_j \cup \text{Nonempty}(R_{j-1}, j, i_1, \ldots, i_{|I|}, \{(a_j, b_{i_1}, \ldots, b_{i_{|I|}})\})\).
6: for all \((i, a, b) \in \text{Sig}(R_{j-1})\) with \(i > j\) and \(a, b\) a minority pair (i.e. \(d(b, a) = b\)) do
7: Let \(t_a, t_b \in R_{j-1}\) witness the triple \((i, a, b)\).
8: Let \(t \leftarrow \text{Nonempty}(R_{j-1}, j, i, \{(a_j, a)\})\).
9: if \(t \neq \emptyset\) then
10: Set \(R_j \leftarrow R_j \cup \{t, p(t_b, t_a, t)\}\).
11: return \(R_m\).

Example 12.1. There is an example of an algebra \(A = \{a, b, c\}, g\) with \(g\) a ternary operation such that \(A\) has a 3-edge term, but is not in the variety generated by generalized majority-minority algebras of any arity. The ternary operation \(g\) is the idempotent symmetric function given by

\[
g(a, b, b) = b, g(a, a, b) = a, g(a, c, c) = a, g(a, a, c) = c, g(b, c, c) = a, g(b, b, c) = c, g(a, b, c) = c.
\]

You can understand this as follows: the subset \(\{a, b\}\) is a majority subalgebra, the subset \(\{a, c\}\) is a pure minority subalgebra, and there is a congruence with equivalence classes \(\{a, b\}, \{c\}\) so that the quotient is a pure minority algebra. Also, the only way to get \(b\) out of an application of \(g\) is if at least two of the inputs are \(b\) (this property is called “absorption”: the subalgebra \(\{a, c\}\) absorbs \(\{a, b, c\}\) with respect to \(g\)).

To see that this isn’t in the variety generated by generalized majority-minority algebras, recall that in any gmm algebra there are functions \(s, p, d\) as in Theorem 12.8 where \(d\) satisfies the additional identity \(d(x, d(y, x)) \approx x\) since \(d\) either acts as first or second projection for any particular pair \(x, y\). Since the quotient corresponding to \(\{a, b\}, \{c\}\) is a pure minority algebra, we must have \(d(c, b) = c\), so by the extra identity we have \(d(b, c) = d(b, d(c, b)) = b\). Then the function \(p\) would satisfy

\[
p \left( \begin{bmatrix} b & c & c \\ c & c & b \end{bmatrix} \right) = \begin{bmatrix} d(b, c) \\ b \end{bmatrix} \neq \begin{bmatrix} b \\ b \end{bmatrix}.
\]

But this is impossible: the subalgebra of \(A^2\) generated by \((b, c), (c, c), (c, b)\) doesn’t contain \((b, b)\), because of the absorption property of \(\{a, c\}\) with respect to \(g\).

To see that \(A\) has a 3-edge term, we define an auxiliary 4-ary term \(f\) by

\[
f(u, x, y, z) = g(g(u, x, z), g(u, y, z), g(u, z, z)),
\]

and then define our 3-edge term by

\[
e(u, x, y, z) = g(g(f(u, x, y, z), x, x), g(f(u, x, y, z), y, y), g(f(u, x, y, z), z, z)).
\]

If we define functions \(s, p, d\) from the 3-edge term \(e\) as in Theorem 12.8 then \(d\) is given by
and the minority pairs are \((a,c),(c,a),(b,c)\). The fact that \((c,b)\) is not a minority pair is witnessed by the fact that the relation \(\text{Sg}_{A^2}\{(b,c),(c,c),(c,b)\}\) contains \((b,c)\) but does not contain \((b,b)\), even though it has \((2,b,c)\) in its signature.

The associated relational clone is generated by the order two automorphism \(\{(a,b),(b,a)\}\) of \(\{a,b\}\), the partial order \(\{(a,a),(a,b),(b,b),(c,c)\}\), the binary relation \(\{(a,a),(a,b),(a,c),(b,a),(b,c)\}\) which witnesses the fact that \(\{a,c\}\) is a “central” subalgebra in Zhuk’s terminology [112] (which is closely related to \(\{a,c\}\) being a ternary absorbing subalgebra), and the affine ternary relation \(\{(a,a,c),(a,c,a),(c,a,a),(c,c,c)\}\).

12.1 Some connections with congruence modularity

**Theorem 12.14.** If an algebra has an edge term, then it generates a congruence modular variety.

**Proof.** By Theorem A.50 from Appendix A we just need to check that an algebra with an edge term has directed Gumm terms, that is, terms \(f_1,\ldots,f_k,p\) satisfying the system of identities

\[
\begin{align*}
    f_1(x,x,y) &\approx x, \\
    f_i(x,y,x) &\approx x \text{ for all } i, \\
    f_i(x,y,y) &\approx f_{i+1}(x,x,y) \text{ for all } i, \\
    f_k(x,y,y) &\approx p(x,y,y), \\
    p(x,x,y) &\approx y.
\end{align*}
\]

If the reader wants to understand why this system of identities implies congruence modularity *without* reading all of Appendix A then they can take the following path: first, read the discussion before Theorem A.50 to see why the existence of directed Gumm terms implies the existence of Gumm terms, then read part of the proof of Theorem A.49 to see how to construct Day terms from Gumm terms, and finally, read Section A.1 of Appendix A to see why the existence of Day terms is equivalent to congruence modularity.

Suppose that \(e\) is a \(k\)-edge term. Define terms \(f_i(x,y,z)\) for \(i < k\) by

\[
    f_i(x,y,z) = e(x,\ldots,x,y,z,\ldots,z),
\]

such that there are \(i-1\) zs, a single \(y\), and \(k+1-i\) xs. Then we have

\[
    f_1(x,x,y) = e(x,\ldots,x,x) = x,
\]

and for \(i < k\) we have

\[
    f_i(x,y,x) = e(x,\ldots,x,y,x,\ldots,x) = x.
\]

From the construction of the \(f_i\)'s we have

\[
    f_i(x,y,y) = e(x,\ldots,x,y,y,\ldots,y) = f_{i+1}(x,x,y)
\]

91
for \( i + 1 < k \). Finally, if we define \( f_k(x, y, z) \) by
\[
f_k(x, y, z) = e(y, x, y, z, \ldots, z)
\]
and \( p(x, y, z) \) by
\[
p(x, y, z) = e(y, x, z, z, \ldots, z),
\]
then
\[
f_{k-1}(x, y, y) = e(x, x, x, y, \ldots, y) = f_k(x, x, y)
\]
and
\[
f_k(x, y, x) = e(y, x, y, x, \ldots, x) = x
\]
by the \( k \)-edge identities, while
\[
f_k(x, y, y) = e(y, x, y, y, \ldots, y) = p(x, y, y)
\]
and
\[
p(x, x, y) = e(x, x, y, y, \ldots, y) = y
\]
by the \( k \)-edge identities again. Thus \( f_1, \ldots, f_k, p \) are a sequence of directed Gumm terms. \( \square \)

**Theorem 12.15.** For \( k \geq 3 \), an algebra has a \( k \)-edge term and generates a congruence distributive variety iff it has a \( k \)-ary near-unanimity term.

**Proof.** First the easy direction. If an algebra \( \mathcal{A} \) has a \( k \)-ary near-unanimity term \( t \), then adding an extra variable at the beginning of \( t \) produces a \( k \)-edge term. Additionally, the discussion before Theorem A.50 shows that we can construct a sequence of Jónsson terms from \( t \), and then Theorem A.46 shows that \( \mathcal{A} \) generates a congruence distributive variety.

Now the harder direction: assume that \( \mathcal{A} \) generates a congruence distributive variety and has a \( k \)-edge term \( e \). By Theorem A.50, there is a sequence of directed Jónsson terms \( f_1, \ldots, f_m \), that is, a sequence satisfying the system of identities
\[
f_1(x, x, y) \approx x,
\]
\[
f_i(x, y, x) \approx x \text{ for all } i,
\]
\[
f_i(x, y, y) \approx f_{i+1}(x, x, y) \text{ for all } i,
\]
\[
f_m(x, y, y) \approx y.
\]

Let \( \mathcal{F} = \mathcal{F}_{V(\mathcal{A})}(x, y) \leq \mathcal{A}^2 \) be the free algebra on two generators in the variety generated by \( \mathcal{A} \). Let \( \mathcal{S} \leq \mathcal{F}^k \) be generated by the vectors \( (x, \ldots, x, y, x, \ldots, x) \) with all but one entry equal to \( x \) and the remaining entry equal to \( y \). Note that \( \mathcal{S} \) is symmetric under permuting its coordinates. We just need to prove that \( (x, \ldots, x) \in \mathcal{S} \).

**Claim:** For all \( i \), we have \( (f_i(y, x, x), x, \ldots, x) \in \mathcal{S} \).

**Proof of Claim:** We induct on \( i \), taking \( (y, x, \ldots, x) \in \mathcal{S} \) as our base case. By the induction hypothesis, we have
\[
(f_i(y, y, x), x, \ldots, x) = (f_{i-1}(y, x, x), x, \ldots, x) \in \mathcal{S}.
\]
Additionally, the tuples
\[
\begin{bmatrix}
f_i(y, x, x) \\
\vdots \\
x
\end{bmatrix}
= f_i \begin{bmatrix}
y & x & x \\
x & x & y \\
x & y & x \\
\vdots & \vdots & \vdots \\
x & x & x
\end{bmatrix}
\]
and
\[
\begin{bmatrix}
f_i(y, y, x) \\
\vdots \\
x
\end{bmatrix}
= f_i \begin{bmatrix}
y & y & x \\
x & x & y \\
x & x & x \\
\vdots & \vdots & \vdots \\
x & x & x
\end{bmatrix}
\]
are both in \( S \). Now we apply the \( k \)-edge term \( e \):
\[
e \begin{bmatrix}
f_i(y, y, x) & f_i(y, y, x) & f_i(y, x, x) & f_i(y, x, x) & \cdots & f_i(y, x, x) \\
f_i(x, x, y) & x & f_i(x, x, y) & x & \cdots & x \\
x & x & x & f_i(x, x, y) & \cdots & x \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x & x & x & x & \cdots & f_i(x, x, y)
\end{bmatrix}
= \begin{bmatrix}
f_i(y, x, x) \\
x \\
\vdots \\
x
\end{bmatrix}.
\]

To finish the proof, we apply the Claim with \( i = m \) to see that \((x, x, \ldots, x) = (f_m(y, x, x), x, \ldots, x) \in S\).

**Example 12.2.** We give an example of a congruence distributive algebra without few subpowers. Recall from Example 7.6 that for each \( n \), the relational structure \((\{0,1\}, \{0\}, \{0,1\}^n \setminus \{(0,\ldots,0)\})\) has strict width exactly \( n \). The limiting relational clone on \(\{0,1\}\) generated by the relations \(\{0\}\) and \(\{0,1\}^n \setminus \{(0,\ldots,0)\}\) for all \( n \in \mathbb{N} \) is preserved by the ternary operation
\[
f(x, y, z) = x \lor (y \land z).
\]
Since the \( n \)-ary critical relation \(\{0,1\}^n \setminus \{(0,\ldots,0)\}\) doesn’t have the parallelogram property and is preserved by \( f \) for all \( n \), the clone generated by \( f \) can’t have few subpowers by Theorem 13.4.

To check that the algebra \( \mathcal{A} = (\{0,1\}, x \lor (y \land z)) \) generates a congruence distributive variety, consider the sequence of ternary terms given by
\[
f_1(x, y, z) = x \lor (y \land z), \quad f_2(x, y, z) = (x \land y) \lor z.
\]
To see that this is a sequence of directed Jónsson terms, note that they satisfy \( f_1(x, y, x) = x \lor (y \land x) = x \), are connected by
\[
f_1(x, y, y) = x \lor (y \land y) = x \lor y = (x \land x) \lor y = f_2(x, x, y),
\]
and have \( f_1(x, x, y) = x, f_2(x, y, y) = y \). By Theorem A.46 and the discussion before Theorem A.50 this implies that \( \mathcal{A} \) is congruence distributive.
**Example 12.3.** We’ve seen earlier that the two-element semilattice $A = (\{0, 1\}, \max)$ does not have few subpowers. Here we will check that the two-element semilattice does not generate a congruence modular variety. In fact, the congruence lattice $\text{Con}(A^2)$ already fails to be modular. It turns out that every congruence on $A^2$ is generated (as a congruence) by just one pair of elements $a, b \in A^2$, so we can label the nontrivial congruences on $A^2$ by pairs of elements $a, b \in A^2$, yielding the following congruence lattice.

To see that this isn’t modular, note that the sublattice generated by $\ker \pi_1, \ker \pi_2, \Theta_2$ is isomorphic to the pentagon lattice $\mathcal{N}_5$. Considered as an abstract lattice, $\text{Con}(A^2)$ is the standard example of a lattice which is meet-semidistributive (recall from Example 10.6 and Proposition 10.32 that the variety of semilattices is $\text{SD}(\land)$) but not join-semidistributive (we have $\Theta \lor \ker \pi_1 = \Theta \lor \ker \pi_2 = 1_{A^2}$, but $\Theta \lor (\ker \pi_1 \land \ker \pi_2) = \Theta \neq 1_{A^2}$).

Although congruence modularity is slightly weaker than having few subpowers, the concepts are quite close. One hint at the connection between them comes from counting congruences on subpowers of $A$.

**Definition 12.16.** If $A$ is an algebra, then we define the function $c_A(n)$ to be the base-2 logarithm of the maximum size of $\text{Con}(R)$ over all $R \leq A^n$.

**Proposition 12.17.** A variety $\mathcal{V}$ is congruence distributive iff for all subdirect products $R \leq_{\text{sd}} A_1 \times \cdots \times A_n$ in $\mathcal{V}$, every congruence on $R$ can be written as a product of congruences on the $A_i$’s.

**Proof.** Suppose first that $\mathcal{V}$ is congruence distributive. Then for any congruence $\theta$ on $R$, by distributivity and $\bigwedge_i \ker \pi_i = 0_R$ we have

$$\bigwedge_i (\theta \lor \ker \pi_i) = \theta \lor \bigwedge_i \ker \pi_i = \theta \lor 0_R = \theta,$$

so $\theta$ is the product of the congruences $\pi_i(\theta \lor \ker \pi_i) \in \text{Con}(A_i)$.

Conversely, suppose that $A \in \mathcal{V}$, and suppose that $\alpha, \beta, \gamma \in \text{Con}(A)$. Then $A/(\beta \land \gamma)$ is a subdirect product of $A/\beta$ and $A/\gamma$, so the congruence

$$\alpha \lor (\beta \land \gamma),$$

is congruence distributive.
considered as a congruence on \( A/(\beta \land \gamma) \), is a product congruence iff it is equal to
\[
(\alpha \lor \beta) \land (\alpha \lor \gamma).\]

\[\square\]

**Corollary 12.18.** If \( \mathcal{V}(A) \) is congruence distributive, then \( c_A(n) = nc_A(1) \).

If a variety is congruence modular but not congruence distributive, then it necessarily contains a (finitely generated) nontrivial affine algebra. So we need to understand \( c_A(n) \) for \( A \) a finite affine algebra, and since the congruence lattice only depends on the polynomial clone, we may assume that \( A \) is a module over a ring. In this case, there is a bijection between congruences on \( A^n \) and submodules of \( A^n \).

**Proposition 12.19.** If \( A \) is a nontrivial finite module over a ring, then \( c_A(n) \geq n^2 - \frac{1}{4} \).

**Proof.** We may as well assume that \( A \) is simple. Let \( c \) be any nonzero element of \( A \). For \( n = 2m \), the span of the columns of the \( n \times m \) matrix \( \begin{bmatrix} cI \\ M \end{bmatrix} \) completely determines the \( m \times m \) matrix \( M \), so \( c_A(2m) \geq m^2 \log_2(|A|) \geq m^2 \).

\[\square\]

**Corollary 12.20.** If \( A \) is finite and \( \mathcal{V}(A) \) is congruence modular but not congruence distributive, then \( c_A(n) = \Omega(n^2) \).

How can we get an upper bound on \( c_A(n) \) when \( A \) is congruence modular? The trick is to use the fact that in modular lattices, the *height* of the lattice is well-behaved. We can relate the height of a congruence lattice to its size using the following elementary bound.

**Proposition 12.21.** If \( A \) is a finite algebra such that \( \text{Con}(A) \) has height \( h \), then
\[
|\text{Con}(A)| \leq \sum_{i=0}^{h} \left( \frac{|A|}{2} \right)^i \leq |A|^{2h}.
\]

**Proof.** Consider any congruence \( \alpha \in \text{Con}(A) \). Since every cover of \( \alpha \) is generated (as a congruence) by \( \alpha \) together with some pair \((a,b) \notin \alpha\), the number of covers of \( \alpha \) is bounded by \( \left( \frac{|A|}{2} \right) \). Since every element of \( \text{Con}(A) \) can be reached from \( 0_A \) by repeatedly choosing covers at most \( h \) times, we get the stated bound on \( |\text{Con}(A)| \).

We can get a slightly better bound as follows: the above argument shows that every congruence can be generated (as a congruence) by at most \( h \) pairs in \( \left( \frac{|A|}{2} \right) \). Additionally, there is only one congruence at height \( h \), since \( \text{Con}(A) \) has a top element \( 1_A \). So we have
\[
|\text{Con}(A)| \leq 1 + \sum_{i=0}^{h-1} \left( \frac{|A|}{2} \right)^i. \]

\[\square\]

**Corollary 12.22.** If \( A \) is finite and generates a congruence modular variety, then \( c_A(n) \leq n^2 \cdot 2|A| \log_2(|A|) \).

**Proof.** Let \( c \) be the maximum height of \( \text{Con}(B) \) over all subalgebras \( B \leq A \) (\( c \) is automatically bounded by \( |A| \)). We claim that for any \( R \leq A^n \), the height of \( \text{Con}(R) \) is bounded by \( nc \). Since \( \text{Con}(R) \) is modular, we can compute its height by looking at the size of any maximal chain in \( \text{Con}(R) \).

95
We will choose our maximal chain to be an maximal extension of the chain

$$0_R \leq \ker \pi_{[n-1]} \leq \cdots \leq \ker \pi_1 \leq 1_R.$$  

By the Diamond Isomorphism Theorem \([A.27]\), the interval \([\ker \pi_{[i]}, \ker \pi_{[i-1]}]\) is isomorphic to the interval \([\ker \pi_i, \ker \pi_{[i-1]} \lor \ker \pi_{[i]}]\), so its height is bounded by the height of the interval \([\ker \pi_i, 1_R]\), which is isomorphic to \(\text{Con}(R/\ker \pi_i)\). Since \(R/\ker \pi_i \cong \pi_i(R) \leq A\), the height of \(\text{Con}(R/\ker \pi_i)\) is bounded by \(c\), and putting these intervals together we see that the height of \(\text{Con}(R)\) is bounded by \(cn\).

Using the previous bound, we get

$$\log_2(\text{Con}(R)) \leq \log_2(2^{cn}) \leq 2cn \log_2(|A|^n) = 2cn^2 \log_2(|A|).$$

\[\square\]

**Theorem 12.23** (Few congruences on subpowers iff congruence modular \([20]\)). Let \(A\) be a finite algebra with at least two elements, and let \(\mathcal{V}(A)\) be the variety it generates.

- If \(\mathcal{V}(A)\) is congruence distributive, then \(c_A(n) = \Theta(n)\).
- If \(\mathcal{V}(A)\) is congruence modular but not congruence distributive, then \(c_A(n) = \Theta(n^2)\).
- If \(\mathcal{V}(A)\) is not congruence modular, then \(c_A(n) = 2^{\Theta(n)}\).

**Proof.** By the previous results, all we need to check is that if \(\mathcal{V}(A)\) is not congruence modular, then \(c_A(n) = 2^{\Omega(n)}\). Let \(F = F_{\mathcal{V}(A)}(x,y,z,w) \leq A^4\) be the free algebra on four generators. We will show that if \(c_F(2n) < 2^n\) for any \(n\), then \(A\) has Day terms, and is therefore congruence modular by Appendix \([A.1]\).

Define congruences on \(F\) as in Corollary \([A.15]\) let \(\theta_{ab}\) be the congruence generated by the pair \((a,b)\) for any pair of variables \(a,b\), set \(\alpha = \theta_{xy} \lor \theta_{zw}\), \(\beta = \theta_{xz} \lor \theta_{yw}\), and \(\gamma = (\alpha \land \beta) \lor \theta_{zw}\). This is the generic Shifting Lemma configuration:

```
+----+------+
|     |      |
| x   | z    |
+----+------+
|     |      |
| y   | w    |
+----+------+
|     |      |
| \alpha|     |
| \beta |      |
|      | \gamma|
```

To show the existence of Day terms, we just need to show that \((x,y) \in \gamma\).

Pick an \(n\) such that \(c_F(2n) < 2^n\), and consider the subalgebra \(R \leq F^{2n}\) consisting of tuples such that every pair of coordinates are related by \(\beta\) (it helps to imagine elements of \(R\) written out horizontally as row vectors, following the convention that variables which are related by \(\beta\) are laid out on horizontal lines). We will define a family of \(2^n\) pairs of elements of \(R\) as follows.

First, we define elements \(x^0, x^1, y^0, y^1 \in F^2\) by \(x^0 = (x,z), x^1 = (z,x)\) and similarly \(y^0 = (y,w), y^1 = (w,y)\). Then, for any \(i = (i_1, \ldots, i_n) \in \{0,1\}^n\), we define \(f_i, g_i \in R\) by

\[f_i = (x^{i_1}, \ldots, x^{i_n}),\]
\[g_i = (y^{i_1}, \ldots, y^{i_n}).\]
For each $i \in \{0,1\}^n$, we define a congruence $\Theta(i)$ to be the congruence of $\mathbb{R}$ generated by the pair $(f_i, g_i)$. Since $c_\mathbb{R}(2n) < 2^n$, there must be some $i \in \{0,1\}^n$ such that

$$\Theta(i) \leq \bigvee_{j \neq i} \Theta(j),$$

and by permuting the coordinates of $\mathbb{R}$, we see that in fact this must hold for every $i$, and in particular for $i = (0,\ldots,0)$. By dropping half of the coordinates of $\mathbb{R}$ to get a similar algebra $\mathbb{R}' \leq \mathbb{F}^n$ such that $f_0$ becomes the vector $f_0' = (x,\ldots,x)$ and $g_0$ becomes the vector $g_0' = (y,\ldots,y)$, and defining elements $f_j', g_j'$ by dropping half the coordinates of $f_j, g_j$, we see that

$$(f_0', g_0') \in \bigvee_{j \neq (0,\ldots,0)} \Theta'(j),$$

where $\Theta'(j)$ is the congruence of $\mathbb{R}'$ generated by the pair $(f_j', g_j')$.

Each $\Theta'(j)$ has the following property: if $(a, b) \in \Theta'(j)$ and every pair of coordinates of $a$ are related by $\alpha \land \beta$, then every pair of coordinates of $b$ are also related by $\alpha \land \beta$. To see this, just note that for each coordinate $i \leq n$ we have $(a_i, b_i) \in \alpha$, since this holds in the case where $(a, b) = (f_j', g_j')$.

For $j \neq (0,\ldots,0), \Theta'(j)$ has the following additional property: there exists some coordinate $i \leq n$ such that if $(a, b) \in \Theta'(j)$, then $(a_i, b_i) \in \gamma$. In fact, we can take the coordinate $i$ to be the first coordinate of $j$ such that $j_i = 1$, and note that the $i$th coordinates of $f_j', g_j'$ are $z, w$ respectively, with $(z, w) \in \gamma$ by the definition of $\gamma$.

Putting the above properties together, and using $\alpha \land \beta \leq \gamma$, we see that $(f_0', b) \in \bigvee_{j \neq (0,\ldots,0)} \Theta'(j)$ implies that every coordinate of $f_0'$ is congruent modulo $\gamma$ to every coordinate of $b$, and taking $b = g_0'$ we see that $(x, y) \in \gamma$, which completes the proof.

**Example 12.4.** Consider the two-element semilattice $\mathbb{A} = \{0,1\}$, max once again. In this case, we can check directly that $c_\mathbb{A}(n) \geq \binom{n}{n/2}$. To see this, note that for every nonempty upwards closed subset $U \leq \mathbb{A}^n$, there is a congruence $\theta_U$ which collapses all elements of $U$ into a single top element of $\mathbb{A}^n/\theta_U$, and which does not identify any pair of elements $a \neq b$ such that $\{a, b\} \not\subseteq U$. In other words, $\theta_U = U^2 \cup \Delta_{\mathbb{A}^n}$.

We just need to check that the number of distinct nonempty upwards closed subsets $U$ of $\{0,1\}^n$ is at least $2^{\binom{n}{n/2}}$: for this, note that upwards closed sets $U$ are in a one-to-one correspondence with antichains (every upwards closed set $U$ is determined by its antichain of minimal elements), and every set of elements of $\mathbb{A}^n$ which each have exactly $n/2$ coordinates equal to 1 forms an antichain.
13 Parallelogram terms

Examining the proof of Theorem 12.11, we can extract useful terms known as parallelogram terms, which we can use to give a better description of the relational clone corresponding to an algebra with few subpowers.

Definition 13.1. If \( k = m + n \), then an \( m,n \)-parallelogram term is a \( k+3 \)-ary term \( r \) which satisfies the identities

\[
\begin{pmatrix}
y y x z & \cdots & x x & \cdots & x \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
y y x x & \cdots & z x & \cdots & x \\
x y y x & \cdots & x z & \cdots & x \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
x y y x & \cdots & x x & \cdots & z \\
\end{pmatrix}
= \begin{pmatrix}
x \\
\vdots \\
x \\
x \\
x \\
\end{pmatrix},
\]

where the upper left \( m \times 3 \) block has all rows given by \( y,y,x \), the lower left \( n \times 3 \) block has all rows given by \( x,y,y \), and the right \( k \times k \) block has \( z \)s on the diagonal and \( x \)s elsewhere.

Theorem 13.2 (Edge term implies parallelogram terms \([75]\)). For any \( m,n > 0 \) with \( m + n = k \), a variety has a \( k \)-edge term \( e \) iff it has an \( m,n \)-parallelogram term \( r \).

Proof. It’s clear that every \( m,n \)-parallelogram term is a \( \Delta \)-cube term for

\[
\Delta = \{\{1,\ldots,m\}, \{1,\ldots,k\}, \{m+1,\ldots,m+n\}, \{1\},\ldots,\{k\}\},
\]

so by Theorem 12.7 if \( A \) has a parallelogram term then it has an edge term.

Now suppose that \( e \) is a \( k \)-edge term. We will build \( m,n \)-parallelogram terms \( r_m \) by induction on \( m \). For \( m = 1 \), we need to show that the vector in \( F(x,y,z)^k \) of all \( x \)s is in the subalgebra generated by the columns of the matrix defining a \( 1,k-1 \)-parallelogram term. These vectors are the vectors where all entries other than one are \( x \)s and the last is a \( z \), the vector of all \( y \)s, and the vectors \( (x,y,\ldots,y),(y,x,\ldots,x) \).

Letting \( d = d(y,x) = e(x,y,x,\ldots,x) \), we have

\[
e\begin{pmatrix}
x y x x & \cdots & x \\
z z x x & \cdots & x \\
x x x z & \cdots & x \\
\vdots & \vdots & \ddots & \vdots \\
x x x x & \cdots & z \\
\end{pmatrix}
= \begin{pmatrix}
d \\
x \\
x \\
\vdots \\
x \\
\end{pmatrix},
\]

and

\[
e\begin{pmatrix}
x y x x & \cdots & x \\
y y z z & \cdots & z \\
y y x x & \cdots & x \\
\vdots & \vdots & \ddots & \vdots \\
y y x x & \cdots & x \\
\end{pmatrix}
= \begin{pmatrix}
d \\
z \\
x \\
\vdots \\
x \\
\end{pmatrix}.
\]
so the vectors \((d, x, x, \ldots, x)\) and \((d, z, x, \ldots, x)\) are in the subalgebra of \(\mathcal{F}(x, y, z)^k\) generated by the columns of the matrix defining a \(1, k - 1\)-parallelogram term. Note that the previous two applications of the edge term \(e\) correspond to applications of the terms

\[s(x_1, \ldots, x_k) = e(x_2, x_1, x_2, \ldots, x_k)\]

and

\[p(x, y, z) = e(y, x, z, \ldots, z)\]

which act like near-unanimity and Mal’cev terms, respectively. To get the vector of all \(x\)s, we apply \(e\) one more time:

\[
e \begin{pmatrix}
  d & d & x & x & \cdots & x \\
  z & x & z & x & \cdots & x \\
  x & x & x & z & \cdots & x \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  x & x & x & x & \cdots & z
\end{pmatrix} =
\begin{pmatrix}
  x \\
  x \\
  x \\
  \vdots \\
  x
\end{pmatrix}.
\]

Explicitly, our \(1, k - 1\)-parallelogram term \(r_1\) is defined from the edge term \(e\) by

\[r_1(x, y, z, u_1, \ldots, u_k) = e(p(y, z, u_2), s(x, u_2, \ldots, u_k), u_2, \ldots, u_k) = e(e(z, y, u_2, \ldots, u_2), e(u_2, x, u_2, \ldots, u_k), u_2, \ldots, u_k).
\]

For \(m > 1\), we construct the \(m, k - m\)-parallelogram term \(r_m\) using the previous term \(r_{m-1}\). Here we focus on the \(m\)th rows of our matrices. Let

\[a = r_{m-1}(y, y, x, \ldots, x, z, x, \ldots, x),\]

where the \(z\) occurs in the \(m + 3\)rd entry. We want to construct tuples \((x, \ldots, x, a, x, \ldots, x)\) and \((x, \ldots, x, a, y, \ldots, y)\) from the columns of the defining matrix for an \(m, k - m\)-parallelogram term. We construct these tuples via

\[
r_{m-1} \begin{pmatrix}
  y & y & x & z & \cdots & x & x & \cdots & x \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  y & y & x & x & \cdots & z & x & \cdots & x \\
  y & y & x & x & \cdots & x & z & \cdots & x \\
  x & y & y & x & \cdots & x & x & z & \cdots & x \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  x & y & y & y & \cdots & y & x & y & \cdots & y
\end{pmatrix} = \begin{pmatrix}
  x \\
  \vdots \\
  x
\end{pmatrix},
\]

and

\[
r_{m-1} \begin{pmatrix}
  y & y & x & x & \cdots & x & x & \cdots & x \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  y & y & x & x & \cdots & x & x & \cdots & x \\
  y & y & x & x & \cdots & x & z & \cdots & x \\
  y & y & y & y & \cdots & y & x & y & \cdots & y \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  y & y & y & y & \cdots & y & x & y & \cdots & y
\end{pmatrix} = \begin{pmatrix}
  y \\
  \vdots \\
  y
\end{pmatrix}.
\]
To get to the vector of all $x$s, we use

$$r_{m-1} \left( \begin{bmatrix}
    x & x & x & x & \cdots & z & x & \cdots & x \\
    \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
    x & x & x & z & \cdots & x & x & \cdots & x \\
    a & a & x & z & \cdots & x & x & \cdots & x \\
    x & y & x & x & \cdots & x & z & \cdots & x \\
    \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
    x & y & y & x & x & \cdots & x & x & \cdots & z
\end{bmatrix} \right) = \begin{bmatrix}
    x \\
    \vdots \\
    x \\
    x \\
    x \\
    \vdots \\
    x
\end{bmatrix},$$

where the middle row works out because $m > 1$. Explicitly, $r_m$ is defined in terms of $r_{m-1}$ by

$$r_m(x, y, z, u_1, ..., u_k) = t_{m-1}(t_{m-1}(x, y, z, u_1, ..., u_k), t_{m-1}(y, y, x, z, ..., u_m, ..., z), z, u_m, ..., u_1, u_{m+1}, ..., u_k).$$

To understand what parallelogram terms tell us, it is necessary to restrict to certain special relations, known as critical relations.

**Definition 13.3.** A subalgebra $\mathbb{R} \in A_1 \times \cdots \times A_n$ is critical if it is $\cap$-irreducible, that is, if it can’t be written as an intersection of strictly larger subalgebras, and if furthermore the relation $\mathbb{R}$ has no dummy variables (that is, it depends on all of its inputs).

A standard result in the theory of algebraic lattices (Proposition A.56 from Appendix A) shows that every relation can be written as an intersection of critical relations (possibly of lower arity). The following result shows that every relation in an algebra with $k$-parallelogram terms can be written as an intersection of relations of arity less than $k$ and relations with the parallelogram property.

**Theorem 13.4 (Parallelogram terms constrain critical relations [75]).** A variety $\mathcal{V}$ has $k$-parallelogram terms iff for all critical $\mathbb{R} \leq A_1 \times \cdots \times A_n$ with $A_i \in \mathcal{V}$, either $n < k$ or $\mathbb{R}$ has the parallelogram property.

**Proof.** First suppose that $\mathcal{V}$ has $k$-parallelogram terms, and let $\mathbb{R} \leq A_1 \times \cdots \times A_n$ be a critical relation. Let $\mathbb{R}^*$ be the cover of $\mathbb{R}$, i.e., $\mathbb{R}^*$ is the intersection of all relations which properly contain $\mathbb{R}$, and let $a = (a_1, ..., a_n) \in \mathbb{R}^* \setminus \mathbb{R}$. Then a relation $\mathbb{S}$ which contains $\mathbb{R}$ will properly contain $\mathbb{R}$ iff $\mathbb{S}$ contains $a$. Following Zhuk [114], we call $a$ a key tuple for the critical relation $\mathbb{R}$.

Since $\mathbb{R}$ is critical, $\mathbb{R}$ is properly contained in its existential projections onto any proper subset of the coordinates $1, ..., n$. Thus, there must exist elements $b_1, ..., b_n$ such that the tuples $(b_1, a_2, ..., a_n), (a_1, b_2, ..., a_n), ..., (a_1, a_2, ..., b_n)$ are all in $\mathbb{R}$.

Now suppose, for contradiction, that $n \geq k$ and that $\mathbb{R}$ does not have the parallelogram property when considered as a binary relation on $(A_1 \times \cdots \times A_i) \times (A_{i+1} \times \cdots \times A_n)$. Then the are $x_1, ..., x_n, y_1, ..., y_n$ such that the three tuples $(y_1, ..., y_n), (y_1, ..., y_i, x_{i+1}, ..., x_n), (x_1, ..., x_i, y_{i+1}, ..., y_n)$ are in $\mathbb{R}$, but $(x_1, ..., x_n)$ is not in $\mathbb{R}$. Since $x = (x_1, ..., x_n)$ is not in $\mathbb{R}$, the subalgebra generated by $\mathbb{R} \cup \{x\}$ must properly contain $\mathbb{R}$, so

$$a \in \text{Sg}(\mathbb{R} \cup \{x\}).$$
Thus there are tuples $c^1, ..., c^m \in \mathbb{R}$ and an $m + 1$-ary term $t$ such that

$$t(x, c^1, ..., c^m) = a.$$  

Defining a tuple $d$ by

$$t(y, c^1, ..., c^m) = d,$$

we see that the three tuples $(d_1, ..., d_n), (d_1, ..., d_i, a_{i+1}, ..., a_n), (a_1, ..., a_i, d_{i+1}, ..., d_n)$ are all in $\mathbb{R}$. But then we can use an $i, n - i$-parallelogram term $r$ (which exists because $n \geq k$) to see that

$$\begin{bmatrix}
a_1 \\
\vdots \\
a_i \\
a_{i+1} \\
\vdots \\
a_n
\end{bmatrix} = r
\begin{bmatrix}
d_1 & d_1 & a_1 & b_1 & \cdots & a_1 & a_1 & \cdots & a_1 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
a_{i+1} & a_{i+1} & a_{i+1} & b_{i+1} & \cdots & a_{i+1} & b_{i+1} & \cdots & a_{i+1} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
a_n & a_n & a_n & a_n & \cdots & a_n & a_n & \cdots & b_n
\end{bmatrix}\in \mathbb{R},$$

contradicting the assumption that $a \notin \mathbb{R}$.

For the converse direction, suppose that $\mathcal{V}$ is a variety such that every critical $k$-ary relation has the parallelogram property, and suppose that $m + n = k$. Let $\mathcal{F} = \mathcal{F}_\mathcal{V}(x, y, z)$ be the free algebra on three generators in $\mathcal{V}$. Suppose for contradiction that $\mathcal{V}$ doesn’t have an $m, n$-parallelogram term. Then

$$\begin{bmatrix}x \\
x \\
x \\
\end{bmatrix} \notin \text{Sg}_{\mathcal{F}^k}\begin{bmatrix}y & y & x & \cdots & x & x & \cdots & x \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
y & y & x & \cdots & z & x & \cdots & x \\
x & y & x & \cdots & z & x & \cdots & x \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x & y & x & \cdots & x & x & \cdots & z
\end{bmatrix},$$

so by Zorn’s Lemma there exists a maximal $k$-ary relation $\mathcal{R}$ on $\mathcal{F}$ which contains the right hand side but does not contain the tuple $(x, ..., x)$. The relation $\mathcal{R}$ is then a critical $k$-ary relation on $\mathcal{F}$, since every relation which properly contains $\mathcal{R}$ must contain $\mathcal{R}^* = \text{Sg}(\mathcal{R} \cup \{(x, ..., x)\})$ and since every existential projection of $\mathcal{R}$ onto a proper subset of the coordinates contains a vector of all $x$s (by the last $k$ columns of the matrix of generators above). However, $\mathcal{R}$ does not have the parallelogram property when considered as a binary relation on $\mathcal{F}^m \times \mathcal{F}^n$, by the first three columns of the matrix of generators above, contradicting our assumption on $\mathcal{V}$. □

**Corollary 13.5.** A variety $\mathcal{V}$ has $k$-parallelogram terms iff for every relation $\mathcal{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ with $\mathbb{A}_i \in \mathcal{V}$, there exists a relation $\mathcal{R}' \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ such that $\mathcal{R}'$ has the parallelogram property and

$$\mathcal{R} = \mathcal{R}' \cap \bigcap_{I \subseteq [n], |I| < k} \pi_I(\mathcal{R}).$$

The relation $\mathcal{R}'$ from the corollary need not be so mysterious: we can take it to be the least relation $\mathcal{R}'$ which contains $\mathcal{R}$ and has the parallelogram property, since any intersection of relations which have the parallelogram property also has the parallelogram property. This choice of $\mathcal{R}'$ can
also be “generated” from $\mathbb{R}$, by repeatedly adjoining tuples which are required to be inside in order for the parallelogram property to hold.

More explicitly, for any $I \subseteq [n]$, we can find the least relation $R'$ which contains $R$ and has the (binary) parallelogram property when considered as a subalgebra of

$$\left( \prod_{i \in I} A_i \right) \times \left( \prod_{j \notin I} A_j \right),$$

by finding the linking congruence of $R$ when considered as a subalgebra of the above, which restricts to a congruence $\alpha_I \in \text{Con}(\pi_I(\mathbb{R}))$, and taking $R'$ to be the relation $\alpha_I \circ R$. We can then take

$$R' = \bigcup_{I_1, I_2, \ldots \subseteq [n]} R_{I_1 I_2 \ldots}.$$

In particular, if all of the algebras $A_i$ are finite, then $R'$ is contained in the (multisorted) relational clone generated by $\mathbb{R}$.

### 13.1 Critical rectangular relations in congruence modular varieties

Using the commutator theory for congruence modular varieties, we can give a more detailed structure theory for the high-arity critical relations preserved by algebras with few subpowers. In fact, this structure theory applies more generally in congruence modular varieties, so long as we restrict our attention to critical relations with a weak form of the parallelogram property.

**Definition 13.6.** A relation $R \leq A_1 \times \cdots \times A_k$ is said to have the $1, k-1$-parallelogram property, or alternatively is called rectangular, if for any $i \leq k$, when we regard $R$ as a binary relation on

$$(A_1 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_k) \times A_i,$$

it has the (binary) parallelogram property.

The main property of subdirect rectangular relations which we need - and which holds in complete generality, not just in the context of congruence modularity - is that if we define a congruence $\theta_i$ on $A_i$ from the linking congruence of $R$ (considered as a binary relation on $(\cdots) \times A_i$), then we have $x \in R$ iff $x/\prod_i \theta_i \in \mathbb{R}/\prod_i \theta_i$. Thus we may as well study the relation

$$R/\prod_i \theta_i \leq_{sd} A_1/\theta_1 \times \cdots \times A_k/\theta_k$$

instead of studying $R$. The reduced relation is critical iff the original $R$ is critical, is still rectangular, and has trivial linking congruences on each $A_i/\theta_i$, so it can be viewed as the graph of a surjective homomorphism

$$\pi_{[k] \setminus \{i\}} \left( R/\prod_i \theta_i \right) \twoheadrightarrow A_i/\theta_i$$

for each $i$.

**Definition 13.7.** A subdirect rectangular relation $R \leq A_1 \times \cdots \times A_k$ is called reduced if for each $i \leq k$, $R$ is the graph of a surjective homomorphism

$$\pi_{[k] \setminus \{i\}}(R) \rightarrow A_i,$$
or equivalently, for each \( i \) the map

\[
\pi[k]\setminus\{i\} : \mathbb{R} \rightarrow \pi[k]\setminus\{i\}(\mathbb{R})
\]

is an isomorphism, i.e. \( \ker \pi[k]\setminus\{i\} = 0_{\mathbb{R}} \).

**Proposition 13.8.** If \( \mathbb{R} \leq_{sd} A_1 \times \cdots \times A_k \) is a reduced subdirect critical rectangular relation, then each \( A_i \) is subdirectly irreducible.

**Proof.** Let \( \mathbb{R}^* \) be the cover of \( \mathbb{R} \) is the lattice of subalgebras of \( A_1 \times \cdots \times A_k \), and let \( a = (a_1, \ldots, a_k) \) be a key tuple for \( \mathbb{R} \), that is, an element of \( \mathbb{R}^* \setminus \mathbb{R} \). Since \( \mathbb{R} \) is critical, for every \( i \) there is some \( b_i \in A_i \) such that \((a_1, \ldots, a_{i-1}, b_i, a_{i+1}, \ldots, a_k) \in \mathbb{R} \) (and this \( b_i \) is unique, since \( \mathbb{R} \) is reduced). The claim is that for each \( i \), every nontrivial congruence on \( A_i \) contains the pair \((a_i, b_i)\) - that is, each \( A_i \) is subdirectly irreducible with monolith equal to the congruence generated by the pair \((a_i, b_i)\).

Let \( \psi_i \in \text{Con}(A_i) \) be any nontrivial congruence. Then the relation

\[
\exists y_i \ ( (x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_k) \in \mathbb{R} ) \land (x_i \equiv_{\psi_i} y_i)
\]

strictly contains \( \mathbb{R} \) (since \( \mathbb{R} \) is reduced), so it contains \( \mathbb{R}^* \), and in particular contains the key tuple \( a \). Using the fact that \( \mathbb{R} \) is reduced again, we see that the pair \((a_i, b_i)\) must be contained in \( \psi_i \). \( \square \)

As it turns out, reduced critical rectangular relations are closely related to the concept of *similarity* between subdirectly irreducible algebras (see Appendix [A.5.1](#)). We won’t need the full theory of similarity, just the following definition.

**Definition 13.9.** If \( A_1, \ldots, A_k \) are subdirectly irreducible algebras, then we say that an algebra \( \mathbb{R} \) is the graph of a joint similarity between the \( A_i \)s if for each \( i \), \( \mathbb{R} \) has a (critical) congruence \( \alpha_i \) with \( \mathbb{R}/\alpha_i \cong A_i \), and for each pair \( i, j \) there are congruences \( \gamma_{ij}, \delta_{ij} \in \text{Con}(\mathbb{R}) \) such that

\[
[\alpha_i, \alpha_j^*] \downarrow [\gamma_{ij}, \delta_{ij}] \not\supset [\alpha_j, \alpha_j^*].
\]

More explicitly, this means that \( \alpha_i \lor \delta_{ij} = \alpha_j^* \), \( \alpha_j \lor \delta_{ij} = \alpha_j^* \), and \( \alpha_i \land \delta_{ij} = \alpha_j \land \delta_{ij} \).

Note that by Proposition [A.86](#), \( \mathbb{R}/(\alpha_1 \land \cdots \land \alpha_k) \) is also a graph of a joint similarity, so there is no real loss in restricting to the case where \( \mathbb{R} \) is a subdirect product of the \( A_i \)s, with \( \alpha_i = \ker \pi_i \).

**Theorem 13.10** (Kearnes, Szendrei [75]). If \( \mathbb{R} \leq_{sd} A_1 \times \cdots \times A_k \) is a reduced subdirect critical rectangular relation of arity \( k \geq 3 \) in a congruence modular variety, then

(a) \( \mathbb{R} \) is the graph of a joint similarity between the \( A_i \)s,

(b) for each \( i, j \), the image of \( \pi_{i,j}(\mathbb{R}) \) in \( A_i/(0_{A_i} : 0_{A_i}^*) \times A_j/(0_{A_j} : 0_{A_j}^*) \) is the graph of an isomorphism

\[
A_i/(0_{A_i} : 0_{A_i}^*) \cong A_j/(0_{A_j} : 0_{A_j}^*),
\]

(c) each monolith \( 0_{A_i}^* \) is abelian, and

(d) the cover \( \mathbb{R}^* \) is also rectangular, and the linking congruence of \( \mathbb{R}^* \) on \( A_i \) is the monolith \( 0_{A_i}^* \).

If \( \mathbb{R} \) has the parallelogram property, then so does its cover \( \mathbb{R}^* \).
Proof. (a) Let \( a = (a_1, \ldots, a_k) \in \mathbb{R}^* \setminus \mathbb{R} \) be a key tuple for \( \mathbb{R} \), and for each \( i \) let \( b_i \in A_i \) such that \( (a_1, \ldots, a_{i-1}, b_i, a_{i+1}, \ldots, a_k) \in \mathbb{R} \). Let \( a^i = (a_1, \ldots, a_{i-1}, b_i, a_{i+1}, \ldots, a_k) \). Then for any \( i \neq j \), if we let \( \delta_{ij} \) be the congruence generated by the pair \((a^i, a^j)\), we claim that

\[
[\ker \pi_i, (\ker \pi_i)^*] \setminus [0_{\mathbb{R}}, \delta_{ij}].
\]

The equality \( \ker \pi_i \vee \delta_{ij} = (\ker \pi_i)^* \) was proved in the previous proposition. For the equality \( \ker \pi_i \wedge \delta_{ij} = 0_{\mathbb{R}} \), note that

\[
\ker \pi_i \wedge \delta_{ij} \leq \ker \pi_i \wedge \ker \pi[k] \setminus \{i,j\} = \ker \pi[k] \setminus \{j\} = 0_{\mathbb{R}},
\]

where the last equality follows from the fact that \( \mathbb{R} \) is reduced.

(b) This follows directly from (a) and the Diamond Isomorphism Theorem [A.27] - for details, see Proposition [A.86].

(c) By Proposition [A.86] again, if \( \pi_{i,j}(\mathbb{R}) \) is not the graph of an isomorphism for any pair \( i, j \), then each monolith \( 0_{A_i}^* \) must be abelian.

(d) Suppose that \( u, v, w \in \mathbb{R} \) with \( \pi[k] \setminus \{i\}(u) = \pi[k] \setminus \{i\}(v) \) and \( v_i = w_i \). We need to show that there is some element \( t \in \mathbb{R} \) with \( \pi[k] \setminus \{i\}(t) = \pi[k] \setminus \{i\}(w) \) and \( t_i = u_i \).

Since \( \mathbb{R}^* \) is contained in the relation

\[
\exists y_i \ ((x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_k) \in \mathbb{R}) \wedge (x_i \equiv_{0_{A_i}} y_i)
\]

and \( \mathbb{R} \) is reduced, we have \((u_i, v_i) \in 0_{A_i}^* \). Let \( p(x, y, z) \) be a Gumm difference term as in Theorem [A.35], i.e. a term such that \( p(y, y, x) \approx x \), and such that for \( x, y \in \theta \) and \( \theta \) any congruence we have \( p(x, y, x) \setminus [\theta, \theta] \) \( x \). Then taking \( \theta = 0_{A_i}^* \), we have \( p(u_i, v_i, u_i) = u_i \) by part (c), so we can take \( t = p(u, v, w) \).

For the last claim, suppose that we view \( \mathbb{R}^* \) as a binary relation on \( A_I \times A_{[n]}/I \), where we set \( A_I = \prod_{i \in I} A_i \), and that we have \((a, b), (c, b), (c, d) \in \mathbb{R}^* \). Pick some \( i \in I \) and \( j \not\in I \). Then there is some \( a' \) such that \( \pi_I \setminus \{i\}(a') = \pi_I \setminus \{i\}(a), a'_i \equiv_{0_{A_i}} a_i \), and \((a', b) \in \mathbb{R} \). Similarly find \( c' \) which only differs from \( c \) in the \( i \)th coordinate, has \( c'_i \equiv_{0_{A_i}} a_i \), and has \((c', b) \in \mathbb{R} \). Then \((c', d) \in \mathbb{R}^* \) by part (d), so we can find \( d' \) which only differs from \( d \) in the \( j \)th coordinate, has \( d'_i \equiv_{0_{A_j}} d_i \), and has \((c', d') \in \mathbb{R} \). Then by the parallelogram property for \( \mathbb{R} \), we have \((a', d') \in \mathbb{R} \), so by part (d) we have \((a, d) \in \mathbb{R}^* \).

\[ \square \]

Example 13.1. Consider the generalized majority-minority algebra \( A = \{a, b, c\} \), \( \varphi_2 \) from Example 11.2, which is subdirectly irreducible with abelian monolith \( 0_{A}^* \) corresponding to the partition \( \{a\}, \{b, c\} \) of its elements, and has \( A/0_{A}^* \) isomorphic to a two element majority algebra. We can check that the monolith \( 0_{A}^* \) of \( A \) is equal to its own centralizer by verifying that \([1_A : 0_{A}^*] = 0_{A}^* \) and \([0_{A}^*, 0_{A}^*] = 0_{A} : to see this, note that

\[
\varphi_2 \left( \begin{bmatrix} a & a \\ b & b \end{bmatrix}, \begin{bmatrix} a & a \\ b & b \end{bmatrix}, \begin{bmatrix} b & c \\ b & c \end{bmatrix} \right) = \begin{bmatrix} a & a \\ b & c \end{bmatrix} \in M(1_A, 0_{A}^*),
\]

so \((b, c) \in [1_A, 0_{A}^*] \), while every element of \( M(0_{A}^*, 0_{A}^*) \) either has all entries equal to \( a \), or has all entries in \( \{b, c\} \) with an even number of \( bs \) and an even number of \( cs \).

The ternary relation \( \mathbb{R} \leq_{sd} A^3 \) corresponding to the columns of the matrix

\[
\begin{bmatrix}
a & b & c & c \\
a & b & c & b \\
a & b & c & b
\end{bmatrix}
\]

104
is a reduced subdirect critical rectangular relation of arity 3 (with key tuple \((c,c,c)\)), so by the structure theorem it is the graph of a joint similarity between three copies of \(\mathbb{A}\). Every two-coordinate projection \(\pi_{i,j}(\mathbb{R})\) is equal to the congruence \(0^*_A = (0_A : 0_A^*)\), and the cover \(\mathbb{R}^*\) of \(\mathbb{R}\) in \(Inv_3(\mathbb{A})\) is the relation \(x \ 0^*_A \ y \ 0^*_A \ z\).

More generally, for any \(k\) we can define a relation \(\mathbb{R}_k \leq_{sd} \mathbb{A}^k\) which contains the tuple \((a,\ldots,a)\) together with the \(2^{k-1}\) tuples in \(\{b,c\}^k\) such that the total number of \(c\)s is even, and we see that \(\mathbb{R}_k\) is a reduced critical rectangular relation for each \(k\). We claim that for every \(k \geq 3\), there are exactly four critical relations in \(Inv_k(\mathbb{A})\): \(\mathbb{R}_k\), \(\mathbb{R}_k \setminus \{(a,\ldots,a)\}\), and the two relations we get from these by swapping \(b\)s and \(c\)s in the last coordinate.

To prove the claim, we first note that the only algebras in \(HS(\mathbb{A})\) which have abelian monoliths are \(\mathbb{A}\) and \(\{b,c\}\), and that these two algebras are not similar to each other (since \(\mathbb{A}/0^*_A\) is not isomorphic to any quotient of \(\{b,c\}\)). Thus by the structure theorem, we only need to consider relations which are either subdirect in \(\mathbb{A}^k\) or subdirect in \(\{b,c\}^k\). The interesting case is the case of relations which are subdirect in \(\mathbb{A}^k\).

The next thing we need to check is that no graph of a similarity \(\mathbb{C} \leq_{sd} \mathbb{A}^2\) from \(\mathbb{A}\) to \(\mathbb{A}\) induces the isomorphism \(\mathbb{A}/0^*_A \rightarrow \mathbb{A}/0^*_A\) which corresponds to swapping the equivalence classes \(\{a\}\) and \(\{b,c\}\) of \(0^*_A\). Note that the only candidate for \(\mathbb{C}\) is the relation \(\{(a,b),(a,c),(b,a),(c,a)\}\), and for this choice of \(\mathbb{C}\) the congruence lattice \(Con(\mathbb{C})\) is given by the following picture.

![Congruence Lattice Diagram]

As the reader can see, there is no pair \(\gamma, \delta \in Con(\mathbb{C})\) such that \([\ker \pi_1 \lor \ker \pi_2] \searrow [\gamma, \delta] \nearrow [\ker \pi_2 \lor \ker \pi_1\], so \(\mathbb{C}\) is not the graph of a similarity. Alternatively, we can see that \(\mathbb{C}\) can’t be the graph of a similarity using the characterization in Corollary \(A.88\), since the corresponding congruence classes of \(0^*_A\) which are linked by \(\mathbb{C}\) do not have the same sizes.

Thus, in any subdirect critical relation \(\mathbb{R} \leq_{sd} \mathbb{A}^k\) of arity \(k > 2\), each \(\pi_{i,j}(\mathbb{R})\) must be the congruence \(0^*_A\), so \(\mathbb{R}\) will consist of the tuple \((a,\ldots,a)\) together with some subalgebra of \(\{b,c\}^k\). Since for any \(S \leq \{b,c\}^k\) the set \(S \cup \{(a,\ldots,a)\}\) will always be closed under \(\varphi_2\), if \(\mathbb{R}\) is critical then so is \(\mathbb{R} \setminus \{(a,\ldots,a)\}\), and it’s easy to check that there are only two critical relations \(S \leq_{sd} \{b,c\}^k\). This completes the classification of critical relations in \(Inv_k(\mathbb{A})\) for \(k > 2\).

**Remark 13.1.** Using the structure theorem \(A.10\) and the fact that the centralizer of the monolith \((0:0^*)\) is automatically abelian for subdirectly irreducible algebras in residually small congruence modular varieties (Corollary \(A.80\)), one can easily reduce the subpower membership problem \(A.11\) for residually small congruence modular varieties to the subpower membership problem for abelian groups by taking advantage of the properties of the Gumm difference term (see Corollary \(A.42\)). For details of the reduction, see \(A.33\).

**Example 13.2.** We give an example of a minimal algebra with few subpowers which does not generate a residually small variety. Let \(\mathbb{A} = \{(a,b,c,d),g\}\), where \(g\) is the idempotent ternary symmetric operation which is determined by that fact that it commutes with the cyclic permutation.
\( \sigma = (a \ b \ c \ d) \) and satisfies

\[
\begin{align*}
g(a, a, b) &= a, \\
g(a, a, c) &= c, \\
g(a, a, d) &= c, \\
g(a, b, c) &= c.
\end{align*}
\]

Then \( A \) has a unique nontrivial congruence \( 0^*_A \) corresponding to the partition \( \{a, c\}, \{b, d\} \), and \( A/0^*_A \) is isomorphic to a two element majority algebra. The congruence classes of \( 0^*_A \) are affine over \( \mathbb{Z}/2 \), and the algebra \( S = Sg_{\mathbb{A}^2}(\{(a, b), (b, a)\}) \) has a congruence \( \psi \) corresponding to the partition

\[
\{\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}, \begin{bmatrix} a \\ d \\ b \\ c \end{bmatrix}\}, \{\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}, \begin{bmatrix} a \\ b \\ d \\ c \end{bmatrix}\}.
\]

such that \( S/\psi \) is isomorphic to a two element affine algebra over \( \mathbb{Z}/2 \) (which is isomorphic to \( \{a, c\} \)).

In fact, we have an isomorphism \( S \cong A \times \{a, c\} \).

Then \( e \) acts like the majority operation \( g(x, y, z) \) on \( A/0^*_A \), acts like the minority operation \( g(x, u, y) \) on \( \{a, c\} \), and has

\[
\begin{align*}
e \begin{bmatrix} b & b & a & a \\ b & a & b & a \\ a & a & a & b \end{bmatrix} &= g \begin{bmatrix} b & a & a \\ a & b & a \\ a & a & a \end{bmatrix} = \begin{bmatrix} a \\ a \\ a \end{bmatrix}, \\
e \begin{bmatrix} d & d & a & a \\ d & a & d & a \\ a & a & a & d \end{bmatrix} &= g \begin{bmatrix} d & c & a \\ a & d & c \\ a & a & a \end{bmatrix} = \begin{bmatrix} a \\ a \\ a \end{bmatrix}.
\end{align*}
\]

Thus \( e \) is a 3-edge term.

Note that applying \( \sigma \) to the second coordinate of \( S \) turns it into \( 0^*_A \), and under the isomorphism \((1, \sigma) : S \cong 0^*_A \), one of the congruence classes of \( \psi \) becomes the diagonal \( \{(x, x) \mid x \in A\} \). Thus \( 0^*_A \) is the center of \( A \), and \( A \) is similar to the idempotent reduct of \( \mathbb{Z}/2 \). Since \( 1_A = (0_A : 0^*_A) \) is not abelian, we see that \( A \) can’t generate a residually small variety.

We can check that

\[
\begin{bmatrix} a & b \\ a & b \end{bmatrix} \notin Sg_{\mathbb{A}^{3 \times 2}} \left\{ \begin{bmatrix} a & b \\ b & a \end{bmatrix}, \begin{bmatrix} a & b \\ b & a \end{bmatrix}, \begin{bmatrix} b & a \\ a & b \end{bmatrix} \right\}
\]

by taking the rows modulo \( \psi \). Thus none of the subsets \( \{a, b\}, \{b, c\}, \{c, d\}, \{d, a\} \) (which are taken to each other by powers of the automorphism \( \sigma \)) are closed under any term which acts nontrivially on \( A/0^*_A \). Using this, one can show that \( \text{Clo}(g) \) does not contain any proper Taylor subclones.

What do critical relations on \( A \) look like? Suppose that \( R \leq_{sd} A^m \times \{a, c\}^n \) is critical and subdirect for some \( m, n \) with \( m + n \geq 3 \). By Theorem 13.4 \( R \) has the parallelogram property. All we can conclude from Theorem 13.10 is that \( R^* \) has the parallelogram property and has linking
congruence \((0^*_\mathbb{A})^m \times 1^n_{\{a,c\}}\), so the reduction \(\mathbb{R}_{\text{red}}^*\) of \(\mathbb{R}^*\) is a subdirect \(m\)-ary relation on the two element majority algebra \(\mathbb{A}/0^*_\mathbb{A}\) which has the parallelogram property.

Luckily, it turns out that any such \(\mathbb{R}_{\text{red}}^*\) has \(\pi_{ij}(\mathbb{R}_{\text{red}}^*)\) either a full relation or the graph of an automorphism of \(\mathbb{A}/0^*_\mathbb{A}\) for any \(i, j \in [m]\). This can be proved directly by reasoning about globally consistent instances of 2-SAT whose solution sets have the parallelogram property, or it can be proved more abstractly by using the fact that the two element majority algebra is subdirectly irreducible and generates a congruence distributive variety.

However we prove the claim about \(\mathbb{R}_{\text{red}}^*\), we see that if we assume without loss of generality that \((a, ..., a) \in \mathbb{R}\) (by applying powers of \(\sigma\) to coordinates of \(\mathbb{R}\)), then we can group the coordinates of \(\mathbb{R}\) into groups of size \(m_1, ..., m_k\),

\[
\mathbb{R} \leq_{sd} \mathbb{A}^{m_1} \times \cdots \times \mathbb{A}^{m_k} \times \{a, c\}^n,
\]

such that \(\pi_{ij}(\mathbb{R})\) is full for coordinates \(i, j\) coming from separate groups, and \(\pi_{ij}(\mathbb{R}) = 0^*_\mathbb{A}\) for coordinates \(i, j\) coming from the same group.

Since we have assumed \((a, ..., a) \in \mathbb{R}\), \(\mathbb{R}\) must be closed under the unary polynomial \(\phi : x \mapsto g(a, x, x)\). Since \(\phi(a) = \phi(c) = a\) and \(\phi(b) = \phi(d) = d\), we see that any vector of \(as\) and \(ds\) which is constant on each group of coordinates will be contained in \(\mathbb{R}\). From this we see that in fact, any piecewise-constant vector

\[
((x_1, ..., x_1), (x_2, ..., x_2), ..., (x_k, ..., x_k), (a, ..., a)) \in \mathbb{A}^{m_1} \times \cdots \times \mathbb{A}^{m_k} \times \{a, c\}^n
\]

must be contained in \(\mathbb{R}\). If we now consider the restriction \(\mathbb{R} \cap \{a, c\}^{m+n}\), then we find that it is an affine space defined by a system of linear equations over \(\mathbb{Z}/2\), where the number of coordinates from any single group which show up in any equation must be even, since we may swap \((a, ..., a), (c, ..., c) \in \mathbb{A}^{m_i}\) in any element of \(\mathbb{R}\). Thus we see that \(\mathbb{R}\) can be written as an intersection of relations \(\mathbb{R}'\) where the coordinates pair up in groups \(\{i, j\}\) of size two, such that \(\pi_{ij}(\mathbb{R}') = 0^*_\mathbb{A}\) and the relation \(\mathbb{R}'\) factors through the map \(0^*_\mathbb{A} \to \{a, c\}\) for each such pair of coordinates.

Using the above analysis, we see that the relational clone corresponding to \(\mathbb{A}\) is generated by the graph of the automorphism \(\sigma\), which is \(\text{Sg}_{\mathbb{A}^2}\{(a, b), (d, a)\}\), the critical binary relation \(\text{Sg}_{\mathbb{A}^2}\{(a, a), (a, b), (b, b)\}\), which corresponds to a partial order on the majority algebra \(\mathbb{A}/0^*_\mathbb{A}\), and the ternary relation \(\text{Sg}_{\mathbb{A}^3}\{(a, a, a), (a, c, c), (b, b, a)\}\), which is the graph of the homomorphism \(0^*_\mathbb{A} \to \{a, c\}\).

14 Learnability of relations encoded by compact representations

We'll start off by reviewing some of the standard definitions of learning theory.

**Definition 14.1.** Fix a universe \(U\). We call a collection \(\mathcal{C}\) of subsets of \(U\), together with a rule for encoding the elements of \(\mathcal{C}\), a concept class. An encoding of an element \(C \in \mathcal{C}\) is called a concept (from \(\mathcal{C}\)). The encoding scheme is called polynomially evaluable if there is an algorithm which takes an encoding of a concept \(C \in \mathcal{C}\) and an element \(x \in U\), and determines whether \(x \in U\) in polynomial time.

Generally we imagine a situation in which a teacher knows a target concept \(C \in \mathcal{C}\), and a student tries to learn the target concept \(C\) from the teacher, either by seeing (random) examples of elements in \(U\) and being told whether or not they are in the target concept \(C\), or by asking the
teacher certain types of questions. The teacher is modeled as an oracle which can be queried by the learner.

The main model which we will be examining in this section is the model of exact learning with (improper) equivalence queries from \([3]\). Learnability results in the equivalence query model can be converted directly into learnability results in the probably approximately correct model (which is often abbreviated as PAC-learning).

**Definition 14.2.** Let \(C'\) be a concept class which contains \(C\), and call \(C'\) the hypothesis class. We define an equivalence oracle \(O_C\) with target concept \(C \in C\) to be the function which takes as input a hypothesis \(C' \in C\), returns “true” if \(C = C'\), and otherwise returns an (arbitrary) element of the symmetric difference \(C \Delta C'\).

**Definition 14.3.** An algorithm which makes calls to an oracle \(O\) is said to learn the concept class \(C\) in the exact model with equivalence queries if, when the oracle \(O\) is the equivalence oracle \(O_C\) with target concept \(C \in C\), the algorithm makes finitely many calls to the oracle \(O\) with encodings of hypotheses \(C' \in C\) before finally discovering the concept \(C\). The learning algorithm is called proper if \(C' = C\), and improper otherwise. If there is an algorithm which learns \(C\) in time polynomial in \(\log |U|\), then we say that \(C\) is polynomially learnable.

We are interested in the case where the universe \(U\) is \(\mathbb{A}^n\) for \(n\) large and \(\mathbb{A}\) a fixed algebraic structure, and where the concept class \(C\) consists of the set of subalgebras of \(\mathbb{A}^n\), i.e. \(C = \text{Inv}_n(\mathbb{A})\) (recall \(\text{Inv}_n(\mathbb{A})\) is the set of \(n\)-ary relations which are preserved by the basic operations of \(\mathbb{A}\)). In order for polynomial (in \(n\)) length encodings of the concepts in \(C\) to exist, we need \(\log |C|\) to be bounded by a polynomial in \(n\), that is, we need \(\mathbb{A}\) to have few subpowers.

Suppose that \(\mathbb{A}\) has a \(k\)-edge term, and fix a particular \(k\)-edge term \(e\). In this case, \(n\)-ary relations on \(\mathbb{A}\) are naturally encoded by compact representations, so we will use compact representations as our encoding scheme for the concept class \(C = \text{Inv}_n(\mathbb{A})\).

For the sake of definiteness, we will slightly modify the definition of a compact representation \(R\) by requiring that for each element \(x_I\) of \(\pi_I(S)\) (where \(|I| < k\)), a specific element \(x \in R\) with \(\pi_I(x) = x_I\) has been marked (by \(x_I\)), and similarly for each minority index \((i, a, b)\) of \(R\), a particular ordered pair \((u_a, u_b) \in R^2\) witnessing this index has been marked (by \((i, a, b)\)). We will also require that each element \(x\) of the compact representation \(R\) is marked at least once (i.e., either \(x\) is part of a marked witness to a minority index of \(R\), or \(x\) is a marked witness for some element of a projection of \(R\) onto a small set of coordinates).

One feature which we would like this encoding scheme to satisfy is that there should be a polynomial time procedure to check whether an element \(a \in \mathbb{A}^n\) is contained in the relation \(R\) encoded by the compact representation \(R\). In other words, we want our encoding scheme to be polynomially evaluable. The next lemma can be used to show that our encoding scheme is polynomially evaluable. We use the notation \([i]\) for the set \(\{1, ..., i\}\).

**Lemma 14.4.** Suppose that \(R \subseteq \mathbb{A}^n\) is a compact representation of \(\mathbb{R} \leq \mathbb{A}^n\), \(i \leq n\), \(a \in \mathbb{A}^n\), \(b \in \mathbb{R}\) with \(\pi_{[i-1]}(a) = \pi_{[i-1]}(b)\), and set \(c_i = d(b_i, a_i)\). Suppose that

- for each \(I \subseteq [i]\) with \(|I| < k\) and \(i \in I\), the element \(x^I \in R\) is the marked element of \(R\) witnessing \(\pi_I(x^I) = \pi_I(a)\), and
- the pair \((u_a, u_c) \in R^2\) is the marked witness of the minority index \((i, a_i, c_i)\).
Then there is a term $t^{[i]}$ of $\Lambda$ which can be built out of the terms $e, s, p, d$ of Theorem 12.8 in time polynomial in $n$, such that $b^{[i]} = t^{[i]}(b, u_a, u_c, x^1, \ldots) \in \mathbb{R}$ satisfies $\pi_{[i]}(a) = \pi_{[i]}(b^{[i]})$.

Proof. The proof is a modification of the proof of Theorem 12.11 with the induction over subsets of $[i]$ modified to only involve polynomially many subsets of $[i]$. The trick is to consider sets $I$ of the form $[j] \cup J$, where $j \leq i$, $|J| = k - 1$, and $i \in J$. There are only polynomially many such sets $I$, and we can induct on $j$ to handle them.

So we will show by induction on $j$ that for every set $I = [j] \cup J$ with $|J| = k - 1$ and $i \in J$, there is a term $t^I$ such that $b^I = t^I(b, u_a, u_c, x^1, \ldots)$ satisfies $\pi_I(b^I) = \pi_I(a)$. The base case $j = 0$ is handled by taking $t^I = x^I$ for $|I| = k - 1$.

For the inductive step, note that if $I = [j] \cup J$, then we can also write $I = [j - 1] \cup ([j] \cup J)$. Let $\{j\} \cup J = \{l_1, \ldots, l_{k-1}, i\}$, and define sets $I_1, \ldots, I_{k-1}$ by deleting $l_1, \ldots, l_{k-1}$, respectively, from $I$, and note that each of the sets $I_m$ has the form $I_m = [j - 1] \cup J_m$, where $J_m = ([j] \cup J) \setminus \{l_m\}$ and $i \in J_m$. By the induction hypothesis, we have already constructed terms $t^{I_m}$ and corresponding elements $b^{I_m} \in \mathbb{R}$ with $\pi_{I_m}(b) = \pi_{I_m}(a)$. Then if we consider

$$s(b, b^{I_1}, \ldots, b^{I_{k-1}}),$$

we see that if we restrict to the coordinates in $\{j\} \cup J$, we have

$$s\left(\begin{bmatrix} a_{l_1} & b_{l_1}^{I_1} & a_{l_2} & \ldots & a_{l_2} \\ a_{l_2} & a_{l_2} & b_{l_2}^{I_2} & \ldots & a_{l_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{l_{k-1}} & a_{l_{k-1}} & a_{l_{k-1}} & \ldots & b_{l_{k-1}}^{I_{k-1}} \\ b_i & a_i & a_i & \ldots & a_i \end{bmatrix}\right) = \begin{bmatrix} a_{l_1} \\ a_{l_2} \\ \vdots \\ a_{l_{k-1}} \\ c_i \end{bmatrix}.$$

Additionally, if we consider

$$p(u_c, u_a, b^{I_1}),$$

then if we restrict to the coordinates in $\{j\} \cup J$, we have

$$p\left(\begin{bmatrix} u_{l_1} & u_{l_1} & b_{l_1}^{I_1} \\ u_{l_2} & u_{l_2} & a_{l_2} \\ \vdots & \vdots & \vdots \\ u_{l_{k-1}} & u_{l_{k-1}} & a_{l_{k-1}} \\ c_i & a_i & a_i \end{bmatrix}\right) = \begin{bmatrix} b_{l_1}^{I_1} \\ a_{l_2} \\ \vdots \\ a_{l_{k-1}} \\ c_i \end{bmatrix}.$$

Thus, if we take $t^I$ to be given by

$$t^I = e(p(u_c, u_a, t^{I_1}), s(b, t^{I_1}, \ldots, t^{I_{k-1}}), t^{I_1}, \ldots, t^{I_{k-1}}),$$

then when we restrict to the coordinates in $\{j\} \cup J$, we get

$$e\left(\begin{bmatrix} b_{l_1}^{I_1} & a_{l_1} & b_{l_1}^{I_1} & a_{l_1} & \ldots & a_{l_1} \\ a_{l_2} & a_{l_2} & a_{l_2} & b_{l_2}^{I_2} & \ldots & a_{l_2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{l_{k-1}} & a_{l_{k-1}} & a_{l_{k-1}} & a_{l_{k-1}} & \ldots & b_{l_{k-1}}^{I_{k-1}} \\ c_i & c_i & a_i & a_i & \ldots & a_i \end{bmatrix}\right) = \begin{bmatrix} a_{l_1} \\ a_{l_2} \\ \vdots \\ a_{l_{k-1}} \\ c_i \end{bmatrix},$$

which completes the induction step. \qed
We can now check if an element \( a \in \mathbb{A}^n \) is in the relation encoded by the compact representation \( R \) as follows. First we check that \( \pi_I(a) \in \pi_I(S) \) for each \( I \) with \( |I| < k \), and let \( x^I \) be the marked element of \( R \) with \( \pi_I(x^I) = \pi_I(a) \). Then we try to construct elements \( b^{[i]} \in \text{Sg}_{\mathbb{A}^n}(R) \) iteratively with \( \pi_I(b^{[i]}) = \pi_I(a) \). We start by taking \( b^{[k-1]} = x^{[k-1]} \), and repeatedly invoke Lemma 14.4 to see that if \( (i,a_i,c_i) \in \text{Sig}(R) \), where \( c_i = d(b^{[i-1]},a_i) \), then we can construct \( b^{[i]} \in \text{Sg}_{\mathbb{A}^n}(R) \) in polynomial time. We formalize this procedure as a subroutine which I will call \text{Approximate}(R,a) \) (this is almost the same as the combination of the subroutines \text{Interpolate} and \text{New-Fix-values} from [69]).

**Algorithm 10 Approximate** \((R,a), \varepsilon,s,p,d\) terms as in Theorem 12.8 \( R \subseteq \mathbb{A}^n \) a compact representation such that \( \pi_I(a) \in \pi_I(R) \) for all \( I \) with \( |I| < k \).

1: for all \( I \subseteq [n] \) with \( |I| < k \) do  
2: \quad Let \( x^I \) be the marked element of \( R \) with \( \pi_I(x^I) = \pi_I(a) \).  
3: \quad Set \( b^{[k-1] \cap [n]} = x^{[k-1] \cap [n]} \).  
4: for \( i \) from \( k \) to \( n \) do  
5: \quad Set \( c_i \leftarrow d(b^{[i-1]},a_i) \).  
6: if \( (i,a_i,c_i) \notin \text{Sig}(R) \) then  
7: \quad return \( b^{[i-1]} \).  
8: else  
9: \quad Let \((u^i_a,u^i_c)\) be the marked witness of the minority index \((i,a_i,c_i)\) in \( R \).  
10: for \( j \) from 1 to \( i-k+1 \) do  
11: \quad for all \( l_1,...,l_{k-1} \) with \( j = l_1 < l_2 < \cdots < l_{k-1} < i \) do  
12: \quad \quad Set \( I \leftarrow [j] \cup \{l_2,...,l_{k-1},i\} \).  
13: \quad \quad for \( m \) from 1 to \( k-1 \) do  
14: \quad \quad \quad Set \( I_m \leftarrow I \setminus \{l_m\} \).  
15: \quad \quad Set \( b^l \leftarrow e(p(u^i_a,u^i_c),b^{[i-1]}),s(b^{[i-1]},b^1,...,b^{k-1}),b^1,...,b^{k-1}) \).  
16: return \( b^{[n]} \).

The running time of \text{Approximate} is \( O(n^{k+1}) \): there are less than \( n^k \) choices for \( j = l_1 < \cdots < l_{k-1} < i \), and for each choice, computing the new \( b^l \) takes \( O(n) \) steps (since \( b^l \) has \( n \) coordinates). By only maintaining the values of \( b^{[i-1]} \) and \( b^l \) with \( i \in I \) in the \( i \)th step through the outer loop, the memory required is reduced to \( O(n^k) \), which is the same as the space required to store a typical compact representation \( R \).

**Proposition 14.5.** If \( R \subseteq \mathbb{A}^n \) is a compact representation and \( a \in \mathbb{A}^n \) with \( \pi_I(a) \in \pi_I(R) \) for all \( I \) with \( |I| < k \), then either \text{Approximate}(R,a) \) returns \( a \in \text{Sg}_{\mathbb{A}^n}(R) \), or \text{Approximate}(R,a) \) returns \( b \neq a \) such that \( b \in \text{Sg}_{\mathbb{A}^n}(R) \), and such that if \( i \) is minimal with \( b_i \neq a_i \), then the minority index \((i,a_i,d(b_i,a_i))\) is not witnessed in \( R \).

At this point everything seems wonderful, but there is one major wrinkle: we have no idea how to (efficiently) test whether a given “compact representation” \( R \subseteq \mathbb{A}^n \) is actually a compact representation of the subalgebra \( \mathbb{R} = \text{Sg}_{\mathbb{A}^n}(R) \) it generates - in other words, we don’t know how to test whether \( R \) is a valid encoding of a concept from the concept class \( \mathcal{C} = \text{Inv}_{\mathbb{A}^n}(\mathbb{A}) \). While it’s easy to test whether \( R \) and \( \mathbb{R} \) have the same projections onto small subsets of the coordinates (just check whether \( \pi_I(R) \) is closed under the operations of \( \mathbb{A} \) for all \( I \) with \( |I| < k \)), what is missing is a way to test whether \( R \) witnesses every minority index which is witnessed in \( \mathbb{R} \).

110
Let’s think for a moment about the problem of checking whether \( R \) and \( \text{Sg}(R) \) witness the same minority indices. Since there are only \( n!|A|^2 \) possible minority indices, we may as well focus on one particular minority index \((i, a, b)\). By replacing \( R \) with \( \pi_\lfloor i \rfloor(R) \) and \( n \) with \( i \), we may reduce to the case \( i = n \).

**Proposition 14.6.** Suppose that the minority index \((i, a, b)\) is witnessed by some pair \((u_a, u_b)\) in a relation \( R \subseteq A^n \). Then for any tuple \( t_a \in R \) with \( \pi_i(t_a) = a \), there is a tuple \( t_b \in R \) such that the pair \((t_a, t_b)\) also witnesses the minority index \((i, a, b)\). If \( i = n \), then \( t_b \) is uniquely determined by \( t_a \).

**Proof.** Take \( t_b = p(u_b, u_a, t_a) \). Then the identity \( p(y, y, x) \approx x \) implies that \( \pi_{i-1}(t_b) = \pi_{i-1}(t_a) \), and the fact that \( a, b \) are a minority pair (that is, that \( d(b, a) = b \)) and the identity \( p(x, y, y) \approx d(x, y) \) imply that \( \pi_i(t_b) = p(b, a, a) = b \).

So we can check whether a minority index \((i, a, b)\) is witnessed by \( \text{Sg}(R) \) as follows. First we pick any tuple \( t_a \in R \) with \( \pi_i(t_a) = a \). Then we modify it to make a tuple \( t_b \) by replacing the \( i \)th coordinate with \( b \). Finally, we check whether \( \pi_i(t_b) \in \text{Sg}(\pi_i(R)) \). By the above results, we have \( \pi_i(t_b) \in \text{Sg}(\pi_i(R)) \) if and only if \((i, a, b) \in \text{Sig}(\text{Sg}(R)) \). We find ourselves naturally led to consider the subpower membership problem.

**Problem 14.1** (Subpower Membership Problem). Given a finite subset \( S \subseteq A^n \) and an element \( x \in A^n \), determine if \( x \) is in the subalgebra of \( A^n \) generated by \( S \).

**Theorem 14.7** (Bulatov, Mayr, Szendrei [11]). For a fixed finite algebra \( A \) with few subpowers, the following problems are polynomial time reducible to eachother:

- *The subpower membership problem for \( A \):* determine if \( x \in \text{Sg}_{A^n}(S) \), given \( x \in A^n \) and \( S \subseteq A^n \).

- *Find a compact representation for \( \text{Sg}_{A^n}(S) \), given a subset \( S \subseteq A^n \).*

- *The subpower intersection problem for \( A \):* given subsets \( R, S \subseteq A^n \), find a set of generators for \( \text{Sg}_{A^n}(R) \cap \text{Sg}_{A^n}(S) \).

If \( A \) has few subpowers and has a finite number of basic operations, then the subpower membership problem for \( A \) is in NP.

**Proof.** Left as an exercise to the reader. The hardest bit is the claim that the subpower membership problem is in NP: for this, imagine that we have a set \( R \) which looks like a compact representation, and consider the set \( C \) of all \( a \in A^n \) such that \( \text{Approximate}(R, a) \) returns \( a \). If \( C \) is not closed under the basic operations of \( A \), then there should be a witness to the fact that \( C \) is not closed, and a nondeterministic algorithm can guess such a witness, verify that it works, and use it to enlarge \( R \).

Unfortunately, whether the subpower membership problem is in P for algebras with few subpowers is currently an open problem (even in the special case of quasigroups). So we need to find a workaround for this issue.

The workaround is to enlarge the concept class \( C = \text{Inv}_n(A) \) to a larger concept class \( C' \), where concepts in \( C' \) are encoded by “compact representations” \( R \subseteq A^n \), where we allow sets \( R \) which are not compact representations of the subalgebra \( \mathbb{R} = \text{Sg}_{A^n}(R) \) which they generate. In order to be precise about exactly what concept \( C \) is encoded by \( R \), we use the \text{Approximate} subroutine.
Definition 14.8. If \( R \subseteq A^n \) is a “compact representation”, then the corresponding concept \( C \subseteq A^n \) encoded by \( R \) is defined by the following rule. An element \( a \in A^n \) is in \( C \) iff the following two conditions are satisfied:

- for every \( I \subseteq [n] \) with \(|I| < k\), we have \( \pi_I(a) \in \pi_I(R) \), and
- the subroutine \texttt{Approximate}(\( R, a \)) returns \( a \).

The penalty we will pay for this workaround is that since the new concept class \( C' \) is larger than \( C = \text{Inv}_n(\mathcal{A}) \), our learning algorithm will now be making improper equivalence queries. If the subpower membership problem for \( \mathcal{A} \) can be proved to be in \( P \), then we will be able to upgrade to a learning algorithm which makes only proper equivalence queries.

Now we can finally describe the learning algorithm, which is remarkably simple.

Algorithm 11 \texttt{Learn}(O), \( O \) an equivalence oracle for an unknown target concept \( C \in \text{Inv}_n(\mathcal{A}) \).

1: Set \( R \leftarrow \emptyset \).
2: \textbf{while} \( O(R) \) does not return “true” \textbf{do}
3: \hspace{1em} Set \( a \leftarrow O(R) \).
4: \hspace{2em} for all \( I \subseteq [n] \) with \(|I| < k\) such that \( \pi_I(a) \) has no designated witness in \( R \) do
5: \hspace{3em} Set \( R \leftarrow R \cup \{a\} \).
6: \hspace{2em} Mark \( a \) as the designated witness for \( \pi_I(a) \).
7: \hspace{2em} while \texttt{Approximate}(\( R, a \)) does not return \( a \) \textbf{do}
8: \hspace{3em} Set \( b \leftarrow \texttt{Approximate}(\( R, a \)) \).
9: \hspace{4em} Let \( i \) be minimal such that \( a_i \neq b_i \).
10: \hspace{3em} Set \( R \leftarrow R \cup \{a, d(b, a)\} \).
11: \hspace{2em} Mark the pair \( (a, d(b, a)) \) as the designated witness for the minority index \( (i, a_i, d(b_i, a_i)) \).
12: Optionally, enlarge \( R \) further to make it closer to a compact representation of \( \text{Sg}(R) \).

Proposition 14.9. For a fixed algebra \( \mathcal{A} \) with few subpowers, the algorithm \texttt{Learn}(O) takes time polynomial in \( n \) to find an encoding \( R \) for the target concept \( C \in \text{Inv}_n(\mathcal{A}) \).

Proof. At every step of the algorithm, we have \( \text{Sg}(R) \subseteq C \): this is true at the beginning, and if it is true before we call \( O(R) \), then since the concept \( C' \) encoded by \( R \) has \( C' \subseteq \text{Sg}(R) \subseteq C \), the value \( a \) returned by \( O(R) \) will be contained in \( C \Delta C' = C \setminus C' \subseteq C \), so \( \text{Sg}(R \cup \{a\}) \subseteq C \).

Furthermore, every time we process a new \( a \in C \setminus C' \), we strictly enlarge \( R \) to make the new concept encoded by \( R \) contain \( a \), either by adding \( a \) as a designated witness to \( \pi_I(a) \in \pi_I(R) \) for some \( I \), or by adding new minority indices which were not present in the original \( R \). Since \( R \) can only increase in size polynomially many times (as a compact representation has size bounded by a polynomial in \( n \)), we can only call the oracle polynomially many times before the process must terminate.

Remark 14.1. If we did not insist on polynomial evaluability of the encoding scheme (or if we could solve the subpower membership problem), then we could instead encode relations via generating sets. The learning algorithm would then be even simpler: at every step, the learner guesses that the target concept is the relation generated by all the examples it has seen so far. This learning algorithm is known as the \textit{closure algorithm}. The issue is that now the equivalence oracle becomes
hard to implement, as the teacher is forced to determine whether a given set generates the target relation they have in mind.

Now we will explain how all of this relates to Valiant’s PAC-learning model [108]. In the PAC-learning model, the teacher (oracle) has access to both a target concept $C \in \mathcal{C}$ and a probability distribution $\mu$ over the universe $U$, both of which are unknown to the learner. The learner is allowed to request random classified examples, sampled from the distribution $\mu$ (by a “classified” example, I mean that the learner is given an example and told whether or not it is in the target concept $C$).

**Definition 14.10.** If $C \in \mathcal{C}$ is a target concept and $\mu$ is a probability distribution on the universe $U$, then the **sampling oracle** for the pair $C, \mu$ is a randomized oracle which samples a random element $a \in U$ drawn from the distribution $\mu$, and returns the ordered pair $(a, a \in C)$, where by “$a \in C$” we mean either “true” or “false” based on whether $a$ is in the target concept $C$.

In the PAC-learning model, the goal of a learning algorithm is to output an encoding of a concept $C'$ in the hypothesis class $\mathcal{C}'$, such that the $\mu$-measure $\mu(C \Delta C')$ of the symmetric difference between $C$ and $C'$ is small. We can’t hope to do better than this, since the chance of seeing an example which lets us distinguish between $C$ and $C'$ is at most $\mu(C \Delta C')$ times the number of classified examples we request.

**Definition 14.11.** We say that an algorithm with access to a sampling oracle learns a concept class $\mathcal{C}$ in the probably approximately correct model with error $\epsilon$ and confidence $1 - \delta$ if for any target concept $C \in \mathcal{C}$ and any probability distribution $\mu$ over the universe, the algorithm eventually returns a hypothesis $C' \in \mathcal{C}'$ such that

$$\Pr[\mu(C \Delta C') \leq \epsilon] \geq 1 - \delta.$$ 

The probability here is taken over the random choices made by the oracle (and possibly the learning algorithm) - the target concept $C$ is *not* being randomized here, we require this for all $C \in \mathcal{C}$ and all $\mu$.

We say that a concept class $\mathcal{C}$ is **efficiently PAC-learnable** if there is an algorithm which learns $\mathcal{C}$ in the PAC-model and takes time polynomial in $\log(|U|)$, $\frac{1}{\epsilon}$, and $\log(\frac{1}{\delta})$ for $\epsilon, \delta > 0$.

The standard learning algorithm in the PAC model is to request a large number of classified examples, and then choose any hypothesis $C' \in \mathcal{C}'$ which is consistent with all of the classified examples we have seen so far. For this to work, it is necessary that the hypothesis class $\mathcal{C}'$ is in some sense “small”, and we also need to have a way to efficiently find at least one hypothesis which is consistent with the examples. First we will define a measure of the “size” of the concept class, known as the VC-dimension.

**Definition 14.12.** If $\mathcal{C}$ is a collection of subsets of some universe $U$, then we say that a set $S$ is **shattered by $\mathcal{C}$** if for all $X \subseteq S$, there is some $C \in \mathcal{C}$ with $C \cap S = X$. We define the **Vapnik-Chervonenkis dimension of $\mathcal{C}$**, written $\text{VC}(\mathcal{C})$, to be the size of the largest set $S$ which is shattered by $\mathcal{C}$.

To see that the VC-dimension is a good measure of the complexity of a concept class, we recall the Sauer-Shelah Lemma.
Lemma 14.13 (Sauer-Shelah Lemma). If $C$ is a collection of subsets of $U$ with VC-dimension $d$, then
\[
|C| \leq \sum_{i=0}^{d} \binom{|U|}{i}.
\]
In fact, we have the stronger result that the number of sets $S \subseteq U$ which are shattered by $C$ is at least $|C|$.

Proof. We show that $C$ shatters at least $|C|$ sets by induction on $|C|$. For the base case, note that the empty set is shattered by $C$ as long as $|C| \geq 1$. For the inductive step, let $x \in U$ be an element which is in some of the sets in $C$ but not all of them, and let $C_x$ be the collection of $C \in C$ with $x \in C$ and $C'_x = C \setminus C_x$. Inductively, $C_x$ shatters at least $|C_x|$ sets and $C'_x$ shatters at least $|C'_x|$ sets, and any set shattered by $C_x$ or $C'_x$ must not contain $x$.

To finish the induction, we just need to check that for any set $S$ which is shattered by both $C_x$ and $C'_x$, the set $S \cup \{x\}$ is shattered by $C$. \hfill \Box

If the set $S$ is shattered by $C$, then the sampling oracle could sample from a uniform distribution on $S$, and in this case the learner is faced with the problem of learning an arbitrary subset $X = C \cap S$ of $S$ given an oracle which returns uniformly random classified examples. If the learner examines $o(|S|)$ classified examples, then clearly they can’t hope to succeed. The following result makes this precise.

Proposition 14.14. If an algorithm learns a concept class $C$ with error $\epsilon$ and confidence $1 - \delta$ after requesting at most $m$ classified examples, then
\[
m \geq (2(1 - \epsilon)(1 - \delta) - 1) \text{VC}(C).
\]

Proof. Let $S$ be a set with $|S| = \text{VC}(C)$ which is shattered by $C$, and for each $X \subseteq S$ consider the sampling oracle $O_X$ which samples from the uniform distribution $\mu$ on $S$, and has target concept some $C_X \in C$ with $C_X \cap S = X$. If we average the performance of the learning algorithm over the sampling oracles $O_X$ (with $X$ chosen as a uniformly random subset of $S$), we see that if it outputs a hypothesis $C'$, then
\[
\mathbb{E}[\mu(C_X \Delta C')] \geq \frac{1}{2} \left( 1 - \frac{m}{|S|} \right).
\]
By Markov’s inequality, this implies that
\[
(1 - \epsilon)\mathbb{P}[\mu(C_X \Delta C') \leq \epsilon] \leq \frac{1}{2} + \frac{m}{2|S|},
\]
so
\[
\frac{1}{2} + \frac{m}{2|S|} \geq (1 - \epsilon)(1 - \delta). \hfill \Box
\]

Conversely, if the VC-dimension of $C$ is small, then the standard learning algorithm in the PAC model performs well, so long as it can be implemented.

Theorem 14.15 (VC-dimension determines sample-complexity [25]). If $\text{VC}(C') = d$, then any algorithm which takes
\[
m \geq \max \left( \frac{4}{\epsilon} \log \left( \frac{2}{\delta} \right), \frac{8d}{\epsilon} \log \left( \frac{13}{\epsilon} \right) \right).
\]
samples from a sampling oracle and outputs any hypothesis $C' \in C'$ consistent with the data will learn $C$ with error $\epsilon$ and confidence $1 - \delta$. 114
Sketch. Consider the following process: pick 2m samples from the probability distribution $\mu$, permute them randomly, feed the first m samples (after permuting) to the learning algorithm, and count how many of the last m samples are classified incorrectly by the hypothesis $C'$ chosen by the learning algorithm.

If the learning algorithm fails to learn $C$ with error $\epsilon$ and confidence $1 - \delta$, then for some choice of target concept $C$ and distribution $\mu$, the process described will incorrectly classify at least $\frac{\epsilon m}{2}$ of the last m samples with probability at least $\delta$, by Chebyshev’s inequality (at least for $m \geq \frac{\delta}{\epsilon}$). Thus there will be some specific set $X$ of size $2m$, such that at least a $\frac{\delta}{2}$ fraction of its permutations lead to an incorrect classification of at least $\frac{\epsilon m}{2}$ of its last m elements.

By the Sauer-Shelah Lemma [14.13], the number of distinct subsets of $X$ which can be written as $C' \cap X$ for some $C' \in C'$ is bounded by $\sum_{i \leq d} \binom{2m}{i}$. For each possible intersection $C' \cap X$, the chance of the first m samples from $X$ being consistent with $C'$ and the last m samples from $X$ having at least $\frac{\epsilon m}{2}$ inconsistencies with $C'$ is at most $2^{-\epsilon m/2}$. Thus if the learning algorithm fails, then by the union bound we must have

$$2^{-\epsilon m/2} \sum_{i \leq d} \binom{2m}{i} \geq \frac{\delta}{2},$$

and plugging in the assumed bounds on $m$ and chugging through the inequalities gives a contradiction.

Note that if $A$ is an algebraic structure and we take $U = A^n$, $C = \text{Inv}_n(A)$, then a set $S$ is shattered by $\text{Inv}_n(A)$ iff $S$ is an independent subset of $A^n$. Thus the VC-dimension of $\text{Inv}_n(A)$ is exactly the same thing as the number $i_A(n)$, so if the concept classes $\text{Inv}_n(A)$ are efficiently PAC-learnable for all $n$, then $A$ must have few subpowers.

We can convert learning algorithms in the equivalence query model into learning algorithms in the PAC model by using the sampling oracle to simulate an equivalence oracle.

**Proposition 14.16 (Angluin [3]).** If a concept class $C$ is efficiently learnable in the (improper) equivalence query model using a hypothesis class $C'$ which has a polynomially evaluable encoding scheme, then $C$ is also efficiently learnable in the PAC model.

**Proof.** Given a sampling oracle $O$, we simulate an equivalence oracle as follows. The $i$th time the equivalence oracle is called by the learner, say to determine whether the target concept $C$ is equivalent to a hypothesis $C' \in C'$, we call the sampling oracle $O$ some number $q_i$ times to get $q_i$ random classified examples, and we check whether the way they are classified agrees with the hypothesis $C'$ (here is where we are using polynomial evaluability). If their classifications do agree with $C'$, then we pretend that the equivalence oracle returned “true”, and otherwise we pick one of the examples $a$ whose classification does not agree with $C'$ and return $a$ as the counterexample in $C \Delta C'$.

By the union bound, the probability that the simulated equivalence oracle ever returns “true” for a hypothesis $C'$ with $\mu(C \Delta C') \geq \epsilon$ is at most

$$\sum_i (1 - \mu(C \Delta C'))^{q_i} \leq \sum_i (1 - \epsilon)^{q_i}.$$

If we take

$$q_i \geq \frac{1}{\epsilon} (\ln(1/\delta) + i \ln(2)),$$

115
for instance, then we get
\[
\sum (1 - \epsilon) q_i \leq \sum e^{-\epsilon q_i} \leq \sum e^{\ln \delta - i \ln(2)} = \sum \frac{\delta}{2^i} \leq \delta.
\]

Remark 14.2. Another learning model is the on-line learning model described by Littlestone [86]. In this model, the learner is repeatedly presented with examples, and for each example must guess its classification before being told whether its guess is correct. The goal of the learner is to have an upper bound on the number of incorrect guesses it makes, even if the sequence of examples is chosen adversarially. It is easy to convert a learnability result in the (improper) equivalence query model into an algorithm for on-line learning.

Remark 14.3. There is a variant of the PAC learning model in which the learner is also allowed to use membership queries: in a membership query, the learner picks an element \( x \in U \), and asks the teacher (oracle) whether \( x \) is in the target concept. In [2], several situations are given where the addition of membership queries can be shown not to help with learning, under some standard cryptographic assumptions. In [30], there is a claim that some of the impossibility results for PAC learning of \( \text{Inv}_n(\mathbb{A}) \) when \( \mathbb{A} \) doesn’t have few subpowers can be generalized to impossibility results in the model of PAC learning with membership queries (under cryptographic assumptions), but the exact statement and the proof are left to a “full version” of the paper which I have been unable to track down.

15 Algebras with few subpowers are finitely related

Suppose a clone \( \mathcal{O} \) on a finite domain \( A \) has a \( k \)-edge term \( e \). We want to show that there exists some finite set of relations \( R_1, ..., R_m \) which generate the relational clone which is dual to \( \mathcal{O} \). This is equivalent to \( \mathcal{O} \) being exactly the set of operations \( \text{Pol}(R_1, ..., R_m) \) which preserve the relations \( R_1, ..., R_m \). If \( R_1, ..., R_m \) are all preserved by \( \mathcal{O} \), then the clone \( \text{Pol}(R_1, ..., R_m) \) will certainly contain \( \mathcal{O} \), but might end up being too large. In this case, \( \text{Pol}(R_1, ..., R_m) \) will still contain the \( k \)-edge term \( e \), and we can use this to our advantage.

To understand the structure of a clone \( \mathcal{O} \) with a \( k \)-edge term, we go back to the explicit representation of the set \( \mathcal{O}_n \) of \( n \)-ary operations of \( \mathcal{O} \) as the free algebra over \( \mathbb{A} = (A, \mathcal{O}) \) on \( n \) generators, which is concretely given by the subalgebra
\[
\mathcal{O}_n = \mathcal{F}_k(x_1, ..., x_n) \leq \mathbb{A}^n
\]
genereated by the elements \( \pi_i : \mathbb{A}^n \to \mathbb{A} \) given by \( \pi_i(a_1, ..., a_n) = a_i \), where \( x_i \in \mathcal{F}_k(x_1, ..., x_n) \) is identified with the element \( \pi_i \in \mathbb{A}^{\mathbb{A}^n} \). Similarly, recall that the set of \( n \)-ary operations \( f \in \text{Pol}_n(R_1, ..., R_m) \), considered as a subalgebra of \( \mathbb{A}^{\mathbb{A}^n} \), is given by the primitive positive formula
\[
f \in \text{Pol}_n(R_1, ..., R_m) \iff \bigwedge_{i \leq m} \bigwedge_{M \in R_i^n} f(M) \in R_i.
\]

To check that these two subalgebras of \( \mathbb{A}^{\mathbb{A}^n} \) are equal, by Theorem 12.11 and the fact that one is contained within the other, it suffices to check that they have the same projections onto subsets \( I \subseteq \mathbb{A}^n \) of the coordinates with \( |I| < k \), and to check that they have the same forks. If \( R_1, ..., R_m \) generate all relations of \( \text{Inv}(\mathbb{A}) \) with arity less than \( k \), then the first condition will be satisfied. The hard part is dealing with the forks.
In order to make precise statements about the set of forks in \( A^n \), we first need to choose an ordering on the coordinates of \( A^n \), that is, an ordering on the elements of \( A^n \). A natural choice is to first fix any total order \( \leq \) on the set \( A \), and to extend this to the lexicographic order on \( A^n \).

**Definition 15.1.** If \((A, \leq)\) is a set with a total order, then we define the lexicographic order \( \leq_{\text{lex}} \) on \( A^n \) by \( a \leq_{\text{lex}} b \) iff either \( a = b \) or there is some \( i \leq n \) such that \( a_j = b_j \) for \( j < i \) and \( a_i < b_i \). In other words, \( a <_{\text{lex}} b \) if \( a_i < b_i \) at the first coordinate \( i \) where \( a \) and \( b \) differ.

**Definition 15.2.** If \((I, \leq)\) is a totally ordered set and \( R \subseteq A^I \) is a relation on \( A \), then for \( i \in I \) we define the set of forks of \( R \) at the \( i \)th coordinate to be the set of pairs \((a, b) \in A^2\) given by

\[
\text{Forks}(R, i) := \{(a, b) \mid \exists t_a, t_b \in R, \pi_{\leq i}(t_a) = \pi_{\leq i}(t_b), \pi_i(t_a) = a, \pi_i(t_b) = b\}.
\]

So in order to understand a clone \( O \) with a \( k \)-edge term, we need to understand the relations of arity less than \( k \), together with the set of forks \( \text{Forks}(O_n, a) \) for all \( a \in A^n \) and all \( n \). The issue is that while each set \( \text{Forks}(O_n, a) \) is given by a finite collection of pairs of elements, there are infinitely many elements \( a \in A^n \), \( n \in \mathbb{N}^+ \) to consider. So we need a way to relate \( \text{Forks}(O_n, a) \) to \( \text{Forks}(O_m, b) \) for some choices of \( a \in A^n \), \( b \in A^m \).

**Proposition 15.3.** Suppose \( a \in A^n, b \in A^m \). If there is a map \( \phi : [m] \to [n] \) such that the associated function \( \phi^* : A^n \to A^m \) given by \( \phi^*(x_1, \ldots, x_n) = (x_{\phi(1)}, \ldots, x_{\phi(m)}) \) satisfies the conditions

- \( \phi^*(a) = b \) and
- for all \( c <_{\text{lex}} a \) we have \( \phi^*(c) <_{\text{lex}} b \),

then for any clone \( O \) on \( A \), we have \( \text{Forks}(O_m, b) \subseteq \text{Forks}(O_n, a) \).

**Proof.** Letting \( \mathbb{A} = (A, O) \), \( \phi \) induces a natural map of free algebras \( O_m \to O_n \) given by \( x_i \mapsto x_{\phi(i)} \). We will write this natural map as \( f \mapsto f_\phi \). Considering \( O_m, O_n \) as subalgebras of \( \mathbb{A}^m, \mathbb{A}^n \), respectively, we see that for \( c \in A^n \) and \( f \in O_m \), the \( i \)th coordinate of the image \( f_\phi \) of \( f \) under this map is given by

\[
\pi_{\leq i}(f_\phi(c)) = \pi_{\leq i}(f_\phi(c)).
\]

In particular, if \( t, t' \in O_m \) with \( \pi_{\leq i}(t) = \pi_{\leq i}(t') \), then

\[
\pi_{\leq i}(t_\phi) = \pi_{\leq i}(t'_\phi),
\]

and

\[
\begin{bmatrix}
  \ell_\phi(a)
  \\
  \ell'_\phi(a)
\end{bmatrix}
= \begin{bmatrix}
  \ell(\phi^*(a))
  \\
  \ell'(\phi^*(a))
\end{bmatrix}
= \begin{bmatrix}
  \ell(b)
  \\
  \ell'(b)
\end{bmatrix},
\]

so every fork in \( \text{Forks}(O_m, b) \) is also a fork in \( \text{Forks}(O_n, a) \). \( \square \)

**Proposition 15.4.** A map \( \phi \) as in the previous proposition exists if there is a strictly increasing function \( h : [n] \to [m] \) such that

- the same elements of \( A \) occur in both \( a \) and \( b \),
- \( h^*(b) = a \), that is, \( a_i = b_{h(i)} \) for all \( i \in [n] \), and
for all \( s \in A \), if the index of the first occurrence of \( s \) in \( a \) is \( i \), then \( h(i) \) is the index of the first occurrence of \( s \) in \( b \).

If no coordinate \( a_i \) of \( a \) is minimal or maximal with respect to the order \(<\) on \( A \), then the converse is true: such a \( \phi \) exists iff such an \( h \) exists.

**Proof.** Given such an \( h \), we define \( \phi \) as follows. We set \( \phi(h(i)) = i \), and for \( j \) not in the image of \( h \) let \( \phi(j) \) be the first index \( i \) such that \( a_i = b_j \), so that \( h(\phi(j)) \leq j \) for all \( j \). Then for any \( c <_{\text{lex}} a \), if \( i \) is the first index where \( a_i \neq c_i \), and if \( j \) is the first coordinate where \( \phi^*(a) \) and \( \phi^*(c) \) differ, then we have \( a_{\phi(j)} \neq c_{\phi(j)} \), so \( i \leq \phi(j) \), so \( h(i) \leq \phi(j) \leq j \), so we must have \( h(i) = j \) since \( \phi^*(a) \) and \( \phi^*(c) \) also differ at \( h(i) \). Thus \( \phi^*(c) <_{\text{lex}} \phi^*(a) = b \).

Now suppose that no coordinate \( a_i \) of \( a \) is minimal or maximal with respect to the order \(<\) on \( A \). Then the map \( \phi \) in the previous proposition must be surjective: if \( i \) is not in the image of \( \phi \), then we can define \( c <_{\text{lex}} a \) which only differs from \( a \) on the \( i \)th coordinate, and \( \phi^*(c) = \phi^*(a) \neq_{\text{lex}} b \), contradicting the choice of \( \phi \). Thus we can define \( h : [n] \to [m] \) by

\[
  h(i) = \min\{j \in [m] \mid \phi(j) = i\},
\]

so \( \phi^*(b) = a \) and we see that \( a \) and \( b \) have the same set of symbols.

For any \( i \), if we define \( c <_{\text{lex}} a \) which matches \( a \) up to the \( i \)th coordinate, has \( c_i < a_i \), and \( c_j > a_j \) for all \( j > i \), then from \( \phi^*(c) < b = \phi^*(a) \), we see that \( h(i) < h(j) \) for all \( j > i \). Thus \( h \) must be strictly increasing. Finally, from the definition of \( h \), we see that if \( i \) is the index of the first occurrence of \( s \) in \( a \), then \( h(i) \) must be the index of the first occurrence of \( s \) in \( b \).

**Definition 15.5.** Let \( A^+ = \bigcup_{n \geq 1} A^n \), and define the partial order \( \leq_E \) on \( A^+ \) by \( a \leq_E b \) iff there exists a map \( h \) as in the previous proposition. Equivalently, \( a \leq_E b \) iff the same set of elements of \( A \) occur in \( a \) and \( b \), and \( b \) can be formed from \( a \) by inserting elements \( s \in A \) after their first occurrences in \( a \).

Note that the partial ordering \( \leq_E \) on \( A^+ \) has no dependence on the arbitrary choice of ordering \(<\) we introduced on the elements of \( A \) (of course, the set \( \text{Forks}(O, a) \) still depends on the choice of \(<\)). The partial order \( \leq_E \) is a refinement of the embeddability partial ordering that occurs in Higman’s Lemma [64]. We can now simplify the description of the sets \( \text{Forks}(O, a) \) using the ordering \( \leq_E \).

**Definition 15.6.** For any pair \((c, d) \in A^2\), we define the set \( \lambda(O, (c, d)) \subseteq A^+ \) to be the set of \( a \in A^+ \) such that \( (c, d) \notin \text{Forks}(O, a) \).

**Corollary 15.7.** For any clone \( O \) on a set \( A \), the set \( \lambda(O, (c, d)) \) is upwards closed in \( A^+ \) with respect to \( \leq_E \), that is, if \( a \in \lambda(O, (c, d)) \) and \( a \leq_E b \), then \( b \in \lambda(O, (c, d)) \).

To describe an upwards closed subset (also called an *upset*) of a *finite* poset, it is enough to describe its minimal elements. We want to show that \( \lambda(O, (c, d)) \) can be described in terms of its minimal elements, but for this to work, it’s necessary to show that \((A^+, \leq_E)\) is a *well* partial order.

**Definition 15.8.** A partial order \((X, \leq)\) is a *well partial order* if for every infinite sequence \( x_1, x_2, \ldots \) of elements of \( X \), there exists an infinite increasing subsequence \( i_1 < i_2 < \cdots \) such that \( x_{i_1} \leq x_{i_2} \leq \cdots \).
Proposition 15.9. A partial order \((X, \leq)\) is a well partial order iff it has no infinite descending chains and no infinite antichains.

Proof. Let \(x_1, x_2, \ldots\) be any infinite sequence of elements of \(X\). Color the edges of the complete graph on \(\mathbb{N}^+\) with three colors, as follows: for \(i < j\), the edge \(\{i, j\}\) is colored red if \(x_i > x_j\), colored blue if \(x_i, x_j\) are incomparable, and colored green if \(x_i \leq x_j\). By Ramsey’s Theorem, there must be some infinite monochromatic clique in this graph, so either there is an infinite descending chain, an infinite antichain, or an infinite subsequence \(i_1 < i_2 < \cdots\) with \(x_{i_1} \leq x_{i_2} \leq \cdots\). \(\square\)

Proposition 15.10. A partial order \((X, \leq)\) is a well partial order iff for all upsets \(U \subseteq X\), every element of \(U\) is \(\geq\) some minimal element of \(U\) and \(U\) has finitely many minimal elements, that is, there exists a finite set of elements \(u_1, \ldots, u_k \in U\) such that \(U = \{x \mid x \geq u_i \text{ for some } i\}\).

Proof. Suppose first that \((X, \leq)\) is a well partial order. If some element \(u \in U\) is not above any minimal element of \(U\), then we can find an infinite descending chain in \(U\). Since any pair of distinct minimal elements of \(U\) are incomparable, the number of minimal elements of \(U\) must be finite. The converse follows from the previous proposition. \(\square\)

So the last ingredient of the argument will be the proof that \(\leq_E\) is a well partial order. While the tools we have available are capable of proving this directly, it is useful to reduce this to the fact that the simpler (and more well-known) embeddability partial ordering \(\leq_e\), due to Higman, is a well partial order - this allows us to transfer other results about Higman’s ordering to the partial order \(\leq_E\).

Definition 15.11. Define the partial order \(\leq_e\) on \(A^+\) by \(a \leq_e b\) if \(b\) can be formed from \(a\) by inserting elements of \(A\).

Proposition 15.12. If \(B\) is the disjoint union of \(A\) with the two-element set of symbols \(\{\#, \prime\}\), then there is an embedding of partial orders

\[ F : (A^+, \leq_E) \hookrightarrow (B^+, \leq_e), \]

i.e. a function \(F\) such that for \(x, y \in A^+\), we have \(x \leq_E y\) iff \(F(x) \leq_e F(y)\).

Proof. We define \(F : A^+ \rightarrow B^+\) to be the function which modifies \(x \in A^+\) by inserting a \(\prime\) after the first occurrence of each symbol within \(x\), inserting a \# at the end of \(x\), and then following that with a \(\prime\) for each symbol which doesn’t occur within \(x\). For instance, if \(A = \{a, b, c, d, e\}\), then

\[ F(adaadca) = a'd'aadc'\#'' , \]

where the two \(\prime\)s at the end keeps track of the fact that \(b, e\) did not occur within the word \(adaadca\).

Note that \(F(x)\) is always formed from \(x\) by inserting exactly 1 copy of \# and exactly \(|A|\) copies of the symbol \(\prime\). Thus, if \(F(x) \leq_e F(y)\), then \(F(y)\) must be obtained by inserting only symbols from \(A\) into the word \(F(x)\). If any symbol \(s \in A\) is inserted before its first occurrence in \(F(x)\), or inserted directly in front of a \(\prime\), then we can see that the resulting word can’t be of the form \(F(y)\), by considering the first location with an invalid insertion. \(\square\)

Theorem 15.13. If \(A\) is a finite set, then the partial order \(\leq_e\) on \(A^+\) is a well partial order.
Proof. We prove this by induction on $|A|$. Since $\leq_e$ clearly has no infinite descending chains (as $a <_e b$ implies $|a| < |b|$), we just need to prove that $\leq_e$ has no infinite antichains. Suppose for contradiction that $\leq_e$ has an infinite antichain, and let $x_1, x_2, \ldots$ be a lexicographically minimal infinite antichain, that is, suppose that $x_1$ is minimal such that there exists an infinite antichain containing $x_1$, that $x_2$ is minimal such that there exists an infinite antichain containing $\{x_1, x_2\}$, etc.

By the infinite pigeonhole principle, we see that there is an infinite subsequence $i_1 < i_2 < \ldots$ such that every element $x_{ij}$ ends in the same element of $A$, say $a$. Let $x'_{ij}$ be the element of $A^+$ we obtain by deleting the $a$ in the last coordinate of $x_{ij}$, then from the definition of $\leq_e$ we see that $x_{ij} \leq_e x_{ik} \iff x'_{ij} \leq_e x'_{ik}$. Let $j$ be minimal such that $x_j >_e x'_{ik}$ for some $k$, and note that $j \leq i_1$ so $j$ is well-defined. Then the sequence

$$x_1, x_2, \ldots, x_{j-1}, x'_i, x'_{ik+1}, \ldots$$

is also an infinite antichain, and is lexicographically smaller than $x_1, x_2, \ldots, x_{j-1}, x_j, \ldots$, a contradiction. \hfill $\Box$

**Corollary 15.14.** If $A$ is a finite set, then the partial order $\leq_E$ on $A^+$ is a well partial order.

**Theorem 15.15** (Few subpowers implies inherently finitely related [1]). If a clone $O$ contains a $k$-edge term, then it is finitely related. In fact, a set $\Gamma \subseteq \text{Inv}(O)$ generates $\text{Inv}(O)$ iff the following two conditions are satisfied:

- every relation of arity strictly less than $k$ in $\text{Inv}(O)$ is contained in $(\Gamma)$, and

- for each minority pair $(c, d) \in A^2$ and each minimal element $a \in \lambda(O, (c, d))$, if we set $n = |a|$, then the relation $\text{Pol}_n(\Gamma)$ on $A^n$ defined by the primitive positive formula

$$\bigwedge_{R \in \Gamma} \bigwedge_{M \in R^n} f(M) \in R$$

has $(c, d) \not\in \text{Forks}(\text{Pol}_n(\Gamma), a)$.

Proof. By the fact that $\leq_E$ is a well partial order, we see that there is a finite set $\Gamma \subseteq \text{Inv}(O)$ which satisfies the conditions given: for instance, we may take $\Gamma$ to consist of the collection of all relations in $\text{Inv}(O)$ of arity less than $k$, together with the relations $O_n \preceq A^n$ for every $n$ such that some minimal element $a \in \lambda(O, (c, d))$ has $|a| = n$ for some $(c, d) \in A^2$.

Now suppose that $\Gamma$ satisfies the given conditions. Then for any $(c, d) \in A^2$ and any $b \in \lambda(O, (c, d))$, there exists some minimal $a \in \lambda(O, (c, d))$ with $a \leq_E b$. Thus if $|a| = n, |b| = m$, then $\text{Forks}(\text{Pol}_m(\Gamma), b) \subseteq \text{Forks}(\text{Pol}_n(\Gamma), a)$, and by the second condition on $\Gamma$ we have $(c, d) \not\in \text{Forks}(\text{Pol}_n(\Gamma), a)$. Thus for any $b \in A^+$ with $|b| = m$, we have

$$(c, d) \not\in \text{Forks}(\text{Pol}_m, b) \implies (c, d) \not\in \text{Forks}(\text{Pol}_n(\Gamma), b),$$

so $\text{Forks}(\text{Pol}_m(\Gamma), b) \subseteq \text{Forks}(O_m, b)$. Since $\text{Pol}_m(\Gamma)$ contains $O_m$ (by $\Gamma \subseteq \text{Inv}(O)$), and every projection of $\text{Pol}_m(\Gamma)$ onto fewer than $k$ coordinates of $A^{\lambda_m}$ is contained in the corresponding projection of $O_m$ (by the first condition on $\Gamma$), we can apply Theorem [12.1] to see that $\text{Pol}_m(\Gamma) = O_m$. \hfill $\Box$

**Corollary 15.16.** The number of clones on a finite set which contain an edge term is countable.
Remark 15.1. There is a converse to Theorem 15.15 if $O$ is a clone on a finite set such that every clone $O'$ with $O' \supseteq O$ is finitely related, then $O$ has an edge term. The proof of this relies on the theory of cube term blockers, which roughly states that a clone $O$ fails to contain a cube term iff there is an infinite sequence of invariant relations which look like the relations $\{0, 1\}^n \setminus \{(0, \ldots, 0)\}$ - recall that the clone corresponding to this sequence of relations on $\{0, 1\}$ was our basic example of a clone which was not finitely related (Example 2.3).

Example 15.1. Consider the algebra $A = \{\{a, b, c\}, g\}$ from Example 12.1 which has $\{a, b\}$ a majority subalgebra and $\{a, c\}$ an absorbing minority subalgebra. Recall that the minority pairs of $A$ were $(a, c), (c, a), (b, c).$ Since $a \in Sg_A\{b, c\},$ for any $s \in A^+$ we have

$$(b, c) \in \text{Forks}(\langle g \rangle, s) \implies (a, c) \in \text{Forks}(\langle g \rangle, s).$$

Take the standard alphabetical ordering $<\{a, b, c\}$. It’s easy to check that $\lambda(\langle g \rangle, (a, c))$ contains $a, b, c, ab, ba, bc, ca, cb, abc, acb, acc, bac, bca, cab, cba$ and that $\lambda(\langle g \rangle, (b, c)) = A^+$: for the strings of length 2, the free algebra $F_A(x, y)$ only has six elements so we may compute the forks directly, for permutations of $abc$ we note that a corresponding permutation of $aac$ comes before it and $g$ preserves the congruence corresponding to the partition $\{a, b\}, \{c\},$ and for $acc$ we note that $aac$ and $aaa$ come before it and that $g$ preserves the affine ternary relation $\{(a, a, a), (a, a, c), (a, c, a), (c, a, c), (c, c, a)\}.$

To complete the description of $\lambda(\langle g \rangle, (a, c)),$ we just need to check that for all $2 \leq i \leq n,$ the word $s_{in} = a \cdots aca \cdots a \in A^+$ of length $n$ with a $c$ in the $i$th position and as elsewhere has $(a, c) \in \text{Forks}(\langle g \rangle, s_{in}).$ For this, we take the terms $x_1$ and $g(x_1, x_1, x_1)$ in the free algebra, and check that they make a fork at $s_{in}.$ For $s' <_{\text{lex}} s_{in},$ we have $s'_i = a$ and $s'_i < c,$ so

$$g(s'_i, s'_i) = g(a, a, a) \text{ or } g(a, a, b) = a = s_1,$$

so $x_1$ and $g(x_1, x_1, x_1)$ agree on tuples which come lexicographically before $s_{in}.$ At $s_{in},$ we get the fork $(a, g(a, a, c)) = (a, c).

Example 15.2. Consider the gmn algebra $A_2 = \{(a, b, c), \varphi_2\}$ from Example 11.2 which had majority subalgebras $\{a, b\}, \{a, c\}$ and minority subalgebra $\{b, c\}.$ The only minority pair to worry about is $(b, c),$ and under the standard alphabetical ordering $<\{a, b, c\},$ we find that $\lambda(\langle \varphi_2 \rangle, (b, c))$ contains the following 16 elements of $A^+$:

$$a, b, c, ab, ac, ba, ca, cb, acb, bcc, cab, cba, abcc, bacc, bcac, beca.$$

Again, it is easy to check the strings of length 2 as $F_{A_2}(x, y)$ only has 4 elements, strings which have a $c$ preceding the first $b$ such as $acb$ don’t work because the corresponding word with $bs$ and $cs$ swapped (i.e. $abc$ in this case) comes before it and $\varphi_2$ preserves order two the automorphism swapping $b$ and $c,$ and strings containing $bca$ such as $abcc$ don’t work because the two strings where one of the $cs$ is replaced by a $b$ (i.e. $abcc$ and $abc$ in this case) come before it and $\varphi_2$ preserves the ternary relation corresponding to the columns of the matrix

$$\begin{bmatrix}
a & b & b & c & c 
a & b & c & b & c 
a & b & c & c & b
\end{bmatrix}.$$ 

It’s much harder to show that the remaining elements $s$ which are not $\geq_E$ to one of the 16 strings displayed above all have $(b, c) \in \text{Forks}(\langle \varphi_2 \rangle, s).$ Each such $s$ has at least one $b,$ exactly one
c, and has its first b before its c. We may assume without loss of generality that s begins with a b, and suppose s has its only c at the rth position for some r ≥ 2. We need to show that there is some term t ∈ F(A_2(x_1, ..., x_n)) such that the pair (x_1, t) gives us a fork at s. In other words, we need to show that we can find a term t such that for each s' <_Lex s we have t(s') = s'_1 and t(s) = c.

The only way I know to show the existence of such a term t is to use the analysis of critical relations in Inv_k(A_2) carried out in Example 13.1. By that analysis, we see that every relation R ≤ A_k is the intersection of some family of binary relations and some family of relations R_I ≤ A_I such that for each I and each i, j ∈ I, we have π_{i,j}(R_I) ⊆ 0^*_A, where 0^*_A is the congruence corresponding to the partition \{a\}, \{b, c\}. Thus, if the term t we are looking for does not exist, then either there is some s' <_Lex s such that 
\[
\begin{bmatrix} s'_1 \\ c \end{bmatrix} \notin Sg_{A_2} \begin{bmatrix} s' \\ s \end{bmatrix},
\]
or there is some family s^1, ..., s^k <_Lex s such that for each j, l, we have (s'_j, s_j) ∈ 0^*_A but
\[
\begin{bmatrix} s^1_1 \\ s^1 \\ s^k_1 \\ c \end{bmatrix} \notin Sg_{A_2} \begin{bmatrix} s^1 \\ s^1 \\ s^k \\ s \end{bmatrix}.
\]

To rule out the first possibility, we note that if s' <_Lex s then s'_1 ∈ \{a, b\}, and if (s'_1, s'_1) ≠ (b, c), then (s'_1, s'_1) is a majority pair and taking \varphi_2(x_1, x_1, x_i) does the trick, while if (s'_1, s'_1) = (b, c), then at the first location j where s' and s differ we must have s'_j = a, s_j = b, so taking \varphi_2(x_j, x_1, x_i) does the trick:
\[
\varphi_2 \begin{bmatrix} a & b & c \\ b & b & c \end{bmatrix} = \begin{bmatrix} b \\ c \end{bmatrix}.
\]

To rule out the second possibility, note that if s' <_Lex s and (s'_j, s_j) ∈ 0^*_A for all j, then the first coordinate where s' and s can differ is at the coordinate i with s_i = c, so we must have s'_1 = b, s_i = c and s'_1 = s_1 = b. Thus the term x_i rules out the second possibility.

16 Fourth basic example: the Rock-Paper-Scissors algebra

We’re going to start building intuition for the bounded width case with a detailed investigation of a fourth basic algebra on three elements, which is sometimes called the “rock-paper-scissors” algebra. This algebra is A = (\{a, b, c\}, ·), where · is the binary, commutative, idempotent operation described by the following table.

<table>
<thead>
<tr>
<th>·</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
<td>c</td>
<td>c</td>
</tr>
</tbody>
</table>

The algebra A is not a semilattice, but every two-element subset of A is a semilattice. Thus, the binary operation · satisfies the following identities:

\[ xx \approx x, \quad xy \approx yx, \quad x(xy) \approx xy. \]
Any binary operation satisfying the above identities is known as a 2-semilattice operation, and the algebra $\mathbb{A}$ is the smallest 2-semilattice which is not a semilattice.

As we will see, the corresponding relational clone is generated by the binary relation $\{(a, b), (b, c), (c, a)\}$ (which corresponds to the fact that the algebra has a cyclic automorphism) and the ternary relation $R_{a,b}$ given by the formula

$$R_{a,b}(x, y, z) := (x \in \{a, b\}) \land (x = a \implies y = z).$$

The ternary relation $R_{a,b}$ has a special role, which is closely connected to the fact that $\{a, b\}$ is a semilattice subalgebra of $\mathbb{A}$.

**Proposition 16.1.** If $\mathbb{R}, \mathbb{S} \subseteq \mathbb{A}^n$ are any $n$-ary relations with $\mathbb{S} \subseteq \mathbb{R}$, then the $n + 1$-ary relation

$$((x, y) \in \mathbb{R} \times \{a, b\}) \land (y = a \implies x \in \mathbb{S})$$

can be defined by a primitive positive formula in $\mathbb{R}, \mathbb{S}$, and $R_{a,b}$.

**Proof.** Just use the following primitive positive formula:

$$\exists z \in \mathbb{A}^n \ x \in \mathbb{R} \land z \in \mathbb{S} \land \bigwedge_{i \leq n} R_{a,b}(y, x_i, z_i).$$

**Proposition 16.2.** If $\mathbb{R} \leq_{sd} \mathbb{A}^n$ is a subdirect $n$-ary relation, then $\mathbb{R}$ is the intersection of its two-variable projections, each of which is either a full relation or the graph of an automorphism of $\mathbb{A}$ which is either the identity or is cyclic. In particular, there is some subset of the coordinates $I \subseteq \{1, \ldots, n\}$ such that the projection $\pi_I$ is an isomorphism from $\mathbb{R}$ to $\mathbb{A}^I$.

**Proof.** We prove this by induction on $n$. The base case, $n = 2$, is easily verified: since $\mathbb{A}$ is simple, every subdirect binary relation on $\mathbb{A}$ is either the graph of an automorphism or is linked, and we can check that every connected subgraph of the complete bipartite graph $K_{3,3}$ either contains a bipartite matching or is a tree with two leaves on both parts (e.g. using Hall’s Marriage Lemma). Therefore up to automorphisms of $\mathbb{A}$ we just need to consider relations which contain $\{(a, a), (a, b), (b, b), (c, c)\}$, $\{(a, a), (b, c), (c, b)\}$, or $\{(a, b), (a, c), (b, a), (c, a)\}$, and all three of these generate $\mathbb{A}^2$.

Now consider the case $n > 2$. By the induction hypothesis, we may assume without loss of generality that $\pi_{[n] \setminus \{i\}}(\mathbb{R}) = \mathbb{A}^{n-1}$ for every $i \leq n$. Suppose for contradiction that $\mathbb{R} \neq \mathbb{A}^n$.

Since the automorphism group of $\mathbb{A}$ is transitive, we may assume without loss of generality that $(a, \ldots, a) \not\in \mathbb{R}$. Since $\mathbb{A}$ is idempotent, the set $\mathbb{R}'$ of triples $(x, y, z)$ such that $(x, y, z, a, \ldots, a) \in \mathbb{R}$ is a subalgebra of $\mathbb{A}^3$, and by the inductive hypothesis every projection of $\mathbb{R}'$ onto any pair of coordinates is full. So we can reduce to the case $n = 3$.

If any two of $(a, a, c), (a, c, a), (c, a, a)$ are in $\mathbb{R}$, then we can combine them to obtain $(a, a, a)$. So we may suppose that $(a, a, b) \in \mathbb{R}$. If we consider the binary relation consisting of pairs $(y, z)$ with $(a, y, z) \in \mathbb{R}$, then by the $n = 2$ case, we must have $(a, c, a) \in \mathbb{R}$. Similar reasoning with the roles of the first and second coordinates reversed then shows that we must have $(c, a, a) \in \mathbb{R}$, a contradiction.

**Proposition 16.3.** If $\mathbb{R} \leq_{sd} \mathbb{A}^n \times \{a, b\}^k$ has full projection onto $\mathbb{A}^n$, then we have $\mathbb{A}^n \times \{b, \ldots, b\} \subseteq \mathbb{R}$.  

123
Proof. For any \( x \in \mathbb{A}^n \), let \( x^- \) be the tuple obtained from \( x \) by applying the cyclic permutation (\( a \; c \; b \)) componentwise. Then it’s easy to check that for any \( x, y \in \mathbb{A}^n \), we have

\[
(xy^-)y = y.
\]

By multiplying all of the elements of \( \mathbb{R} \) together in any order (with parentheses placed arbitrarily), we see that there is some \( x \in \mathbb{A}^n \) such that \((x_1,\ldots,x_n, b,\ldots,b) \in \mathbb{R} \). For any \( y \in \mathbb{A}^n \), there are tuples \( c, d \in \{a, b\}^n \) such that \((y, c), (y^-, d) \in \mathbb{R} \) by the assumption \( \pi_{[n]}(\mathbb{R}) = \mathbb{A}^n \). Thus

\[
(y, b) = ((x, b) \cdot (y^-, d)) \cdot (y, c) \in \mathbb{R}.
\]

The previous two propositions are enough to describe an algorithm which solves CSP(\( \mathbb{A} \)). The algorithm first establishes arc-consistency, reducing some of the domains of the variables until every constraint relation becomes subdirect. Then for each variable with a two element domain, the last proposition shows that we may as well take that variable equal to the top/absorbing element of that domain. After this restriction, if we consider the remaining variables, each relation decomposes into binary relations, each of which is either an equality relation or the graph of a cyclic automorphism. This final problem can be solved by checking that no cycle of binary relations implies that any variable is related to itself by a nontrivial cyclic automorphism.

Definition 16.4. An instance of a CSP is cycle-consistent if for every sequence of variables \( v_1,\ldots, v_n \) and relations \( R_1,\ldots, R_n \) and pairs of coordinates \((i_k, j_k)\) such that \( v_k, v_{k+1} \) are related by \( \pi_{(i_k,j_k)}(R_k) \) for each \( k \) (indices taken modulo \( n \)), the composition

\[
\pi_{(i_1,j_1)}(R_1) \circ \cdots \circ \pi_{(i_n,j_n)}(R_n)
\]

contains the equality relation on the domain of the variable \( v_1 \).

Corollary 16.5. Any cycle-consistent instance of CSP(\( \mathbb{A} \)) has a solution.

If we want to understand the complete structure of a general relation \( \mathbb{R} \subseteq \mathbb{A}^n \), things become more complicated. A typical relation we need to consider has the form

\[
x_1 \in \{a, b\} \land (x_1 = a \implies x_2 \in \{a, b\}) \land (x_1 = x_2 = a \implies x_3 \in \{a, b\}) \\
\land \cdots \land (x_1 = \cdots = x_k = a \implies y = z).
\]

The final \( y = z \) in the last implication can also be replaced with any unary relation on \( y \), and for any subset of the variables we can apply a cyclic automorphism of \( \mathbb{A} \). We call any such relation a basic relation on \( \mathbb{A} \).

Theorem 16.6. Suppose \( \mathbb{R} \subseteq \mathbb{A}^n \). Then \( x \in \mathbb{R} \) iff \( x \) satisfies every basic relation on \( \mathbb{A} \) which contains \( \mathbb{R} \). In particular, \( \mathbb{R} \) is contained in the relational clone generated by \( \{(a, b), (b, c), (c, a)\} \) and \( R_{a,b} \).

Proof. Suppose \( x \in \mathbb{R} \). Let \( I = \{i_1,\ldots,i_k\} \subseteq [n] \) be maximal such that, after applying cyclic automorphisms to coordinates in \( I \), we have \( x_{ij} = a \) for all \( j \leq k \), and such that the basic relation

\[
y_{i_1} \in \{a, b\} \land (y_{i_1} = a \implies y_{i_2} \in \{a, b\}) \land (y_{i_1} = y_{i_2} = a \implies y_{i_3} \in \{a, b\}) \\
\land \cdots \land (y_{i_1} = \cdots = y_{i_{k-1}} = a \implies y_{i_k} \in \{a, b\})
\]

124
contains $\mathbb{R}$. Assume without loss of generality that the coordinates are ordered such that $I = \{n - k + 1, \ldots, n\}$ and such that the $n - k$-ary relation $\mathbb{R}'$ defined by

$$(y_1, \ldots, y_{n-k}) \in \mathbb{R}' \iff (y_1, \ldots, y_{n-k}, a, \ldots, a) \in \mathbb{R}$$

has $\mathbb{R}' \leq_{sd} \mathbb{A}^m \times \{a, b\}^{n-m-k}$ for some $m$ (possibly after further applications of cyclic automorphisms). Then by the maximality of $I$, we have $x = (x_1, \ldots, x_m, b, \ldots, b, a, \ldots, a)$. By Propositions 16.2, 16.3 and our assumption that $x$ satisfies all basic relations containing $\mathbb{R}$, we have $(x_1, \ldots, x_m) \in \pi_{[m]}(\mathbb{R}')$ and $\pi_{[m]}(\mathbb{R}') \times \{(b, \ldots, b)\} \subseteq \mathbb{R}'$, so $(x_1, \ldots, x_m, b, \ldots, b) \in \mathbb{R}'$, so $x \in \mathbb{R}$. \hfill $\square$

**Proposition 16.7.** Suppose that $f : \mathbb{A}^n \to \mathbb{A}$ is any idempotent operation which depends on all of its inputs and preserves the relation $R_{a,b}$. Then the restriction of $f$ to $\{a, b\}$ must be the $n$-ary semilattice operation on $\{a, b\}$, that is, for any $(x_1, \ldots, x_n) \in \{a, b\}^n \setminus \{(a, \ldots, a)\}$, we have $f(x_1, \ldots, x_n) = b$.

**Proof.** Suppose for contradiction that there is some $(x_1, \ldots, x_n) \in \{a, b\}^n \setminus \{(a, \ldots, a)\}$ with $f(x_1, \ldots, x_n) \neq b$. Since $\{a, b\} = \pi_1(R_{a,b})$ is preserved by $f$, we must then have $f(x_1, \ldots, x_n) = a$. We will show that for all $i$ with $x_i = b$, $f$ does not depend on its $i$th input.

Let $y, z \in \mathbb{A}^n$ be any pair of tuples with $y_i = z_i$ whenever $x_i = a$. Then each $(x_i, y_i, z_i) \in R_{a,b}$, so

$$\begin{bmatrix} a \\ f(y) \\ f(z) \end{bmatrix} = f \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ z_1 & z_2 & \cdots & z_n \end{bmatrix} \in R_{a,b},$$

so $f(y) = f(z)$. \hfill $\square$

**Theorem 16.8.** An $n$-ary operation $f$ is contained in $\text{Clo}_n(\mathbb{A})$ iff it preserves the relations $\{(a, b), (b, c), (c, a)\}$ and $R_{a,b}$. If $f$ depends on all its inputs, this occurs iff $f$ preserves the cyclic automorphism of $\mathbb{A}$ and $f|_{\{a,b\}}$ is the $n$-ary semilattice operation on $\{a, b\}$.

**Proof.** We just need to check this in the case where $f$ depends on all of its inputs. Let $\mathcal{F} = \mathcal{F}_{\pi_t(\mathbb{A})}(x_1, \ldots, x_n) \leq \mathbb{A}^n$ be the subalgebra generated by $\pi_1, \ldots, \pi_n : \mathbb{A}^n \to \mathbb{A}$. The projection $\pi_x(\mathcal{F})$ of $\mathcal{F}$ onto the coordinate of $\mathbb{A}^n$ corresponding to $x \in \mathbb{A}^n$ is the subalgebra of $\mathbb{A}$ generated by $\{\pi_1(x), \ldots, \pi_n(x)\} = \{x_1, \ldots, x_n\}$.

If $x$ is a diagonal tuple, say $x = (a, \ldots, a)$, then $\pi_x(\mathcal{F}) = \{a\}$, corresponding to the fact that any $f \in \mathcal{F}$ must be idempotent, with $f(a, \ldots, a) = a$. If exactly two elements of $\mathbb{A}$ occur in $x$, say $x \in \{a, b\}^n$, then $\pi_x(\mathcal{F}) = \{a, b\}$, and if $f$ depends on all its inputs and preserves $R_{a,b}$, this implies that we must have $f(x) = b$, i.e. $\pi_x(f) = b$. Thus, if $I \subseteq \mathbb{A}^n$ is the set of $x$ such that all three of $a, b, c$ show up in the coordinates of $x$, we see that $\pi_I(\mathcal{F}) \leq_{sd} \mathbb{A}^I$, and by Proposition 16.3 we have $f \in \mathcal{F} \iff \pi_I(f) \in \pi_I(\mathcal{F})$.

By Proposition 16.2, $\pi_I(\mathcal{F})$ is the intersection of its two-variable projections, each of which is either full or the graph of a cyclic automorphism of $\mathbb{A}$. A two variable projection $\pi_{x,y}(\mathcal{F})$ will only be the graph of a cyclic automorphism $\sigma \in \text{Aut}(\mathbb{A})$ if $\pi_i(x), \pi_i(y)$) is in the graph of $\sigma$ for all $i$, that is, if $y_i = \sigma(x_i)$ for all $i$. Thus, $\pi_I(f) \in \pi_I(\mathcal{F})$ iff whenever $y = \sigma(x)$, we have $f(y) = \sigma(f(x))$. \hfill $\square$

Note that one of the key steps behind the analysis of the rock-paper-scissors algebra was Proposition 16.2 which classified the subdirect powers of the algebra, and that the method of proof depended only on checking properties of subdirect binary and ternary relations on $\mathbb{A}$. The general pattern behind this is best understood in terms of a property of the polynomial clone known as polynomial completeness.
Definition 16.9. An algebra is \emph{polynomially complete} if its polynomial clone is the clone of all operations, that is, if every operation on the underlying set can be expressed using the basic operations of the algebra together with the constant operations.

Theorem 16.10. A finite idempotent algebra $A$ is polynomially complete if every binary relation on $A$ which contains the diagonal is either the equality relation or the full relation, and every ternary relation $R \leq_{sd} A^3$ such that every two variable projection of $R$ is full is equal to the full relation $A^3$.

Proof. We will show by induction on $n$ that every $n$-ary relation $R \leq A^n$ which contains the subalgebra of diagonal tuples $(x, \ldots, x)$, $x \in A$ is given by a conjunction of equalities between pairs of coordinates. The base case $n = 2$ follows from our assumption on $A$. By the inductive hypothesis, we may assume without loss of generality that $\pi_i|\{i\} R = A^{n-1}$ for each $i$.

If $n = 3$, then our assumption on $A$ implies that $R = A^3$. Otherwise, suppose for contradiction that $(x_1, \ldots, x_n) \notin A^n$, and consider the ternary relation $R'$ consisting of triples $(u, v, w)$ such that $(u, v, w, x_1, \ldots, x_n) \in R$. Since $A$ is idempotent, $R'$ is a subalgebra of $A^3$, and every two-variable projection of $R'$ is full, so by the $n = 3$ case we must have $(x_1, x_2, x_3) \in R'$, a contradiction.

Note that we have shown that the relational clone corresponding to the polynomial clone of $A$ is generated by the equality relation. The general Inv – Pol Galois duality now shows that $A$ is polynomially complete. To see this concretely, consider the subalgebra of $A^{A^n}$ generated by the functions $\pi_i$ and the constant (diagonal) tuples. Then this subalgebra is described by a conjunction of equalities between pairs of coordinates. But no two-variable projection of this subalgebra can be an equality relation: if $x \neq y \in A^n$, then there is always some $i$ such that $\pi_i(x) \neq \pi_i(y)$. Thus this subalgebra of $A^{A^n}$ must be the full set of operations $A^n \to A$.

Corollary 16.11. The rock-paper-scissors algebra is polynomially complete.

As far as relations go, the main impact of polynomial completeness is that it strongly constrains subdirect relations where each factor is polynomially complete. As we have seen, if some factors are not polynomially complete, then the structure of an arbitrary relation can be quite intricate. In the case of the rock-paper-scissors algebra, we are able to side-step this intricacy by restricting each factor which is a proper subalgebra of $A$ to its top/absorbing element. This is a general strategy that can be used in the study of bounded width algebras, as well as finite Taylor algebras.

We conclude this section with a few classical results about polynomial completeness.

Definition 16.12. The ternary \emph{discriminator} function is the function $t$ defined by

$$t(x, y, z) = \begin{cases} z & x = y, \\ x & x \neq y. \end{cases}$$

Proposition 16.13. A finite algebra is polynomially complete iff it has the ternary discriminator as a polynomial operation.

Proof. One direction is obvious. For the other direction, it’s enough to show that the idempotent algebra $A = (A, t)$ whose only basic operation is the ternary discriminator $t$ is polynomially complete. We may assume that the underlying set $A$ contains at least two distinct elements $a, b$. Suppose
first that \( R \leq_{sd} A^2 \) is a relation properly containing the diagonal of \( A^2 \), and assume without loss of generality that \((a, b) \in R\) with \( a \neq b \). Then for any \( c \in A \), we have
\[
\begin{bmatrix}
a \\
c
\end{bmatrix} = t \begin{bmatrix}
a & b & c \\
 b & b & c
\end{bmatrix} \in R,
\]
and similarly \((d, b) \in R\) for any \( d \in A \). Then for any \( c, d \in A \) we have
\[
\begin{bmatrix}
d \\
c
\end{bmatrix} = t \begin{bmatrix}
a & a & d \\
 c & b & b
\end{bmatrix} \in R,
\]
so \( R = A^2 \).

To finish, we just need to show that any ternary relation \( R \leq_{sd} A^3 \) such that every two variable projection is full must be the full relation \( A^3 \). Since \( A \) has full automorphism group, if \( R \neq A^3 \) then we may assume without loss of generality that \((a, a, a) \notin R\), while all three of \((a, a, b), (a, b, a), (b, a, a)\) are in \( R \). Then we have
\[
\begin{bmatrix}
b \\
 a \\
b
\end{bmatrix} = t \begin{bmatrix}
a & a & b \\
a & b & a \\
b & a & a
\end{bmatrix} \in R,
\]
so
\[
\begin{bmatrix}
a \\
a \\
a
\end{bmatrix} = t \begin{bmatrix}
a & b & b \\
a & a & a \\
b & b & a
\end{bmatrix} \in R,
\]
contradicting the assumption \((a, a, a) \notin R\). \(\square\)

**Example 16.1.** We can give an alternative proof of the fact that the rock-paper-scissors algebra is polynomially complete by expressing the ternary discriminator as a polynomial. First, we can define the unary polynomial \( x^+ \) corresponding to the cyclic permutation \((a b c)\) by
\[
x^+ = ((xa)c)(xb),
\]
and we can define the inverse of this by \( x^- = (x^+)^+ \). Note that we now have
\[
xy^+ = \begin{cases} x^+ & x = y, \\
x & x \neq y. \end{cases}
\]
Thus if we set \( u(x, y, z) = (z(xy^+)^-)x \), then we have
\[
u(x, y, z) = (z(xy^+)^-)x = \begin{cases} xz & x = y, \\
x & x \neq y, \end{cases}
\]
so we may take
\[
t(x, y, z) = ((u(x, y, z)u(x, y, z^+)^-)u(x, y, z^-)^+)^-.
\]
To see that this works, note that if \( x = y \), then two of \( xz, (xz^+)^-, (xz^-)^+ \) are equal to \( z \) while the third is equal to \( z^+ \), so since \( \{z, z^+\} \) is a semilattice we see that in this case \( t(x, y, z) \) is given by
\[
(((xz)(xz^+)^-)(xz^-)^+)^- = (zzz)^- = (z^+)^- = z,
\]
while if \( x \neq y \) then \( u(x, y, z) = x \), so \( t(x, y, z) \) is given by
\[
((xx^-)^+)^- = (xx^+)^- = (x^+)^- = x.
\]
The ternary discriminator \( t \) satisfies the system of identities
\[
\begin{align*}
t(x, y, y) &\approx x, \\
t(x, y, x) &\approx x, \\
t(y, y, x) &\approx x.
\end{align*}
\]
Any ternary term satisfying this system of identities is known as a Pixley term. Note that any Pixley term is automatically a Mal’cev term, and that the term \( d \) defined from \( t \) by
\[
d(x, y, z) = t(x, t(x, y, z), z)
\]
is automatically a majority term. In the case where \( t \) is the ternary discriminator, \( d \) becomes the dual discriminator of Example 7.5.

**Theorem 16.14** (Pixley [96]). An algebra \( \mathbb{A} \) generates a variety which is both congruence permutable and congruence distributive iff it has a Pixley term. If \( \mathbb{A} \) is also simple, then it is polynomially complete.

**Proof.** If \( \mathbb{A} \) has a Pixley term, then it has both a Mal’cev term and a majority term, so it generates a congruence permutable and congruence distributive variety. Conversely, suppose that \( \mathbb{A} \) generates a congruence permutable and congruence distributive variety. Let \( \mathcal{F} = \mathcal{F}_V(\mathbb{A})(x, y, z) \) be the free algebra on three generators in this variety, and for \( a, b \in \{x, y, z\} \) let \( \theta_{ab} \) be the smallest congruence with \( a \equiv_{\theta_{ab}} b \). Then \((x, z) \in \theta_{xz} \land (\theta_{xy} \circ \theta_{yz})\), so by congruence distributivity and permutability, we have
\[
(x, z) \in \theta_{xz} \land (\theta_{xy} \lor \theta_{yz}) = (\theta_{xz} \land \theta_{xy}) \lor (\theta_{xz} \land \theta_{yz}) = (\theta_{xz} \land \theta_{yz}) \circ (\theta_{xz} \land \theta_{xy}).
\]
Thus there is some \( t \in \mathcal{F} \) such that
\[
x \ (\theta_{xz} \land \theta_{yz}) \ t \ (\theta_{xz} \land \theta_{xy}) \ z.
\]
Thus \( t \) is a ternary term which satisfies the Pixley identities.

Now suppose that \( \mathbb{A} \) is simple. Since \( \mathbb{A} \) is Mal’cev, every binary relation on \( \mathbb{A} \) is the graph of an isomorphism modulo the linking congruence, and the linking congruence is necessarily either \( 0_\mathbb{A} \) or \( 1_\mathbb{A} \). Thus every binary relation on \( \mathbb{A} \) which contains the diagonal is either full or equal to the diagonal. Since \( \mathbb{A} \) has a majority term, every ternary relation on \( \mathbb{A} \) whose two variable projections are all full must itself be a full relation. Thus \( \mathbb{A} \) is polynomially complete. \( \Box \)

Varieties which are both congruence distributive and congruence permutable are known as arithmetical varieties. The name arithmetical comes from the theory of arithmetical rings, which are rings where the “Chinese remainder condition” holds: for any ideals \( I_1, ..., I_n \) and elements \( a_1, ..., a_n \) with \( a_i \equiv a_j \pmod{I_i + I_j} \) for all \( i, j \), there exists some \( x \) with \( x \equiv a_i \pmod{I_i} \) for all \( i \).

**17 Partial semilattice operations and the digraph of semilattice subalgebras**

In this section we will go over a binary analogue of a standard result about iterating unary functions to make (compositionally) idempotent functions, that is, functions satisfying \( e \circ e = e \). First we review the case of unary iteration.
Definition 17.1. If $f : A \rightarrow A$ is a unary function, we define $f^m$ to be $f \circ \cdots \circ f$, with $n$ copies of $f$. If $(A, f)$ is either finite or profinite, we define $f^\infty$ by

$$f^\infty(x) := \lim_{n \to \infty} f^{\circ n}(x).$$

Alternatively, we can define $f^\infty$ as the limit of $f^m$ over the net of positive integers $n$, ordered by divisibility. Similarly, we define $f^{\infty-1}$ by

$$f^{\infty-1}(x) := \lim_{n \to \infty} f^{\circ(n-1)}(x).$$

Proposition 17.2. If $(A, f)$ is profinite, then the limit defining $f^\infty$ exists, and $f^\infty$ satisfies the identity

$$f^\infty(f^\infty(x)) \approx f^\infty(x).$$

Furthermore, if $A$ is finite, then

$$f^\infty = f^{\circ \text{lcm}(1, \ldots, |A|)},$$

and the graph of $f^\infty$ can be computed from the graph of $f$ in time linear in $|A|$.

Proof. It’s enough to prove this in the case where $A$ is finite. Let $m, m'$ be any positive multiples of $\text{lcm}\{1, \ldots, |A|\}$, we will show that $f^{\circ m} = f^{\circ m'}$: this will show that the limit is equal to $f^m$, and taking $m' = 2m$ will show that $f^\infty \circ f^\infty = f^\infty$. To see that $f^{\circ m} = f^{\circ m'}$, note that for any $x$, the sequence $x, f(x), f(f(x)), \ldots, f^{\circ k}(x), \ldots$ must be eventually periodic with period $p$ at most $|A|$, and the periodic behavior must begin within the first $|A|$ steps, so for any $k \geq |A|$ we have $f^{\circ k}(x) = f^{\circ(k+p)}(x)$. Since $m - m'$ is a multiple of $p$ and $m, m' \geq |A|$, this implies that $f^{\circ m} = f^{\circ m'}$.

In order to compute the graph of $f^\infty$ efficiently, we will also compute the function $f^{\infty-1}$ simultaneously. First, make a list of elements of $A$, and mark all of them as “unprocessed”. In each round, we pick the next unprocessed element $x$ from the list, and compute the sequence of iterates $x, f(x), f(f(x)), \ldots$, marking each one as “processed” as we compute it, until the first time we compute $f^{\circ k}(x)$ and find that it has already been marked as “processed”. There are two cases: either $f^{\circ k}(x)$ is equal to $f^{\circ i}(x)$ for some $i < k$, or $f^{\circ k}(x)$ was processed in some previous round.

We can distinguish between the two cases by checking whether the value of $f^\infty(f^{\circ k}(x))$ has already been computed.

In the case where $f^{\circ k}(x) = f^{\circ i}(x)$ for some $i < k$, we first set $f^\infty(f^{\circ j}(x)) := f^{\circ j}(x)$ and $f^{\infty-1}(f^{\circ j}(x)) := f^{\circ(j-1)}(x)$ for $i < j < k$. For $j < i$, we iterate downwards, setting

$$f^\infty(f^{\circ j}(x)) := f^{\infty-1}(f^{\circ(j+1)}(x))$$

and

$$f^{\infty-1}(f^{\circ j}(x)) := f^{\infty-1}(f^\infty(f^{\circ j}(x))).$$

In the case where $f^{\circ k}(x)$ was processed in a previous round, we iterate downwards using the above rules to handle all $j < k$.

Since the number of steps needed for each round is linear in the number of elements which are marked as processed in that round, and since each element of $A$ is marked as processed at most once, the entire procedure for computing $f^\infty$ and $f^{\infty-1}$ runs in time linear in $|A|$.
In the context of CSPs, the reduction to the case of core structures was based on the observation that any non-surjective unary polymorphism \( f : A \to A \) allows us to replace the underlying set \( A \) by the smaller set \( f(A) \) to obtain a homomorphically equivalent CSP on a smaller domain. In this case, the map \( f^\infty : A \to A \) will also be non-surjective, and in fact we have the guarantee that

\[
f^\infty(A) \subseteq f^\con(A)
\]

for all \( n \geq 0 \). So whenever we shrink the domain of a non-core CSP using a unary polymorphism, we may as well assume that the unary polymorphism in question is (compositionally) idempotent.

On the algebraic side, if \( e \circ e = e \) and \( e \in \text{Clo}(A) \), we can define a reduct \( A_e \) of \( A \) as follows. For every \( n \)-ary operation \( f \in \text{Clo}_n(A) \), we define the corresponding operation \( f_e : A^n \to A \) by

\[
f_e(x_1, \ldots, x_n) = e(f(e(x_1)), \ldots, e(x_n)).
\]

Then we define \( A_e \) to be the algebraic structure \((A, \{f_e \mid f \in \text{Clo}(A)\})\) having a basic operation \( f_e \) for each term \( f \) of \( A \).

Each operation \( f_e \) only depends on the restriction of \( f \) to \( e(A) \), and takes values in \( e(A) \). Also, if \( f \) preserves \( e(A) \), then \( f \) and \( f_e \) agree when they are restricted to \( e(A) \). The reduct \( A_e \) has \( e(A) \) as a subalgebra, and is completely determined by its restriction to the subalgebra \( e(A) \) together with the description of the map \( e : A \to e(A) \). So the reduct \( A_e \) and its subalgebra \( e(A) \) are essentially interchangeable, and the subalgebra \( e(A) \) of \( A_e \) has as its basic operations the terms of \( A \) which preserve \( e(A) \).

As a special case of the general result relating reflections to height 1 identities, we have the following basic result.

**Proposition 17.3.** If a system of height 1 identities is satisfied by terms \( f_1, \ldots, f_k \) of \( A \), then the same system of height 1 identities is satisfied by the corresponding operations \( f_1, f_1, \ldots, f_k, f_k \) of \( A_e \) (defined as above).

Note that identities which involve nesting functions may not survive the process of passing from \( A \) to the reduct \( A_e \).

Now we return to the world of idempotent operations, and describe a surprisingly powerful binary analogue of unary iteration. Rather than (compositionally) idempotent operations, we will produce a type of binary operation which I call a **partial semilattice** operation.

**Definition 17.4.** We say that an idempotent binary operation \( s \) is a partial semilattice if it satisfies the identity

\[
s(x, s(x, y)) \approx s(s(x, y), x) \approx s(x, y).
\]

Equivalently, \( s \) is a partial semilattice if for all \( x, y \), the set \( \{x, s(x, y)\} \) is closed under \( s \), and acts like a semilattice subalgebra with absorbing element \( s(x, y) \) under \( s \).

Note that unlike semilattices and 2-semilattices, partial semilattices are not necessarily Taylor operations. The binary projection \( \pi_1 \) is an extreme example of a partial semilattice operation which is not Taylor. This is a necessary feature of the definition, since we will show that any idempotent binary operation can be used to produce a partial semilattice operation (in a nontrivial way).

In order to produce partial semilattice operations, we will start by treating our binary operation as a unary function of the second variable, with the first variable treated as a (constant) parameter.
Definition 17.5. If \( t : \mathbb{A}^2 \to \mathbb{A} \) is a binary function and \( \mathbb{A} \) is finite (or profinite), then we define \( t^\infty \) to be the pointwise limit
\[
t^\infty(x, y) := \lim_{n \to \infty} t^n(x, y),
\]
where \( t^1 := t \) and \( t^{n+1}(x, y) := t(x, t^n(x, y)) \).

Proposition 17.6. For any binary term \( t \), we have
\[
t^\infty(x, t^\infty(x, y)) \approx t^\infty(x, y).
\]
If \( t \) is idempotent, then so is \( t^\infty \).

The function \( t^\infty \) now satisfies one of the two defining identities for a partial semilattice. Note that \( t^\infty \) can be computed from \( t \) in time linear in \( |A|^2 \). To find a term \( u \) which satisfies the second identity \( u(u(x, y), x) \approx u(x, y) \), we plug \( t^\infty \) into itself in a surprisingly counterintuitive way.

Proposition 17.7. If \( f \) is an idempotent binary term which satisfies the identity
\[
f(x, f(x, y)) \approx f(x, y),
\]
and if we define a term \( u \) by
\[
u(x, y) := f(x, f(y, x)),
\]
then \( u \) satisfies the identity
\[
u(u(x, y), x) \approx u(x, y).
\]

Proof. We have
\[
f(x, u(x, y)) \approx f(x, f(x, f(y, x))) \approx f(x, f(y, x)) \approx u(x, y),
\]
so
\[
u(u(x, y), x) \approx f(u(x, y), f(x, u(x, y))) \approx f(u(x, y), u(x, y)) \approx u(x, y).
\]

Finally, to get a term which satisfies both defining identities of a partial semilattice, we iterate the function \( u \) on its second variable.

Proposition 17.8. If \( u \) is an idempotent binary term which satisfies the identity
\[
u(u(x, y), x) \approx u(x, y),
\]
then \( s := u^\infty \) satisfies the identity
\[
s(x, s(x, y)) \approx s(s(x, y), x) \approx s(x, y).
\]

Proof. Define \( u^n \) as in the definition of \( u^\infty \). Then for any \( m \) we have
\[
u^m(u(x, y), x) \approx u(x, y),
\]
and on replacing \( y \) by \( u^{n-1}(x, y) \), we get
\[
u^m(u^n(x, y), x) \approx u^n(x, y)
\]
for any \( m, n \).
The full process, going from \( t \) to \( f = t^\infty \) to \( u(x, y) = f(x, f(y, x)) \) to \( s = u^\infty \), is functorial, and the final function \( s : A^2 \to A \) can be computed from \( t \) in time linear in \( |A|^2 \). Since \( s \) was defined from \( t \) in a nontrivial way, we get the following result.

**Proposition 17.9.** If \( t \) is a binary idempotent term and \( a, b \) are such that \( t(a, b) = t(b, a) = b \), then the partial semilattice term \( s \in \text{Clo}(t) \) defined by the above process also satisfies \( s(a, b) = s(b, a) = b \).

More generally, if \( B, C \) are subsets of \( \mathbb{A} \) such that for any \( x \in B \cup C \) and any \( y \in C \) we have \( t(x, y), t(y, x) \in C \), then the same holds for \( s \).

**Corollary 17.10.** If \( (b, b) \in \text{Sg}_{\mathbb{A}^2}\{ (a, b), (b, a) \} \), then there is a partial semilattice term \( s \in \text{Clo}(\mathbb{A}) \) such that \( s(a, b) = s(b, a) = b \).

Once we have a partial semilattice term \( s \) with \( s(a, b) = s(b, a) = b \), we can use it to preprocess the inputs to other \( n \)-ary functions to force them to preserve the subset \( \{a, b\} \) and act like the \( n \)-ary semilattice operation on this subset. To do this, we first need to find terms \( s_n \in \text{Clo}(s) \) which act like the \( n \)-ary semilattice operation.

**Proposition 17.11.** If \( s \) is a partial semilattice operation, then for all \( n \) there are terms \( s_n \in \text{Clo}(s) \) of arity \( n \) such that if \( \{x, x_2, ..., x_n\} = \{x, y\} \), then

\[ s_n(x, x_2, ..., x_n) \approx s(x, y). \]

**Proof.** If \( \{x, x_2, ..., x_n\} = \{x, y\} \), then the expressions \( s(x, x_2), ..., s(x, x_n) \) are all equal to either \( x \) or \( s(x, y) \), and at least one of them is equal to \( s(x, y) \), so since \( \{x, s(x, y)\} \) acts like a semilattice oriented from \( x \) to \( s(x, y) \) under \( s \), we can combine these expressions in any order to produce such a term \( s_n \).

For concreteness, we define \( s_n \) inductively, as follows: \( s_1(x) = x \), \( s_2(x, y) = s(x, y) \) and

\[ s_n(x_1, ..., x_n) = s(s_{n-1}(x_1, ..., x_{n-1}), s(x_1, x_n)). \]

Now we can use the terms \( s_n \) to preprocess the inputs to \( n \)-ary functions. If \( f \) is an \( n \)-ary term of \( \mathbb{A} \), define the term \( f_s \) by

\[ f_s(x_1, ..., x_n) = f(s_n(x_1, ..., x_n), s_n(x_2, ..., x_n, x_1), ..., s_n(x_n, x_1, ..., x_{n-1})). \]

As in the case of unary operations, we will consider the reduct \( \mathbb{A}_s \) with basic operations \( f_s \) for every term \( f \) of \( \mathbb{A} \). This reduct will be simpler in the sense that for any \( a, b \) with \( s(a, b) = s(b, a) = b \), each term \( f_s \) will act like the \( n \)-ary semilattice operation on \( \{a, b\} \). Additionally, every two-variable height 1 identity which holds in \( \mathbb{A} \) will also hold in \( \mathbb{A}_s \).

**Proposition 17.12.** Let \( \mathbb{A} = (A, (f^i)_{i \in I}) \) be a finite idempotent algebra, and let \( \Sigma \) be the set of all two-variable height 1 identities which involve both variables on each side and are satisfied in \( \mathbb{A} \). Then the operations \( (f^i_s)_{i \in I} \) of \( \mathbb{A}_s \) will also satisfy the identities in \( \Sigma \).

Additionally, if \( B, C \) are subalgebras of \( \mathbb{A} \) such that for any \( x \in B \) and any \( y \in C \) we have \( s(x, y), s(y, x) \in C \), then for any \( n \)-ary term \( f \) of \( \mathbb{A} \) and any \( x_1, ..., x_n \in B \cup C \) such that at least one \( x_i \in C \), we have \( f_s(x_1, ..., x_n) \in C \).

**Proof.** Suppose we have an identity

\[ f^i(a_1, ..., a_m) \approx f^j(b_1, ..., b_n), \]

132
with \( \{a_1, \ldots, a_m\} = \{b_1, \ldots, b_n\} = \{x, y\} \). Define \( a_1', \ldots, a_m' \) by \( a_k' = s(x, y) \) if \( a_k = x \) and \( a_k' = s(y, x) \) if \( a_k = y \), and define \( b_1', \ldots, b_n' \) similarly. Then for each \( k \), we have

\[
s_m(a_k, \ldots, a_m, a_1, \ldots, a_{k-1}) \approx a_k',
\]

and similarly for the \( b_i's \), so

\[
f^i_s(a_1, \ldots, a_m) \approx f^i(a_1', \ldots, a_m') \approx f^j(b_1', \ldots, b_n') \approx f^j(b_1, \ldots, b_n).
\]

For the last statement, we just need to check that for any \( x_1, \ldots, x_n \in \mathcal{B} \cup \mathcal{C} \) with at least one of the \( x_i's \) in \( \mathcal{C} \) we have \( s_n(x_1, \ldots, x_n) \in \mathcal{C} \) (since \( \mathcal{C} \) is closed under each term \( f \) of \( \mathcal{A} \)). This follows from the fact that \( s_n \) is defined from \( s \) in a way that involves all of its variables.

Since an algebra \( \mathcal{A} \) is Taylor if it satisfies a nontrivial system of two-variable height 1 identities, if \( \mathcal{A} \) is Taylor then \( \mathcal{A}_s \) will also be Taylor. Later, we will see that algebras with bounded width are also characterized by two-variable height 1 identities, so the same sort of implication (i.e. \( \mathcal{A} \) has bounded width implies \( \mathcal{A}_s \) has bounded width) will hold in that case as well. Algebras with few subpowers are not characterized by height 1 identities, essentially because no semilattice can have few subpowers, so such an implication fails in that case.

There are two other interesting cases which are not characterized by two-variable height 1 identities: algebras of width 1, and algebras such that the associated CSP is solved by the linear programming relaxation. It turns out that we can still prove a similar result in these cases.

**Proposition 17.13.** If \( \mathcal{A} \) has symmetric terms \( f_n \) of every arity, then it has symmetric terms \( f^s_n \) which act like the semilattice operation on each set \( \{a, b\} \) with \( s(a, b) = s(b, a) = b \).

**Proof.** Let \( f_n \) be a symmetric term of arity \( n \), for each \( n \). Then for any \( n \), let \( \sigma_1, \ldots, \sigma_n! \) be an enumeration of the permutations of \( \{1, \ldots, n\} \), and define \( f^s_n \) by

\[
f^s_n(x_1, \ldots, x_n) := f_n!\left(s_n(x_{\sigma_1(1)}, \ldots, x_{\sigma_1(n)}), \ldots, s_n(x_{\sigma_n!(1)}, \ldots, x_{\sigma_n!(n)})\right).
\]

Then \( f^s_n \) is a symmetric term of arity \( n \).

**Proposition 17.14.** If \( \mathcal{A} \) has totally symmetric terms \( f_n \) of every arity, then it has totally symmetric terms \( f^s_n \) which act like the semilattice operation on each set \( \{a, b\} \) with \( s(a, b) = s(b, a) = b \).

**Proof.** Fix \( n \). For every \( m \geq 1 \), let \( S^m_n \) be the set of \( n \)-ary terms \( t \) of \( \mathcal{A} \) such that there exist variables \( y_1, \ldots, y_l \) with \( \{y_1, \ldots, y_l\} = \{x_1, \ldots, x_n\} \) and such that for each \( i \), the number of \( j \) with \( y_j = x_i \) is at least \( m \), and

\[
t(x_1, \ldots, x_n) = s_l(y_1, \ldots, y_l).
\]

Note that for \( m' > m \) we have \( S^m_{m'} \subseteq S^m_n \), and each \( S^m_n \) is finite and nonempty, so the intersection \( S^n = \bigcap_m S^m_n \) is also finite and nonempty. Furthermore, for any \( a_1, \ldots, a_n \in \mathcal{A} \), the set of values \( \{t(a_1, \ldots, a_n) \mid t \in S^n\} \)

depends only on the set \( \{a_1, \ldots, a_n\} \). Thus we can take

\[
f^s_n(x_1, \ldots, x_n) := f_{|S^n|}(\{t(x_1, \ldots, x_n) \mid t \in S^n\}).
\]
Remark 17.1. The previous two propositions only used the fact that the restrictions of the \(s_n\)s to \(\{a, b\}\) are symmetric and totally symmetric, respectively. So they can be generalized to show that if an algebra \(A\) has symmetric/totally symmetric operations of each arity, then for every subset \(X\) of \(A\) such that some collection of terms \(t_n\) of \(A\) preserve \(X\) and have symmetric/totally symmetric restrictions to \(X\), we can find symmetric/totally symmetric operations of \(A\) which preserve \(X\) and such that their restrictions to \(X\) agree with the restrictions of the terms \(t_n\). It turns out that a similar general result holds for Taylor clones and clones of bounded width, but the proof of that will need to wait until we show that Taylor algebras have cyclic terms.

Recall that for any \(a, b\), the set \(\{a, b\}\) is a semilattice subalgebra of \(A\) iff the ternary relation 

\[(x \in \{a, b\}) \land (x = a \implies y = z)\]

defines a subalgebra of \(A^3\). We can generalize this somewhat.

**Proposition 17.15.** If \(B, C\) are subsets of \(A\), then the ternary relation 

\[(x \in B \cup C) \land (x \notin C \implies y = z)\]

defines a subalgebra of \(A^3\) iff \(B \cup C\) is a subalgebra of \(A\), and for any \(n\), any \(n\)-ary term \(f \in \text{Clo}_n(A)\) which depends on all of its inputs, and any \(x_1, ..., x_n \in B \cup C\) such that at least one \(x_i \in C\), we have \(f(x_1, ..., x_n) \in C\).

**Definition 17.16.** If \(C \leq B\) are subalgebras of \(A\) such that there exists a term \(t\) with \(t(B, C), t(C, B) \subseteq C\), then we say that \(C\) binary absorbs \(B\), and write \(C \triangleleft_{\text{bin}} B\). If for any \(n\), any \(n\)-ary term \(f \in \text{Clo}_n(A)\) which depends on all of its inputs, and any \(x_1, ..., x_n \in B\) such that at least one \(x_i \in C\), we have \(f(x_1, ..., x_n) \in C\), then we say that \(C\) strongly absorbs \(B\), and write \(C \triangleleft_{\text{str}} B\).

We can summarize the previous results in the following proposition, which shows that binary absorption and strong absorption are very nearly the same thing.

**Proposition 17.17.** If \(C \triangleleft_{\text{bin}} B\), then there is a partial semilattice term \(s\) with \(s(B, C), s(C, B) \subseteq C\), and in the reduct \(A_s\) the subalgebras \(B_s, C_s\) satisfy \(C_s \triangleleft_{\text{str}} B_s\). Furthermore, \(C \triangleleft_{\text{str}} B\) iff the ternary relation 

\[(x \in B) \land (x \notin C \implies y = z)\]

defines a subalgebra of \(A^3\) (and \(C \leq B\)).

By iteratively replacing \(A\) with reduces \(A_s\) for partial semilattice terms \(s\) quadratically many times, we can reduce to the case where for all \(a, b\), we have \((b, b) \in \text{Sg}_{A^2}\{(a, b), (b, a)\}\) iff \(\{a, b\}\) is a semilattice subalgebra of \(A\) with absorbing element \(b\).

**Definition 17.18.** We say that an idempotent algebra \(A\) has been prepared if for every pair \(a, b\) such that \((b, b) \in \text{Sg}_{A^2}\{(a, b), (b, a)\}\), the set \(\{a, b\}\) is a semilattice subalgebra of \(A\).

For algebras which have been prepared, it makes sense to define a digraph whose edges correspond to semilattice subalgebras of \(A\).

**Definition 17.19.** If \(s\) is a partial semilattice operation and \(a, b\) have \(s(a, b) = b\), then we write \(a \to_s b\), or just \(a \to b\) if \(s\) is understood (or if the algebra has been prepared).

**Theorem 17.20.** Let \(s\) be a fixed nontrivial partial semilattice term of an idempotent algebra \(A\). If \(A\) is prepared, then the following are equivalent.

(a) \(s(a, b) = b\), that is, \(a \to b\),

(b) the restriction of \(s\) to \(\{a, b\}\) acts like the semilattice operation on \(\{a, b\}\) with absorbing element \(b\),
(c) there exists \( c \) such that \( s(a, c) = b \),

(d) \( \left[ \begin{array}{c} b \\ a \end{array} \right] \in Sg_2 \left\{ \left[ \begin{array}{c} a \\ b \end{array} \right], \left[ \begin{array}{c} b \\ a \end{array} \right] \right\} \)

(e) there is a binary term \( t \) of \( \mathbb{A} \) with \( t(a, b) = t(b, a) = b \),

(f) there is a partial semilattice term \( s' \) of \( \mathbb{A} \) with \( s'(a, b) = b \),

(g) for every \( n \) and every \( n \)-ary term \( f \in \text{Clo}_n(\mathbb{A}) \) which depends on all its inputs, the restriction of \( f \) to \( \{a, b\} \) acts like the \( n \)-ary semilattice operation on \( \{a, b\} \) with absorbing element \( b \),

(h) the ternary relation \( (x \in \{a, b\}) \land (x = a \implies y = z) \) defines a subalgebra of \( \mathbb{A}^3 \).

If \( \mathbb{A} \) has not been prepared, then (a), (b), (c) are equivalent to each other, (d), (e), (f) are equivalent to each other, (g), (h) are equivalent to each other, and (g) implies (a) implies (d).

Proposition 17.21. If \( \mathbb{A} \) has been prepared and \( a, b, c \in \mathbb{A} \) have \( c \in Sg\{a, b\} \) with \( a \rightarrow c \), then \( \mathbb{A} \) has a partial semilattice term \( s \) with \( s(a, b) = c \).

Proof. Let \( s' \) be an arbitrary nontrivial partial semilattice term of \( \mathbb{A} \), and choose \( p \) a binary term of \( \mathbb{A} \) with \( p(a, b) = c \). Then take \( s(x, y) = s'(x, p(x, y)) \). We clearly have \( s(a, b) = s'(a, p(a, b)) = s'(a, c) = c \), so we just have to check that \( s \) is a partial semilattice.

If \( p \) is second projection then \( s = s' \) and we are done. Otherwise, since \( \mathbb{A} \) has been prepared, \( p \) and \( s' \) both act as the semilattice operation on \( \{x, s'(x, p(x, y))\} = \{x, s(x, y)\} \), so \( s \) also acts as the semilattice operation on \( \{x, s(x, y)\} \).

In any digraph, the strongly connected components have a natural partial order.

Definition 17.22. We say that \( b \) is reachable from \( a \) if there is a sequence \( a = a_0, a_1, \ldots, a_k = b \) such that \( a_i \rightarrow a_{i+1} \) for \( i = 0, \ldots, k - 1 \).

Proposition 17.23. If \( \mathbb{A} \) is prepared and \( s^1, \ldots, s^k \) are partial semilattice terms of \( \mathbb{A} \), then for any \( n \)-ary term \( f \in \langle s^1, \ldots, s^k \rangle \), \( f(x_1, \ldots, x_n) \) is always reachable from at least one of the variables \( x_1, \ldots, x_n \).

Definition 17.24. We say that a subset \( S \) of an algebra \( \mathbb{A} \) which has a partial semilattice operation \( s \) is upwards closed if whenever \( a \in S \) and \( a' \in \mathbb{A} \) have \( a \rightarrow_s a' \), we also have \( a' \in S \).

Definition 17.25. We say that a set \( A \) is strongly connected if for every subset \( S \subseteq A \) with \( S \neq \emptyset \), \( A \) there is an \( a \in S \) and a \( b \in A \setminus S \) such that \( a \rightarrow b \). We say that a set \( A \) is a maximal strongly connected component of an algebra \( \mathbb{A} \) if \( A \) is a strongly connected subset which is upwards closed (note that every finite upwards closed set contains at least one maximal strongly connected component). Finally, we call an element of an algebra \( \mathbb{A} \) maximal if it is contained in any maximal strongly connected component of \( \mathbb{A} \).

The main application of partial semilattice terms to CSPs is the following general idea: if a solvable instance of a CSP is arc-consistent (i.e. all relations are subdirect), then it probably has a solution where each variable is assigned a value in a maximal strongly connected component of the corresponding domain. So a basic case to try to understand is the case where every domain is a strongly connected algebra.
Remark 17.2. The digraph considered in this section is the same as the set of “thin red edges” of Andrei Bulatov’s colored graph [36] attached to any Taylor algebra. Bulatov has a different construction of a partial semilattice operation $s$ from a binary term $t$, which is still based on the counterintuitive idea of taking a function $f$ which satisfies $f(x, f(x, y)) \approx f(x, y)$ and plugging in $f(x, f(y, x))$.

18 Maximal strongly connected components and polynomial completeness

In this section we prove a few results of Andrei Bulatov [35] about the way maximal strongly connected components of partial semilattice algebras interact with binary and ternary relations. A consequence of the results of this section is that simple, strongly connected algebras are always polynomially complete. Throughout this section, we will always fix a partial semilattice operation $s$.

Theorem 18.1. Fix a partial semilattice operation $s$. Suppose $R \leq_{sd} A \times B$ is subdirect and $A, B$ are maximal strongly connected subsets of $A, B$, respectively.

(a) The set of $a$ such that $(\{a\} \times B) \cap R \neq \emptyset$ is upwards closed. In particular, if $(A \times B) \cap R$ is nonempty, then it is subdirect in $A \times B$.

(b) The set of $a$ such that $\{a\} \times B \subseteq R$ is upwards closed.

(c) If $A$ is contained in a linked component of $R$ (that is, a connected component of $R$ considered as a bipartite graph on $A \sqcup B$), $(A \times B) \cap R \neq \emptyset$, and $A, B$ are finite, then $A \times B \subseteq R$.

Additionally, the product $A \times B$ is a maximal strongly connected subset of $A \times B$.

Proof. For part (a), suppose that $(a, b) \in R$ and $b \in B$, and let $a \to a'$. Since $R$ is subdirect, there is some $b'$ with $(a', b') \in R$. Then

$$\begin{bmatrix} a' \\ s(b, b') \end{bmatrix} = s\left( \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} a' \\ b' \end{bmatrix} \right) \in R,$$

and $b \to s(b, b')$, so $s(b, b') \in B$.

For part (b), suppose that $\{a\} \times B \subseteq R$ and $a \to a'$. Let $S$ be the set of $b \in B$ such that $(a', b) \in R$, that is, $S = \pi_2(\{a'\} \times B) \cap R)$. By part (a), $S$ is nonempty. To finish, we just have to show that $S$ is upwards closed. Suppose $b \in S$ and $b \to b'$. Then by assumption we have $(a, b') \in R$, so

$$\begin{bmatrix} a' \\ \end{bmatrix} = s\left( \begin{bmatrix} a' \\ b \end{bmatrix}, \begin{bmatrix} a \\ b' \end{bmatrix} \right) \in R.$$

For part (c), suppose first that $A \times A \subseteq R \circ R^-$, where $R^- \leq B \times A$ is the reverse of $R$. Let $a$ be any element of $A$, and let $X$ be the set of $b \in B$ such that $(a, b) \in R$, that is, $X = \pi_2(\{a\} \times B) \cap R)$. By part (a), $X \cap B \neq \emptyset$, and by the finiteness of $B$, the intersection $X \cap B$ has a maximal strongly connected component $S$. Since $B$ is a maximal strongly connected component of $B$, $S$ is a maximal strongly connected component of $X$. 
By the assumption $A \times A \subseteq R \circ R^-$ and the definition of $X$, we see that $(A \times X) \cap R$ is subdirect in $A \times X$. Thus by part (b) and the fact that $\{a\} \times S \subseteq (A \times X) \cap R$, we see that $A \times S \subseteq (A \times X) \cap R$, so $A \times S \subseteq R$. Then by part (b) applied to $R^-$, we see that $A \times B \subseteq R$.

Now suppose that $A \times A \not\subseteq R \circ R^-$. From the finiteness of $A$ we see that there is some $k$ such that $A \times A \subseteq (R \circ R^-)^k$. Choose $k$ minimal, and let $R' = (R \circ R^-)^{\circ(k-1)} \leq_{sd} A^2$. Then $R'$ is equal to its own reverse $R''$, and $A \times A \subseteq R' \circ R'$ since $2(k-1) \geq k$ for $k \geq 2$. Thus the previous case applied to $R'$ shows that $A \times A \subseteq R'$, contradicting the minimality of $k$.

**Corollary 18.2.** If $\pi : A \to B$ is a surjective homomorphism of finite algebras, then the subalgebra of $A$ generated by the maximal elements of $A$ maps surjectively onto the subalgebra of $B$ generated by the maximal elements of $B$.

**Corollary 18.3.** If we start with any arc-consistent instance of CSP($A_1$, $\ldots$, $A_n$) and replace every domain and every relation by the subalgebra generated by its maximal elements, then the resulting instance will still be arc-consistent.

**Corollary 18.4.** Fix a partial semilattice operation $s$. Suppose that $R \leq_{sd} A \times B$ is a subdirect product of finite algebras $A$, $B$, and that $B$ is simple and $B = Sg(B)$, with $B$ a maximal strongly connected component of $B$. Then:

(a) if $A$ is also simple and $A = Sg(A)$ with $A$ a maximal strongly connected component of $A$, and $R \cap (A \times B) \neq \emptyset$, then $R$ is either the graph of an isomorphism or $A \times B$, and

(b) if $A$ is arbitrary and $R$ is not the graph of a homomorphism from $A$ to $B$, then there is an $a \in A$ with $\{a\} \times B \subseteq R$.

**Proof.** If $B$ is simple, then the linking congruence of $R$ on $B$ must either be the trivial congruence $0_B$, in which case $R$ is the graph of a homomorphism from $A$ to $B$, or the full congruence $1_B$, in which case $R$ is linked. In the second case, the results follow from Theorem 18.1(c). 

**Theorem 18.5.** Fix a partial semilattice operation $s$. Suppose $R \subseteq A \times B \times C$ is closed under $s$, $A$ is strongly connected, $\pi_{23}(R)$ is strongly connected, $\pi_{12}(R) = A \times B$, $\pi_{13}(R) = A \times C$, and $A, B, C$ are finite. Then $R = A \times \pi_{23}(R)$.

**Proof.** By Theorem 18.1(c), we just need to show that $R$ is linked as a subset of $A \times \pi_{23}(R)$. We will do this by showing that for any $a \to a'$ in $A$, some fork of $R$ links $a$ to $a'$ in one step.

Since $\pi_1(R) = A$, there exist $b \in B, c \in C$ such that $\langle a, b, c \rangle \in R$. Since $\pi_{13}(R) = A \times C$, there exists some $b' \in B$ such that $\langle a', b', c \rangle \in R$. Since

$$
\begin{pmatrix}
ad' \\
\quad s(b,b') \\
c
\end{pmatrix}
= s
\begin{pmatrix}
a \\
b \\
c
\end{pmatrix},
\begin{pmatrix}
a' \\
b' \\
c
\end{pmatrix} \in R,
$$

we may assume without loss of generality that $b' = s(b,b')$, that is, that $b \to b'$.

Since $\pi_{12}(R) = A \times B$, there exists some $c' \in C$ such that $\langle a, b', c' \rangle \in R$. Since

$$
\begin{pmatrix}
a \\
b' \\
s(c,c')
\end{pmatrix}
= s
\begin{pmatrix}
a \\
b \\
c
\end{pmatrix},
\begin{pmatrix}
a' \\
b' \\
c'
\end{pmatrix} \in R,
$$

137
we may assume without loss of generality that \( c' = s(c, c') \), that is, that \( c \to c' \).

Since \((a', b', c')\) and \((a, b', c')\) are in \( R \), we have
\[
\begin{bmatrix}
a' \\
b' \\
c'
\end{bmatrix}
= s
\begin{bmatrix}
a' \\
b' \\
c
\end{bmatrix},
\begin{bmatrix}
a' \\
b' \\
c'
\end{bmatrix}
\in R.
\]

Thus both \( a \) and \( a' \) meet \((b', c') \in \pi_{23}(R)\).

\[\text{Corollary 18.6.}\: \text{Fix a partial semilattice operation } s. \text{ Suppose } R \subseteq A_1 \times \cdots \times A_n \text{ is closed under } s, A_1 \text{ is strongly connected, } \pi_{[2,n]}(R) \text{ is strongly connected, } \pi_{1i}(R) = A_1 \times A_i \text{ for } i \in [2, n], \text{ and } A_i \text{ are finite for all } i. \text{ Then } R = A_1 \times \pi_{[2,n]}(R).\]

\[\text{Corollary 18.7.}\: \text{Fix a partial semilattice operation } s. \text{ Suppose } R \subseteq A_1 \times \cdots \times A_n \text{ is closed under } s, \text{ all } A_i \text{ are strongly connected, } \pi_{ij}(R) = A_i \times A_j \text{ for all } i \neq j, \text{ and } A_i \text{ are finite for all } i. \text{ Then } R = A_1 \times \cdots \times A_n.\]

\[\text{Corollary 18.8.}\: \text{Fix a partial semilattice operation } s. \text{ If } A \text{ is simple and is generated by a finite maximal strongly connected component } A, \text{ then } A \text{ is polynomially complete.}\]

\[\text{Proof.}\: \text{We just need to show that every relation } R \leq A^n \text{ which contains the set of constant tuples } \Delta_n = \{(a, ..., a) \mid a \in A\} \text{ is an intersection of equality relations. First consider the case } n = 2. \text{ From the assumption that } A \text{ is simple we see that either } R \text{ is the equality relation, or } R \text{ is linked. If } R \text{ is linked, then Theorem 18.1(c) and the fact that } R \text{ contains } \Delta_2 \text{ implies that } A \times A \subseteq R, \text{ and from the assumption } A = \text{Sg}(A) \text{ we see that } R = A \times A. \]

\[\text{Now consider the case } n \geq 3. \text{ If any two-variable projection } \pi_{ij}(R) \text{ is the equality relation, then we can ignore one of the coordinates } i, j, \text{ so we may assume without loss of generality that } \pi_{i,j}(R) = A \times A \text{ for all } i, j. \text{ Let } a \text{ be any element of } A, \text{ and let } R \text{ be a maximal strongly connected component of } R \text{ which is reachable from } (a, ..., a). \text{ Then for any } i, j \text{ we must have } \pi_{i,j}(R) = A \times A, \text{ so by the previous corollary we have } R = A^n. \text{ Thus } A^n \subseteq R, \text{ and from the assumption } A = \text{Sg}(A) \text{ we see that } R = A^n.\]

\[\text{Example 18.1.}\: \text{The reader may be wondering whether we can weaken the assumption that } \pi_{23}(R) \text{ is strongly connected from Theorem 18.5 to the assumption that } B, C \text{ are strongly connected. It seems plausible that if } B, C \text{ are both strongly connected and } \pi_{23}(R) \text{ is a subdirect product of } B \text{ and } C, \pi_{23}(R) \text{ might automatically be strongly connected.} \]

\[\text{However, there is an example of a strongly connected 2-semilattice } A \text{ and a subdirect product } R \leq_{sd} A^2 \text{ which is not strongly connected. The 2-semilattice } A \text{ is pictured below.}\]
The missing values are given by $s(b,c) = s(b,f) = s(e,c) = a$.

If we let $\theta \leq_{sd} A^2$ be the smallest congruence containing $(b,c)$, then $\theta$ corresponds to the partition \{a, b, c\}, \{d\}, \{e\}, \{f\}, and $A/\theta$ is a four element tournament. Considering $\theta$ as an algebra, we find that $\theta$ is not strongly connected: $(b,c)$ and $(e,b)$ are incomparable minimal elements of $\theta$, and the remaining elements of $\theta$ form a maximal strongly connected component.

Example 18.2. Here we will give an example of a subdirect product of strongly connected algebras which has two maximal strongly connected components (such an example is necessarily not a 2-semilattice, since every 2-semilattice has a unique maximal strongly connected component).

As in the previous example, we will consider a congruence $\theta$ on a six-element algebra $A$. This time $A/\theta$ will be the three-element rock-paper-scissors algebra, and every congruence class of $A$ will have two elements, with $s$ acting as $\pi_1$ on the congruence class. As a digraph, $A$ is just a directed six-cycle, pictured below.

Given the above digraph structure and the assumption that there is a congruence $\theta$ corresponding to the partition \{a1, a2\}, \{b1, b2\}, \{c1, c2\}, there is only one way to fill in the values of the partial semilattice operation $s$. The reader can check that the congruence $\theta$, considered as a subalgebra of $A^2$, has two maximal strongly connected components which are both isomorphic to $A$.

Despite the above examples, we do at least have the following result, which is important for understanding how restricting to maximal strongly connected components interacts with cycle-consistency.

**Theorem 18.9.** Fix a partial semilattice operation $s$. Suppose that $R \subseteq A \times A$ is closed under $s$, $A$ is finite and strongly connected, and $R$ contains the diagonal $\Delta_A = \{(a,a) \, | \, a \in A\}$. Then $R$ has a maximal strongly connected component which contains $\Delta_A$.

**Proof.** Since $\Delta_A$ is strongly connected, it’s enough to show that if $(a,b)$ is reachable from $(a,a)$ in $R$, then some element $(c,c)$ of $\Delta_A$ is reachable from $(a,b)$ in $R$. We will define a unary polynomial $\phi$ of $R$ such that $\phi((a,a)) = (a,b)$ and such that $\phi(x)$ is reachable from $x$ in $R$ for all $x \in R$.

To construct $\phi$, choose some sequence $(a_i, b_i) \in R$ such that $(a,a) = (a_0, b_0)$, $(a_i, b_i) \rightarrow (a_{i+1}, b_{i+1})$ for all $i$, and $(a_k, b_k) = (a,b)$ for some $k$. Then define $\phi$ by

$$\phi(x) = s \left( s \left( \cdots s \left( s \left( x, \begin{bmatrix} a_1 \\ b_1 \end{bmatrix} \right), \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \right) \cdots, \begin{bmatrix} a_k \\ b_k \end{bmatrix} \right), \begin{bmatrix} a_{k+1} \\ b_{k+1} \end{bmatrix} \right) \cdots, \begin{bmatrix} a_{i+1} \\ b_{i+1} \end{bmatrix} \right) \cdots, \begin{bmatrix} a \\ b \end{bmatrix} \right).$$

Note that since $\phi((a,a)) = (a,b)$, we have $\pi_1(\phi((a,x))) = a$ for all $x \in A$.

Since $A$ is finite, we can find $m \geq 1$ such that $\phi^{\circ 2m} = \phi^{\circ m}$. Define another unary polynomial $\phi_\Delta$ of $R$ by

$$\phi_\Delta(x) = s \left( s \left( \cdots s \left( s \left( x, \begin{bmatrix} b_1 \\ b_1 \end{bmatrix} \right), \begin{bmatrix} b_2 \\ b_2 \end{bmatrix} \right) \cdots, \begin{bmatrix} b_k \\ b_k \end{bmatrix} \right), \begin{bmatrix} b_{k+1} \\ b_{k+1} \end{bmatrix} \right) \cdots, \begin{bmatrix} b_{i+1} \\ b_{i+1} \end{bmatrix} \right) \cdots, \begin{bmatrix} b \end{bmatrix} \right).$$
that is, by replacing each \((a_i, b_i)\) in the definition of \(\phi\) by \((b_i, b_i)\). Then if \(\phi^\circ m ((a, a)) = (a, c)\), we have
\[
\phi^\circ_{\Delta} \left( \phi^\circ_{\Delta}^{(m-1)} \left[ \begin{array}{c} a \\ b_i \\ c \\ \end{array} \right] \right) = \phi^\circ_{\Delta} \left[ \begin{array}{c} a \\ c \end{array} \right] = [c].
\]
Thus \((c, c)\) is reachable from \((a, b)\) in \(R\).

**Corollary 18.10.** If we start with any cycle-consistent instance of CSP\((A_1, \ldots, A_n)\) and replace every domain and every relation by the subalgebra generated by its maximal elements, then the resulting instance will still be cycle-consistent.

**Proof.** By Theorem 18.1(a), we just need to check this in the special case where our cycle-consistent instance is a cycle of binary relations \(R_i \leq_{sd} A_i \times A_{i+1}\) with indices taken modulo \(n\). Let \(R \leq_{sd} A_1 \times \cdots \times A_n \times A_1\) be the relation given by the formula
\[
(x_1, x_2) \in R_1 \land \cdots \land (x_n, x_{n+1}) \in R_n.
\]
The assumption that the instance is cycle-consistent implies that \(\Delta_{A_1} \subseteq \pi_{1,n+1} R\). Set \(R_{\Delta} = \pi_{1,n+1} R\).

For any algebra \(A\), let \(A^{\max}\) denote the subalgebra of \(A\) generated by the maximal elements of \(A\). We see from Theorem 18.1(a) that \(\pi_{i,i+1}(R^{\max}) = R^{\max}_i\) for each \(i\) and that \(\pi_{1,n+1}(R^{\max}) = R^{\max}_n\). By Theorem 18.9 we have \(\Delta_{A_1}^{\max} \subseteq R_{\Delta}^{\max}\), so the new instance is cycle-consistent at the first variable.

19 2-semilattices, spirals, and ancestral algebras

In this section we’ll discuss a pretty general class of partial semilattice algebras which are nice enough for the associated CSP to have bounded width, due to Bulatov [29]. Following the strategy of replacing domains of variables with the subalgebras generated by their maximal elements, and noting that many of the structural results proved in the preceding section apply best to strongly connected algebras, we see that it would be quite convenient if every domain of every variable in our CSP has a unique maximal strongly connected component. The most straightforward examples of algebras with this property are 2-semilattices.

**Definition 19.1.** A binary operation \(s\) is a 2-semilattice operation if it satisfies the identities
\[
s(x, y) \approx s(y, x), \quad s(x, s(x, y)) \approx s(x, y), \quad s(x, x) \approx x.
\]
In other words, a 2-semilattice is a partial semilattice operation which is also commutative.

**Proposition 19.2.** An algebra \(A = (A, s)\) is a 2-semilattice iff for all \(a, b \in A\), the subalgebra \(\text{Sg}_A \{a, b\}\) is a semilattice under \(s\).

**Proposition 19.3.** If \(A = (A, s)\) is a finite 2-semilattice, then \(A\) has a unique maximal strongly connected component.

**Proof.** If \(a, b\) are any two maximal elements of \(A\), then \(s(a, b) = s(b, a)\) is reachable from both \(a\) and \(b\), so \(a\) and \(b\) must be in the same maximal strongly connected component.
The first difficult results about bounded width CSPs were proved for 2-semilattices. However, the proofs only depended on the fact that every 2-semilattice has a unique maximal strongly connected component. Bulatov [29] calls this the "maximal red component condition". I’ve chosen to call such algebras “ancestral” instead, because they can be equivalently defined as follows.

**Definition 19.4.** An idempotent algebra \( A \) with a fixed partial semilattice operation \( s \) is called **ancestral** if for all \( a, b \in A \), there is some \( c \in \text{Sg}_{A}\{a, b\} \) which is reachable from both \( a \) and \( b \). We call any such \( c \) a **common ancestor** of \( a \) and \( b \).

**Proposition 19.5.** A finite idempotent algebra \( A \) is ancestral iff every proper subalgebra of \( A \) has a unique maximal strongly connected component.

A nice generalization of 2-semilattices is the collection of algebras which I call “spirals”. Spirals are defined in terms of a single commutative binary operation, so they can be described more rapidly than general ancestral algebras. As we will see later, a minimal Taylor clone is ancestral if and only if it is a minimal spiral, so we would not lose too much generality by restricting the study of ancestral algebras to the study of spirals.

**Definition 19.6.** An algebra \( A = (A, f) \) is a **spiral** if \( f \) is a commutative idempotent binary operation and every subalgebra of \( A \) which is generated by two elements either has size two or has a surjective homomorphism to the free semilattice on two generators.

**Example 19.1.** Here we give an example of a minimal spiral \( A_6 \) which is not a 2-semilattice.

<table>
<thead>
<tr>
<th>( A_6 )</th>
<th>( a )</th>
<th>( b )</th>
<th>( c )</th>
<th>( d )</th>
<th>( e )</th>
<th>( f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>( a )</td>
<td>( c )</td>
<td>( e )</td>
<td>( d )</td>
<td>( e )</td>
<td>( d )</td>
</tr>
<tr>
<td>( b )</td>
<td>( c )</td>
<td>( b )</td>
<td>( c )</td>
<td>( f )</td>
<td>( f )</td>
<td></td>
</tr>
<tr>
<td>( c )</td>
<td>( e )</td>
<td>( c )</td>
<td>( c )</td>
<td>( e )</td>
<td>( c )</td>
<td></td>
</tr>
<tr>
<td>( d )</td>
<td>( d )</td>
<td>( c )</td>
<td>( c )</td>
<td>( d )</td>
<td>( d )</td>
<td></td>
</tr>
<tr>
<td>( e )</td>
<td>( e )</td>
<td>( f )</td>
<td>( e )</td>
<td>( d )</td>
<td>( e )</td>
<td>( f )</td>
</tr>
<tr>
<td>( f )</td>
<td>( d )</td>
<td>( f )</td>
<td>( c )</td>
<td>( d )</td>
<td>( f )</td>
<td>( f )</td>
</tr>
</tbody>
</table>

Every proper subalgebra of \( A_6 \) is a 2-semilattice - in fact, every pair of elements other than \( \{a, b\} \) generates a two or three element semilattice subalgebra of \( A_6 \). The pair \( \{a, b\} \) generates \( A_6 \), and \( A_6 \) has a congruence \( \theta \) corresponding to the partition \( \{a\}, \{b\}, \{c, d, e, f\} \) such that \( A_6/\theta \) is isomorphic to the free semilattice on two generators.

The reader may check that any nonempty subset \( S \) of \( A_6 \) which is closed under multiplication by \( a \) and by \( b \) must necessarily contain all four of \( c, d, e, f \) - using this observation, it is easy to check that \( \text{Clo}(A_6) \) contains no nontrivial proper subclones.

**Theorem 19.7.** If \( A = (A, f) \) is a spiral, then for any partial semilattice term \( s \in \text{Clo}(f) \) which is defined nontrivially in terms of \( f \), the reduct \( A_s = (A, s) \) is ancestral.

**Proof.** We prove this by induction on the size of \( A \). Let \( a, b \) be any two elements of \( A \). If \( \text{Sg}_{A}\{a, b\} \) has size two, then since \( f \) is commutative we must either have \( a \to b \) or \( b \to a \), so one of \( a, b \) is a common ancestor of \( a \) and \( b \).

Otherwise, by the definition of a spiral, there is a surjective homomorphism \( \alpha \) from \( \text{Sg}_{A}\{a, b\} \) to the free semilattice on two generators. Clearly \( a \) and \( b \) must be sent to the two generators of the free semilattice by \( \alpha \), say \( \alpha(a) = x \) and \( \alpha(b) = y \), and every nontrivial binary term \( t \in \text{Clo}(f) \)
must have \( \alpha(t(a, b)) = t(x, y) = f(x, y) \). Thus the kernel of \( \alpha \) has congruence classes \( \{a\}, \{b\}, \) and \( S = Sg_{A}\{a, b\} \setminus \{a, b\} \), and \( S \) is a binary absorbing subalgebra of \( A \) with respect to \( f \).

Since \( S \) is a binary absorbing subalgebra of \( A \) with respect to \( f \) and \( s \in \text{Clo}(f) \) is defined nontrivially, we must have \( s(a, b), s(b, a) \in S \). Since \( |S| \leq |A| - 2 \), we can apply the inductive hypothesis to see that \( s(a, b), s(b, a) \) have a common ancestor in \( Sg(S, a) \{s(a, b), s(b, a)\} \subseteq Sg_{A}\{a, b\} \).

\[ \tag*{□} \]

\textit{Example 19.2.} An example of an ancestral algebra which is not a 2-semilattice or a spiral is the algebra \( A_4 = (\{a, b, c, d\}, s) \), where \( s \) is the partial semilattice operation described below.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>b</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>c</td>
<td>d</td>
<td>c</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>d</td>
<td>a</td>
<td>a</td>
<td>d</td>
<td>d</td>
</tr>
</tbody>
</table>

The algebra \( A_4 \) has the cyclic automorphism \((a b c d)\), and is generated by the pair \( a, c \), since \( s(a, c) = b, s(c, a) = d \). The binary term \( s' \) given by

\[ s'(x, y) := s(x, s(y, x)) \]

is another (nontrivial) partial semilattice term of \( A_4 \), such that \( s'(a, c) = a, s'(c, a) = c \). So the reduct \((\{a, b, c, d\}, s')\) of \( A_4 \) is not an ancestral algebra, as it has the subalgebra \((\{a, c\}, s')\) which has the two maximal strongly components \( \{a\} \) and \( \{c\} \).

It is easy to check that \( A_4 \) is simple, and every proper subalgebra of \( A_4 \) is a two element semilattice. By Corollary 18.8 \( A_4 \) is polynomially complete, and in fact Theorem 18.1 and Theorem 18.5 imply that every subdirect relation \( R \leq_{sd} A_4^n \) can be written as an intersection of two variable relations, each of which is the graph of an automorphism of \( A_4 \). In particular, if we consider the ternary relation

\[ R_{ac} = Sg_{A^3} \left\{ \begin{array}{ccc} a & a & b \\ a & c & b \\ c & a & a \end{array} \right\}; \]

we find that \( R_{ac} = A_4^3 \). Since there is an automorphism of \( A_4 \) which interchanges \( a \) and \( c \), we see in particular that there are ternary terms \( g, g' \in \text{Clo}(A_4) \) such that \( \{a, c\} \) is closed under \( g \) and \( g' \), with \((\{a, c\}, g)\) a two element majority algebra and \((\{a, c\}, g')\) a two element affine algebra. Either of the reducts \((\{a, b, c, d\}, g)\) or \((\{a, b, c, d\}, g')\) defines a Taylor algebra, since \( g \) satisfies the identity

\[ g(x, x, y) \approx g(x, y, x) \approx g(y, x, x) \approx s'(x, y), \]

and \( g' \) satisfies the similar identity

\[ g'(x, x, y) \approx g'(x, y, x) \approx g'(y, x, x) \approx s'(y, x). \]

\textbf{Proposition 19.8.} Every quotient of an ancestral algebra is ancestral.

\textbf{Theorem 19.9.} If \( A_1, \ldots, A_n \) are ancestral algebras with partial semilattice operation \( s \), then so is \( A_1 \times \cdots \times A_n \).
Proof. We prove this by induction on \( n \). Let \( a, b \in A_1 \times \cdots \times A_n \). Since \( A_1 \) is ancestral, there is some \( c_1 \in S_{F_1}(a_1, b_1) \) which is reachable from both \( a_1 \) and \( b_1 \). Lifting the path from \( a_1 \) to \( c_1 \) to a path from \( a \) to some element \( c' \in S_{F_1}(a, b) \) with \( c_1' = c_1 \), and lifting the path from \( b_1 \) to \( c_1 \) to a path from \( b \) to some \( c'' \in S_{F_2}(a, b) \) with \( c_1'' = c_1 \), we see that we just need to find a common ancestor of \( c' \) and \( c'' \). Since \( c_1' = c_1'' \) and \( A_1 \) is idempotent, we see that \( c', c'' \) have a common ancestor so long as \( A_2 \times \cdots \times A_n \) is ancestral, which follows from the inductive hypothesis.

Corollary 19.10. If \( \mathbb{A} \) is ancestral and \( \mathbb{B} \in HSP_{fin}(\mathbb{A}) \), then \( \mathbb{B} \) is also ancestral.

It turns out that ancestral algebras can be defined entirely in terms of collections of partial semilattice operations.

Theorem 19.11. A finite idempotent algebra \( \mathbb{A} \) with a fixed partial semilattice operation \( s \) is ancestral iff for some \( m \geq n \geq 0 \) it has a sequence of partial semilattice terms \( p_1, p_2, \ldots, p_m \) such that

- \( a \rightarrow_{s} p_i(a, b) \) for all \( a, b \in \mathbb{A} \) and all \( i \),
- \( a \rightarrow_{s} b \) implies \( p_i(a, b) = b \) for all \( i \), and
- if we define binary operations \( f_i \) recursively by \( f_0(x, y) := s(x, y) \) and
  \[
  f_i(x, y) := p_i(f_{i-1}(x, y), f_{i-1}(y, x))
  \]
  for \( i \geq 1 \), then \( f_m(x, y) \approx f_n(y, x) \).

Proof. That the existence of such a sequence implies \( \mathbb{A} \) is ancestral follows from the fact that for any \( a, b \), each \( f_i(a, b) \) is reachable from \( a \) and each \( f_j(b, a) \) is reachable from \( b \).

For the converse direction, let \( \mathbb{F} = F_{\mathbb{A}}(x, y) \leq \mathbb{A}^{x, y} \) be the free algebra on two generators in the variety generated by \( \mathbb{A} \). Since \( \mathbb{F} \in SP_{fin}(\mathbb{A}) \), \( \mathbb{F} \) is ancestral, so there is some sequence of elements \( f_0, ..., f_n \in \mathbb{F} \) with \( f_0(x, y) = s(x, y) \), such that each \( f_{i-1} \rightarrow_{s} f_i \), each \( f_i \in S_{\mathbb{F}}(f_{i-1}(x, y), f_{i-1}(y, x)) \), and such that the subset \( S \) of elements of the subalgebra \( \mathbb{S} = S_{\mathbb{F}}(f_n(x, y), f_n(y, x)) \) which are reachable from \( f_n \) in \( \mathbb{S} \) is minimal given these constraints. Then \( S \) must be strongly connected, and for every \( g \in S \) we must have \( S = S_{\mathbb{F}}(g(x, y), g(y, x)) \). Thus we can extend our sequence \( f_0, ..., f_n \) by \( f_{n+1}, ..., f_m \) such that each \( f_{i-1} \rightarrow_{s} f_i \), and \( f_m(x, y) \approx f_n(y, x) \), and we will automatically have \( f_i \in S_{\mathbb{F}}(f_{i-1}(x, y), f_{i-1}(y, x)) \) for each \( i \).

Note that \( f_{i-1} \rightarrow_{s} f_i \) implies \( p_i(a, b) \) and \( f_{i-1}(x, y) = p_i(f_{i-1}(x, y), f_{i-1}(y, x)) \), by the argument of Proposition 17.21.

Note that the reduct with basic operations \( s, f_i \) is ancestral, and has the property that \( a \rightarrow_{s} b \) implies \( f_i(a, b) = f_i(b, a) = b \) for all \( i \), so \( \{a, b\} \) is a semilattice subalgebra with respect to any nontrivial binary term in \( Clo(f_0, ..., f_m) \). Thus we may assume without loss of generality that \( a \rightarrow_{s} b \) implies \( p_i(a, b) = b \) for all \( i \), and then the argument of Proposition 17.21 implies that each \( p_i \) is a partial semilattice term.

In fact, we can go further: every ancestral algebra has an ancestral reduct which is prepared. Recall that \( \mathbb{A} \) is prepared if for all \( a, b \in \mathbb{A} \), we have \((b, b) \in S_{\mathbb{A}^2}\{\{(a, b), (b, a)\}\} \) iff \( \{a, b\} \) is a semilattice subalgebra of \( \mathbb{A} \) with \( a \rightarrow b \).

Theorem 19.12. Every finite ancestral algebra \( \mathbb{A} \) has a reduct which is prepared and ancestral.
Proof. Let $s, f_i$ be as in Theorem [19.11] and assume without loss of generality that these are the basic operations of $\mathbb{A}$. Suppose there is a pair $a, b \in \mathbb{A}$ with $(b, b) \in Sg_{\mathbb{A}^2}\{a, b\}$ but $s(a, b) \neq b$. Let $s'$ be a partial semilattice term with $s'(a, b) = b$. Then $c \rightarrow_s d$ implies $c \rightarrow_{s'} d$, and if we define
\[
f'_0(x, y) := s'(x, y)
\]
and
\[
f'_i(x, y) := f_{i-1}(s'(x, y), s'(y, x))
\]
for $i \geq 1$, then the reduct with basic operations $s', f'_i$ is an ancestral algebra (with respect to $s'$) with strictly more semilattice subalgebras than $\mathbb{A}$.

Due to the structural simplifications we can obtain by passing to reducts, it makes sense to focus on ancestral algebras such that no proper reduct is also ancestral.

Definition 19.13. A finite algebra $\mathbb{A}$ is called a minimal ancestral algebra if $\mathbb{A}$ is ancestral, and no proper reduct of $\mathbb{A}$ is ancestral.

Since every minimal ancestral algebra is automatically prepared, we don’t need to specify a particular choice of partial semilattice operation to define the digraph of semilattice subalgebras.

Proposition 19.14. Every finite ancestral algebra has a reduct which is a minimal ancestral algebra.

Proof. Whether an algebra is ancestral only depends on the collection of partial semilattice operations in its clone. Since there are only finitely many partial semilattice operations on a given finite set, we don’t need to worry about infinite descending chains of smaller and smaller ancestral reducts.

Proposition 19.15. If $\mathbb{A}$ is a minimal ancestral algebra and $\mathbb{B} \in HSP_{\text{fin}}(\mathbb{A})$, then $\mathbb{B}$ is also a minimal ancestral algebra.

Proof. Let $f_i$ be terms for $\mathbb{A}$ as in Theorem [19.11]. If we can find a proper reduct of $\mathbb{B}$ which is ancestral, then there is a sequence of terms $f'_i$ of this reduct such that $f'_0(x, y) \approx x$, $f'_i(x, y) \rightarrow f'_{i+1}(x, y)$, and $f'_m(a, b) = f'_n(b, a)$ holds for all $a, b \in \mathbb{B}$. Then if we define additional terms $f'_{m+i}$ by
\[
f'_{m+i}(x, y) := f_i(f'_m(x, y), f'_i(y, x)),
\]
we see that these terms $f'_0, ..., f'_m, f'_{m+1}, ...$ generate the same reduct on $\mathbb{B}$ as $f'_0, ..., f'_m$, and generate an ancestral reduct of $\mathbb{A}$.

Theorem 19.16. If $\mathbb{A}$ is a minimal ancestral algebra, then for any $a, b \in \mathbb{A}$, if $S$ is the maximal strongly connected component of $Sg_{\mathbb{A}}\{a, b\}$, then we have $Sg_{\mathbb{A}}\{a, b\} = S \cup \{a, b\}$. If $\{a, b\} \not\subseteq S$, then $Sg_{\mathbb{A}}\{a, b\}$ has a semilattice quotient with $S$ as a congruence class which acts as the top element.

Proof. Choose terms $f_i$ as in Theorem [19.11]. Let $F = F_{\mathbb{A}}(x, y)$ be the free algebra on two generators in the variety generated by $\mathbb{A}$. Pick any element $g(x, y)$ in the maximal strongly connected component of $F$, and note that since $g(x, y)$ is reachable from both $x$ and $y$ in $\mathbb{F}$, every term $t(x_1, ..., x_k) \in Clo(g)$ which depends on all its inputs has the property that $t(x_1, ..., x_k)$ is reachable from each $x_i$ in $F_{\mathbb{A}}(x_1, ..., x_k)$.

144
Applying the semilattice iteration argument, we get a partial semilattice term \( s'(x, y) \in \text{Clo}(g) \), which is reachable from each of \( x, y \), and \( g(x, y) \) in \( F \). In particular, we see that \( s'(x, y) \) is contained in the maximal strongly connected component of \( F \), and if we define terms \( f'_i \) by

\[
f'_0(x, y) := s'(x, y)
\]

and

\[
f'_i(x, y) := f_{i-1}(s'(x, y), s'(y, x))
\]

for \( i \geq 1 \), then the reduct with basic operations \( f'_i \) is an ancestral algebra, and each \( f'_i(x, y) \) is contained in the maximal strongly connected component of \( F \). Thus the clone generated by the \( f'_i \)'s must be equal to the clone of \( A \), and we see that every element of \( F \) is either equal to one of \( x, y \) or is contained in the maximal strongly connected component of \( F \).

**Corollary 19.17.** If \( A \) is a minimal ancestral algebra, then the maximal strongly connected component of \( A \) is a strongly absorbing subalgebra of \( A \).

There is a sense in which even the class of minimal ancestral algebras is unnecessarily large: it contains algebras such as the algebra \( A_4 \) from Example [19.2] which have proper Taylor reducts with two element majority or affine subalgebras.

**Theorem 19.18.** Suppose \( A \) is a minimal ancestral algebra which is generated by \( a \) and \( b \), is strongly connected, and is simple. Then there are ternary terms \( g, g' \in \text{Clo}(A) \) such that \( \{a, b\} \) is closed under \( g \) and \( g' \), \( \{a, b\}, g \) is a two element majority algebra, and \( \{a, b\}, g' \) is a two element affine algebra.

**Proof.** Let \( S = \text{Sg}_A \{a, b\} \). If \( S \) is linked, then by Theorem [18.1(c)] we must have \( (b, a) \in S \), so \( a \rightarrow b \), a contradiction. Otherwise, \( S \) is the graph of an automorphism swapping \( a \) and \( b \). In this case, the ternary relation \( R = \text{Sg}_A \{(a, a, b), (a, b, a), (b, a, a)\} \) has \( (a, a), (a, b), (b, a) \in \pi_{i,j}(R) \) for each \( i, j \), so by Theorem [18.1(c)] we have \( \pi_{i,j}(R) = A^2 \), and then by Theorem [18.5] we have \( R = A^3 \). Thus \( (a, a, a) \in R \) and \( (b, b, b) \in R \), and we can take \( g, g' \) to be ternary terms of \( A \) which witness these facts.

Later we will see that the above result implies that a minimal ancestral algebra which is both strongly connected and generated by two elements has a proper Taylor reduct (and, in fact, has a proper bounded width reduct). For now we will show that minimal ancestral algebras which avoid this situation are actually spirals.

**Theorem 19.19.** If \( A \) is a minimal ancestral algebra such that for all \( a, b \) the subalgebra \( \text{Sg}_A \{a, b\} \) has no strongly connected quotient, then \( A \) is term equivalent to a spiral.

**Proof.** Let \( s \) be a nontrivial partial semilattice operation on \( A \). Define a sequence of terms \( f_i \) inductively by \( f_0 := s \) and

\[
f_{i+1}(x) := f_i(s(x, y), s(y, x)).
\]

We will show by induction on \( |A| \) that for each \( a, b \in A \), there is an \( n \) such that \( f_n(a, b) = f_n(b, a) \). To see this, note that by Theorem [19.16] for any \( a, b \) the subalgebra generated by \( s(a, b), s(b, a) \) is contained in the maximal strongly connected component \( S \) of \( \text{Sg}_A \{a, b\} \), so as long as \( S \neq A \) we can apply the induction hypothesis to see that there is some \( i \) such that \( f_i(s(a, b), s(b, a)) = f_i(s(b, a), s(a, b)) \),

145
and for this $i$ we then have $f_{i+1}(a, b) = f_{i+1}(b, a)$.

Thus there is some $n$ such that $f = f_n$ is commutative (in fact, we can take $n = |A|$). To finish, we need to show that if $A$ is generated by two elements $a, b$ with $|A| > 2$, then the maximal strongly connected component $S$ of $A$ does not contain either of $a, b$. To this end, suppose for a contradiction that $S$ contains $b$. Let $S = Sg_A\{(a, b), (b, a)\}$. If $S$ is contained in a linked component of $S$, then by Theorem 18.1(c) we must have $(b, b) \in S$, so $a \rightarrow b$, a contradiction. Otherwise, the linking congruence $\theta \in \text{Con}(A)$ of $S$ has $|S/\theta| > 1$ and $b/\theta \in S/\theta$, and so we may assume without loss of generality that $\theta$ is trivial. But if $\theta$ is trivial, then $A$ has an automorphism which interchanges $a$ and $b$, so $S$ contains both $a$ and $b$, so $A$ is both strongly connected and generated by two elements, a contradiction.

20 Cycle-consistency solves ancestral CSPs

In this section we will prove that any cycle-consistent instance of an ancestral CSP has a solution. This proof is a simple case of Kozik’s proof [78] of the fact that cycle-consistency solves CSPs over templates with bounded width: the main purpose of presenting the argument in this special case is to allow the reader to focus on the overall proof strategy before getting into the technical algebraic details.

The ingredients which we will need for the proof are the following facts about ancestral algebras.

- Every ancestral algebra $A$ has a unique maximal strongly connected component $A^{\text{max}}$ (Proposition 19.5).
- If $\pi : A \rightarrow B$ is a surjective homomorphism, then $\pi(A^{\text{max}}) = B^{\text{max}}$ (Corollary 18.2 to Theorem 18.1(a)).
- If $R \leq sd A \times B$ and $A^{\text{max}}$ is contained in a linked component of $R$, then $R^{\text{max}} = A^{\text{max}} \times B^{\text{max}}$ (Theorem 18.1(c)).
- In particular, if $R \leq sd A \times B$, $A$ is generated by $A^{\text{max}}$, $B$ is generated by $B^{\text{max}}$, and $B$ is simple, then $R$ is either the graph of a homomorphism $A \rightarrow B$ or $R = A \times B$ (Corollary 18.4).
- If $R \leq sd A \times B \times C$ has $\pi_{12}(R) = A \times B$ and $\pi_{13}(R) = A \times C$, then $R^{\text{max}} = A^{\text{max}} \times \pi_{23}(R)^{\text{max}}$ (Theorem 18.5).
- Applying the above inductively, if $R \leq sd A_1 \times \cdots \times A_n$ has $\pi_{ij}(R) = A_i \times A_j$ for all $i \neq j$, then $R^{\text{max}} = A_1^{\text{max}} \times \cdots \times A_n^{\text{max}}$ (Corollary 18.7).
- If $R \leq sd A \times A$ and $R$ contains the diagonal $\Delta_A$, then $\Delta_A^{\text{max}} \subseteq R^{\text{max}}$ (Theorem 18.9).
- If we start with any cycle-consistent instance of CSP$(A_1, \ldots, A_n)$ and replace every domain and every relation by the subalgebra generated by its maximal elements, then the resulting instance will still be cycle-consistent (Corollary 18.10).

If we assume that our algebras are minimal ancestral (rather than just ancestral), then each $A^{\text{max}}$ becomes a subalgebra (Corollary 19.17), which slightly simplifies the arguments. We won’t use this simplification, but the reader should keep it in mind.

The general strategy is to start with a cycle-consistent instance, and to find a way to shrink some of the variable domains and relations to get a strictly smaller cycle-consistent instance. Eventually,
we reach a situation where all the variable domains have size 1 and the instance is still cycle-consistent - at this point, there is obviously a solution to the CSP. We have already seen that by shrinking variable domains, we can reach a situation where each variable domain \( A_x \) is generated by \( A_x^{\max} \) (the last bullet point above).

To finish the argument, we need to find another strategy for reducing the variable domains when each \( A_x = \text{Sg}(A_x^{\max}) \). The intuition is that if \( A_x = \text{Sg}(A_x^{\max}) \), then there is some congruence \( \theta_x \in \text{Con}(A_x) \) such that \( A_x/\theta_x \) is simple, and in fact \( A_x/\theta_x \) will be polynomially complete by Corollary \[18.8\]. Since polynomially complete algebras should have few interesting subdirect relations, it’s plausible that we can replace the domain \( A_x \) with an arbitrary congruence class of \( \theta_x \), and always obtain a cycle-consistent instance.

So fix a variable \( x \) with \( |A_x| > 1 \), a maximal congruence \( \theta_x \) in \( \text{Con}(A_x) \), and a congruence class \( A'_x \) of \( \theta_x \). We now have to restrict the other variable domains in order to, at the very least, get an arc-consistent sub-instance. We will show that a very minimalistic sort of reduction strategy suffices: instead of worrying about all possible issues with ensuring arc-consistency, we will only consider paths from variables \( y \) to \( x \) through the instance.

**Definition 20.1.** If \( X \) is an instance of a CSP and \( x, y \) are variables of \( X \), then a path \( p \) from \( x \) to \( y \) is defined as a sequence \( x = v_0, (R_1, i_1, j_1), v_1, ..., v_{n-1}, (R_n, i_n, j_n), v_n = y \) such that each \( v_i \) is a variable, and each \( R_k \) is a relation such that one of the constraints of the instance \( X \) imposes the relation \( R_k \) on a tuple \( u = (u_1, ...) \) of variables with \( u_{i_k} = v_{k-1} \) and \( u_{j_k} = v_k \).

To every path \( p \) from \( x \) to \( y \), we associate the binary relation \( P_p \leq A_x \times A_y \) which is given by

\[
P_p := \pi_{i_1 j_1}(R_1) \circ \cdots \circ \pi_{i_n j_n}(R_n).
\]

In other words, \( P_p \) is the set of pairs of values in \( A_x \times A_y \) which are consistent with the path \( p \).

We define addition and negation of paths in the natural way, so that if \( p \) is a path from \( x \) to \( y \) and \( q \) is a path from \( y \) to \( z \), then \( p + q \) is a path from \( x \) to \( z \) with \( P_{p+q} = P_p \circ P_q \), and \( -p \) is a path from \( y \) to \( x \) with \( P_{-p} = P_p^{-1} \).

In particular, we see that an instance is arc-consistent iff for all paths \( p \) the associated binary relations \( P_p \) are subdirect, and it is cycle-consistent iff we additionally have \( \Delta_{A_v} \subseteq P_{p} \) for every path \( p \) from a variable \( v \) back to itself.

**Definition 20.2.** Suppose that \( X \) is a cycle-consistent instance such that for all variable domains we have \( A_v = \text{Sg}(A_v^{\max}) \), that \( x \) is any variable with \( |A_x| > 1 \), that \( \theta_x \) is any maximal congruence on \( A_x \), and that \( A'_x \) is any congruence class of \( A_x/\theta_x \).

For each variable \( y \), we say that \( y \) is proper if there is a path \( p \) from \( y \) to \( x \) such that \( P_p/\theta_x \leq A_y \times A_x/\theta_x \) is the graph of a homomorphism \( \iota_y : A_y \rightarrow A_x/\theta_x \). In this case, we define the congruence \( \theta_y \in \text{Con}(A_y) \) to be the kernel of \( \iota_y \), and we define \( A'_y \) to be the preimage of \( A'_x \) under \( \iota_y \). If \( y \) is not proper, then we define \( A'_y \) to be \( A_y \).

We define the reduced instance \( X' \) by replacing the domain of each variable \( v \) by \( A'_v \), and replacing each constraint relation \( R \leq A_{v_1} \times \cdots \times A_{v_m} \) of \( X \) by \( R' = R \cap (A'_{v_1} \times \cdots \times A'_{v_m}) \).

The reason for the name “proper” is that a variable \( v \) is proper iff the reduced domain \( A'_v \) is a proper subalgebra of \( A_v \). First we need to check that the maps \( \iota_y \) for the proper variables \( y \) are well-defined.
**Lemma 20.3.** If $y$ is a proper variable and $p, q$ are two paths from $y$ to $x$ such that $P_p / \theta_x, P_q / \theta_x$ are graphs of homomorphisms $\iota_p, \iota_q : A_y \rightarrow A_x / \theta_x$, then in fact we have $\iota_p = \iota_q$. Thus $\iota_y, \theta_y$, and $A_y$ are all well-defined.

*Proof.* The path $p - q$ connects $y$ to itself, so by cycle-consistency we must have $\Delta_{A_y} \subseteq P_{p - q} = P_p \circ P_q$. Taking the quotient by $\theta_x$, we see that $\Delta_{A_y} \subseteq (P_p / \theta_x) \circ (P_q / \theta_x)^{-}$, so for every element $a \in A_y$ we must have $\iota_p(a) = \iota_q(a)$. □

We sometimes abuse notation, and think of $\iota_y$ as an isomorphism from $A_y / \theta_y$ to $A_x / \theta_x$.

**Lemma 20.4.** Suppose $p$ is a path from $y$ to a proper variable $z$. Then one of the following is true:

- $P_p / \theta_z = A_y \times A_z / \theta_z$, or
- $y$ is also proper, and $P_p / (\theta_y \times \theta_z)$ is the graph of an isomorphism $\iota_p : A_y / \theta_y \cong A_z / \theta_z$ such that $\iota_y = \iota_z \circ \iota_p$.

*Proof.* This follows from Corollary 18.4 and cycle-consistency (note that $A_z / \theta_z$ is simple, since it is isomorphic to $A_x / \theta_x$). □

We have the ingredients necessary to check that the reduced instance $X'$ is cycle-consistent. We start with arc-consistency.

**Lemma 20.5.** Suppose $R \leq_{sd} A_{v_1} \times \cdots \times A_{v_n}$ is a constraint of $X$. Then the reduced constraint $R' = R \cap (A'_{v_1} \times \cdots \times A'_{v_n})$ is subdirect inside $A'_{v_1} \times \cdots \times A'_{v_n}$, that is, $\pi_i(R') = A'_i$ for each $i$.

*Proof.* By symmetry, it’s enough to prove that $\pi_1(R') = A'_1$. In other words, for each element $a \in A'_1$, we want to find a tuple $s \in R$ such that $s_i \in A'_i$ for all $i$. We may ignore variables $v_i$ such that $i \neq 1$ and $v_i$ is not proper, since for such $i$ the restriction from $A_{v_i}$ to $A'_i = A_{v_i}$ has no effect. Similarly, for any two proper variables $v_i, v_j$ such that $\pi_j(R)$ induces an isomorphism between $A_{v_i} / \theta_{v_i}$ and $A_{v_j} / \theta_{v_j}$, we may ignore one of the two variables $v_i, v_j$, since any element $s \in R$ which satisfies $s_i \in A'_i$ will automatically also satisfy $s_j \in A'_j$.

To formalize the process of ignoring variables, we define an equivalence relation $\sim$ on the set of indices of proper variables of $R$, with $i \sim j$ when $\pi_{ij}(R)$ induces an isomorphism between $A_{v_i} / \theta_{v_i}$ and $A_{v_j} / \theta_{v_j}$ (that $\sim$ is an equivalence relation is easy to check). Then we let $I \subseteq [n]$ be a set of variable indices such that each $\sim$-class has exactly one representative in $I$, $1 \in I$, and no index of any non-proper variable other than possibly 1 is in $I$. We then define a relation $S \leq A_{v_1} \times \prod_{i \in I \setminus \{1\}} A_{v_i} / \theta_{v_i}$ by

$$S := \pi_I(R) / \prod_{i \in I \setminus \{1\}} \theta_{v_i}.$$

We just need to show that for every $a \in A'_{v_1}$ there is some $s \in S$ with $s_1 = a$ and $s_i = A'_{v_i} / \theta_{v_i}$ for each $i \in I \setminus \{1\}$. Note that by Lemma 20.4 and the construction of $I$, for every pair $i, j \in I$ the projection $\pi_{ij}(S)$ is full. Thus by Corollary 18.7, we in fact have

$$S^\text{max} = A_{v_1}^\text{max} \times \prod_{i \in I \setminus \{1\}} A_{v_i}^\text{max} / \theta_{v_i},$$

and since each $A_{v_i}$ is generated by $A_{v_i}^\text{max}$, we have

$$S = A_{v_1} \times \prod_{i \in I \setminus \{1\}} A_{v_i} / \theta_{v_i}.$$
Now we can check that cycle-consistency also holds for the reduced instance.

**Lemma 20.6.** Suppose $p$ is a path from $v$ to $v$ in the instance $X$, and let $p'$ be the corresponding path in $X'$. If $\Delta_{\hat{A}_{v'}} \subseteq \mathbb{P}_p$, then $\Delta_{\hat{A}'_{v'}} \subseteq \mathbb{P}_p'$.

**Proof.** Suppose that $p$ is the path $v = v_0, (R_1, i_1, j_1), v_1, \ldots, v_{n-1}, (R_n, i_n, j_n), v_n = v$. Note that in the corresponding path $p'$, we must replace each $R_i$ with $R'_i$, so we must also worry about the proper variables which occur in $R_i$ but do not lie along the path $p$. In order to do this cleanly, we consider the relation $R$ defined by

$$R := \left\{ (v_0, u_1^1, \ldots, u_n^1, v_n) \in A_v \times \prod_{i \leq n} R_i \times A_v \mid v_0 = u_1^1, u_1^{j_1} = u_2^{i_2}, \ldots, u_{n-1}^{j_{n-1}} = u_n^1, u_n^j = v_n \right\}.$$

If each $R_i$ has arity $m_i$, then $R$ is thought of as a relation of arity $m = 2 + \sum_i m_i$, and the indices of $R$ might contain several copies of variables of the instance $X$. Let the $i$th index of $R$ correspond to the variable $y_i$ in $X$, with $y_1 = v_0 = v$ and $y_m = v_n = v$, so

$$R \leq_{sd} A_{y_1} \times \cdots \times A_{y_m}.$$  

Note that by the arc-consistency of the instance $X$, for any two indices $i, j$ of the relation $R$, the projection $\pi_{ij}(R)$ is the same as $\mathbb{P}_q$ for some path $q$ from $y_i$ to $y_j$ formed out of the relations $R_i$, and that $\pi_{1m}(R) = \mathbb{P}_p$, so $\pi_{1m}(R) \supseteq \Delta_{\hat{A}_v}$.

As in the argument for arc-consistency, we define an equivalence relation $\sim$ on the proper indices of $R$ defined by $i \sim j$ when $\pi_{ij}(R)$ induces an isomorphism between $A_{y_i}/\theta_{y_i}$ and $A_{y_j}/\theta_{y_j}$. We let $I \subseteq [m]$ to be a set of indices of $R$ with $1, m \in I$, such that $I$ contains no indices of non-proper variables of $R$ other than possibly 1 and $m$, such that $I \setminus \{1\}$ contains one representative from each $\sim$ class of $\{1, \ldots, m - 1\}$, and such that $I \setminus \{1, m\}$ contains one representative from each $\sim$ class of $\{2, \ldots, m\}$. As before, we define a relation $S$ by

$$S := \pi_I(R)/\prod_{i \in I \setminus \{1, m\}} \theta_{y_i}.$$  

We just need to show that for every $a \in A'_{v'}$, there is some $s \in S$ with $s_1 = s_m = a$ and $s_i = A'_{y_i}/\theta_{y_i}$ for each $i \in I \setminus \{1, m\}$. By Lemma 20.4 and the construction of $I$, for every pair $i, j \in I$ with $\{i, j\} \neq \{1, m\}$ the projection $\pi_{ij}(S)$ is full. Thus by Corollary 18.7 we have

$$\pi_{I \setminus \{m\}}(S) = \prod_{i \in I \setminus \{1, m\}} A_{y_i}/\theta_{y_i}$$

and

$$\pi_{I \setminus \{1\}}(S) = A_{y_m} \times \prod_{i \in I \setminus \{1, m\}} A_{y_i}/\theta_{y_i}.$$  

Thus by Theorem 18.5 we have

$$S_{\text{max}} = \pi_{1m}(S)_{\text{max}} \times \prod_{i \in I \setminus \{1, m\}} A_{y_i}^{\text{max}}/\theta_{y_i},$$

149
and by Theorem 18.9 and the assumption $\pi_{1m}(S) = \pi_{1m}(R) \supseteq \Delta_{A_v}$, we have $\pi_{1m}(S)^{\text{max}} \supseteq \Delta_{A_v}^{\text{max}}$. Since each $A_y$ is generated by $A_y^{\text{max}}$, we have

$$S \supseteq \Delta_{A_v} \times \prod_{i \in I \setminus \{1, m\}} A_{y_i} / \theta_{y_i},$$

so in particular for every $a \in A'_v$ we have $\{a\} \times \prod_{i \in I \setminus \{1, m\}} A'_{y_i} / \theta_{y_i} \times \{a\} \subseteq S$, so $(a, a) \in \pi_{1m}(R') = \mathbb{P}'$.

Thus the reduced instance $X'$ is cycle-consistent. Since we can iteratively shrink our instance whenever some variable $x$ has $A_x \neq \text{Sg}(A_x^{\text{max}})$ or has $A_x = \text{Sg}(A_x^{\text{max}})$ but $|A_x| > 1$, we see that we eventually reach a situation where each $A_x$ consists of a single element, and then arc-consistency proves that this collection of single elements gives a solution to the original instance. We have proved our main result.

**Theorem 20.7.** If $X$ is a cycle-consistent instance of an ancestral CSP, then $X$ has a solution.

In fact, for any variable $x$ of $X$, and for any element $a \in A_x$ such that there is a sequence of subalgebras $A_x \supseteq A_0 \supseteq \cdots \supseteq A_n = \{a\}$ with $A_0 = \text{Sg}(A_x^{\text{max}})$ and such that for each $i$, there is a maximal congruence $\theta_i \in \text{Con}(A_i)$ and a congruence class $A'_i$ of $\theta_i$ with $A_{i+1} = \text{Sg}(A'_i)$, there is a solution to the instance $X$ in which $x$ is assigned the value $a$.

The simple construction of the reduced instance $X'$ can be used to show that we can find a solution to any cycle-consistent instance of an ancestral CSP in linear time.

## 21 Cycle-consistency solves majority CSPs

The paper which prompted the study of cycle-consistency was a preliminary investigation by Chen, Dalmau, and Grußien [40], which studied a slightly stronger consistency notion: singleton arc-consistency. Singleton arc-consistency refers to the strategy of fixing a particular value for some variable, and checking if applying arc-consistency to the remaining variables produces a contradiction. Singleton arc-consistency is clearly at least as powerful as cycle-consistency. One of the main results of [40] showed that singleton arc consistency solves majority CSPs, but in fact their proof strategy was to show that cycle-consistent instances of majority CSPs always have solutions.

The argument for majority algebras is simpler than the argument for ancestral algebras, essentially because the analogue of the case where all the variables domains are strongly connected doesn’t need to be considered. Instead, we are always in the situation where some variable domain $A_x$ has a proper absorbing subalgebra (every singleton is an absorbing subalgebra of a majority algebra), although we need to work slightly harder than we did in the absorbing case of ancestral CSPs since the absorption is no longer binary absorption. Rather than working with absorbing subalgebras, [40] used the closely related concept of an ideal of a majority algebra.

**Definition 21.1.** If $A = (A, m)$ is a majority algebra, then $B \subseteq A$ is called an ideal of $A$ if $m(B, A, B) \subseteq B$.

The word “ideal” comes from the theory of median algebras - a subset $B$ is an ideal of a median algebra $A$ iff there is a congruence $\theta$ of $A$ such that $B$ is a congruence class of $\theta$. The corresponding statement is not true of majority algebras in general: every subset of the dual discriminator algebra...
from Example 7.5 is an ideal, but the dual discriminator algebra on \( n \) elements is simple (and polynomially complete) for \( n \geq 3 \).

The next result shows that ideals interact with standard algebraic constructions (products, quotients, intersections) nicely. A similar result holds for absorbing subalgebras, with the same proof.

**Proposition 21.2.** Suppose that a relation \( \mathcal{R} \) is defined by a primitive positive formula \( \Phi \) involving the relations \( \mathcal{R}_1, \ldots, \mathcal{R}_k \). If we replace each \( \mathcal{R}_i \) with an ideal \( \mathcal{R}_i' \) of \( \mathcal{R}_i \) to make a primitive positive formula \( \Phi' \), then the relation \( \mathcal{R}' \) which is defined by \( \Phi' \) is an ideal of \( \mathcal{R} \).

**Proof.** Let \( \Phi(x) = \exists y \Psi(x, y) \), with \( \Psi \) quantifier-free, and let \( \Psi' \) be the corresponding formula with \( \mathcal{R}_i \)'s replaced by \( \mathcal{R}_i' \)'s. Then for any \( a, b, c \) with \( a, c \in \mathcal{R}' \) and \( b \in \mathcal{R} \), there exist \( d, e, f \) such that \( \Psi'(a, d), \Psi(b, e), \Psi'(c, f) \) hold, so \( \Psi'(m(a, b, c), m(d, e, f)) \) holds since each \( \mathcal{R}_i' \) is an ideal, so \( \Phi'(m(a, b, c)) \) holds. \( \Box \)

Recall the definition of a path in an instance (Definition 20.1). It’s notationally convenient to allow paths to act on subsets of the variable domains.

**Definition 21.3.** If \( p \) is a path connecting variables \( x, y \) of an instance \( \mathbf{X} \), and if \( B \) is a subset of the variable domain \( \mathcal{A}_x \), then we define \( B + p \) to be the subset of \( \mathcal{A}_y \) given by

\[
B + p := \{ c \in \mathcal{A}_y \mid \exists b \in B \text{ s.t. } (b, c) \in \mathcal{P}_p \} = \pi_2(\mathcal{P}_p \cap (B \times \mathcal{A}_y)).
\]

**Proposition 21.4.** If \( \mathcal{B} \leq \mathcal{A}_x \) and \( p \) is a path from \( x \) to \( y \), then \( \mathcal{B} + p \) is a subalgebra of \( \mathcal{A}_y \). If \( \mathcal{B} \) is an ideal of \( \mathcal{A}_x \) and the instance is arc-consistent, then \( \mathcal{B} + p \) is an ideal of \( \mathcal{A}_y \).

Our overall strategy will be to start with a cycle-consistent instance \( \mathbf{X} \), and find a collection of ideals \( \mathcal{A}_x' \) of the variable domains \( \mathcal{A}_x \) such that reducing each domain to \( \mathcal{A}_x' \) produces an arc-consistent instance \( \mathbf{X}' \). Then we will show that any such \( \mathbf{X}' \) is automatically cycle-consistent.

In order to find an arc-consistent family of ideal subdomains, we consider the set \( \mathcal{I} \) of pairs \((x, \mathcal{B})\) where \( x \) is a variable and \( \mathcal{B} \) is a proper ideal of \( \mathcal{A}_x \). Note that \( \mathcal{I} \) is nonempty as long as some \( x \) has \(|\mathcal{A}_x| > 1\), since every singleton is an ideal.

**Definition 21.5.** Let \( \mathcal{I} \) be the set of pairs \((x, \mathcal{B})\) where \( x \) is a variable and \( \mathcal{B} \) is a proper ideal of \( \mathcal{A}_x \). We define a quasiorder \( \preceq \) on \( \mathcal{I} \) by \((x, \mathcal{B}) \preceq (y, \mathcal{B} + p)\) for every path \( p \) from \( x \) to \( y \) with \( \mathcal{B} + p \neq \mathcal{A}_y \).

**Proposition 21.6.** If \( \mathbf{X} \) is a cycle-consistent instance, \( x \) is a variable, and \((x, \mathcal{B}) \preceq (x, \mathcal{C})\), then \( \mathcal{B} \leq \mathcal{C} \).

**Proof.** Suppose \( p \) is a path from \( x \) to itself with \( \mathcal{B} + p = \mathcal{C} \). By cycle-consistency we must have \( \Delta_{\mathcal{A}_x} \subseteq \mathcal{P}_p \), so \( \mathcal{B} \subseteq \mathcal{B} + p \). \( \Box \)

**Definition 21.7.** Suppose \( \mathbf{X} \) is a cycle-consistent instance of a majority CSP, and assume without loss of generality that each constraint of \( \mathbf{X} \) is binary. Fix a maximal element \((x, \mathcal{A}_x')\) of \( \mathcal{I} \) under the quasiorder \( \preceq \).

Call a variable \( y \) proper if there is a path \( p \) from \( x \) to \( y \) such that \( \mathcal{A}_x' + p \neq \mathcal{A}_y \), and in this case set \( \mathcal{A}_y' = \mathcal{A}_x' + p \). If \( y \) is not proper, then set \( \mathcal{A}_y' = \mathcal{A}_y \).

Define the reduced instance \( \mathbf{X}' \) by replacing the domain of each variable \( v \) by \( \mathcal{A}_v' \), and by replacing each constraint \( \mathcal{R} \subseteq \mathcal{A}_u \times \mathcal{A}_v \) with \( \mathcal{R}' = \mathcal{R} \cap (\mathcal{A}_u' \times \mathcal{A}_v') \).
First we need to check that the sets $A'_y$ are well-defined.

**Lemma 21.8.** If there are paths $p, q$ from $x$ to $y$ such that $A'_x + p \neq A_y$ and $A'_x + q \neq A_y$, then $A'_x + p = A'_x + q$.

**Proof.** Since $(x, A'_x)$ is maximal and $(x, A'_x) \preceq (y, A'_x + p)$, we must have $(y, A'_x + p) \preceq (x, A'_x) \preceq (y, A'_x + q)$, so $A'_x + p \preceq A'_y + q$. Similarly we have $A'_x + q \preceq A'_y + p$, so $A'_x + p = A'_x + q$. \qed

Next we check arc-consistency.

**Lemma 21.9.** If $p$ is a path from $y$ to $z$ and $p'$ is the corresponding path in $X'$, then $A'_y + p' = A'_z$.

**Proof.** We just need to check this in the case when $p$ has length 1, corresponding to a binary relation $R \preceq sd A_y \times A_z$. If $A'_y + p \neq A'_z$, then $y, z$ must both be proper with $A'_y + p = A'_z$. Either way we see that $A'_y + p \supseteq A'_z$, and since $R' = R \cap (A'_y \times A'_z)$ we have $A'_y + p' = A'_z$ in the reduced instance. \qed

Finally, we check that arc-consistency of $X'$ and cycle-consistency of $X$ implies cycle-consistency of $X'$. For this, we note that if $p$ is a path from $v$ back to itself in $X$, and if $p'$ is the corresponding path in $X'$, then $P_{p'}$ is an ideal of $P_p$. Since $P_p \supseteq \Delta_{A_x}$, we have

$$m(P_{p'}, \Delta_{A'_x}, P_{p'}) \subseteq P_{p'},$$

so the cycle-consistency of $X'$ follows from the following result.

**Theorem 21.10.** Suppose that $R \preceq sd A \times A$ is subdirect with $m(R, \Delta_A, R) \subseteq R$, where $m$ is a majority operation. Then $\Delta_A \subseteq R$.

**Proof.** Let $a$ be any element of $A$. Since $R$ is subdirect, there are $b, c \in A$ such that $(a, b) \in R$ and $(c, a) \in R$. Then since $(a, a) \in \Delta_A$, we have

$$\begin{bmatrix} a \\ a \\ a \\ c \end{bmatrix} = m\left( \begin{bmatrix} a \\ b \\ a \\ c \end{bmatrix} \right) \in m(R, \Delta_A, R) \subseteq R. \quad \Box$$

**Corollary 21.11.** The reduced instance $X'$ is cycle-consistent.

We have proved the main result of this section.

**Theorem 21.12.** Every cycle-consistent instance $X$ of a majority CSP has a solution.

In fact, for any variable $v$ of $X$ and any value $a \in A_v$, the instance $X$ has a solution in which the variable $v$ is assigned the value $a$.

**Proof.** For the second statement, we note that if $|A_v| > 1$, then $(v, \{a\}) \in I$, so there is some maximal element $(x, A'_x) \in I$ such that $(v, \{a\}) \preceq (x, A'_x)$, and we define the reduction $X'$ in terms of the maximal element $(x, A'_x)$. If $v$ is proper, then $(v, \{a\}) \preceq (x, A'_x) \preceq (v, A'_v)$ we must have $a \in A_v$, and if $v$ is not proper then we have $a \in A_v = A'_v$. Either way, we see by induction that the reduced instance $X'$ has a solution in which the variable $v$ is assigned the value $a$. \qed

**Corollary 21.13.** Suppose $A$ is an algebra with a partial semilattice term $s$ and a ternary term $q$ such that for any subalgebra $B \leq A$, the restriction of $q$ to $Sg(B_{\text{max}})$ is a majority operation. Then every cycle-consistent instance of CSP($A$) has a solution.
Proof. By Corollary 18.10 if we start with a cycle-consistent instance $X$ and restrict all the variable domains $A_i$ to $\text{Sg}(A_i^{\text{max}})$ to create a new instance $X'$, then $X'$ will still be cycle-consistent, and by assumption $X'$ will be preserved by the majority operation $g$. Then by the previous theorem, $X'$ will have a solution.

Example 21.1. Consider $A = (-, 0, +), g$, where $g$ is the idempotent cyclic ternary operation with

\[
\begin{align*}
g(0, 0, -) &= g(0, -, -) = -, \\
g(0, -, +) &= g(-, -, +) = -, \\
g(0, 0, +) &= g(0, +, +) = +, \\
g(0, +, -) &= g(-, +, +) = +.
\end{align*}
\]

This can be described more succinctly as follows: the permutation $(- +)$ is an automorphism of $A$, $\{-, +\}$ is a majority subalgebra of $A$, and $\{0, -\}, \{0, +\}$ are semilattice subalgebras of $A$ with $0 \rightarrow -, +$. The term $s(x, y) := g(x, x, y)$ is a partial semilattice, and $s, g$ satisfy the assumptions of the Corollary above, so every cycle-consistent instance of $\text{CSP}(A)$ has a solution. We give a table for $s$ and draw the graph of two element subalgebras of $A$ (with undirected edges for majority subalgebras and directed edges for semilattice subalgebras) below.

<table>
<thead>
<tr>
<th>s</th>
<th>-</th>
<th>0</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>-</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

The relational clone $\text{Inv}(g)$ is generated by the unary relation $x \neq 0$, the binary relation $x = -y$, the binary relation $x \leq y$, and the ternary relation $x = 0 \implies y = z$.

The clone $\langle g \rangle$ is properly contained in the clone $\langle s_2 \rangle$ from Example 7.8 and it does not contain any proper subclone with a Taylor operation. In some sense the algebra considered in this example is the prototypical example of a bounded width algebra: Bulatov [35] has shown that in every minimal bounded width clone, the maximal strongly connected components behave as if there is a majority operation preserving them, and for every pair of maximal strongly connected components there is a two-element majority subalgebra which connects them.

Remark 21.1. It’s tempting to try to generalize Theorem 21.10 to near-unanimity operations. We say that a subalgebra $B$ absorbs $A$ with respect to a near-unanimity operation $t$ if

$$t(B, ..., B, A, B, ..., B) \subseteq B$$

for each possible location of $A$. Suppose that $R \leq_{sd} A \times A$ absorbs $\Delta_A$ with respect to $t$ - can we conclude that $R$ contains the diagonal?

Unfortunately the answer is no: even if $R$ is subdirect and absorbs $A^2$ with respect to a near-unanimity term, we might not have $\Delta_A \subseteq R$. Consider the threshold function $t_2^R$ from Example 2.3 defined by

$$t_2^R(x_1, ..., x_n) = \begin{cases} 1 & \sum_i x_i \geq 2, \\ 0 & \sum_i x_i \leq 1. \end{cases}$$

For $n \geq 4$, the relation

$$R = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}$$

153
absorbs \{0, 1\}^2 with respect to \(e_2^3\), but does not contain the diagonal element \((0, 0)\). However, \(R\) does intersect the diagonal at \((1, 1)\). In the next section we will see that this weaker claim generalizes: if \(R \leq_{sd} A \times A\) absorbs \(\Delta_A\), then \(R \cap \Delta_A \neq \emptyset\).

## 22 Absorption, Jónsson absorption, and connectivity

Absorption is a common generalization of ideals of majority algebras and maximal strongly connected components of minimal ancestral algebras, and a lot of the theory of absorbing subalgebras applies to general (finite, idempotent) algebras, without assuming the existence of a Taylor term. After introducing absorption, we will show that absorbing subalgebras \(R'\) of binary relations \(R\) retain some of the connectivity properties of the original relations \(R\).

**Definition 22.1.** A subalgebra \(B \leq A\) absorbs \(A\) with respect to an idempotent term \(t\) if

\[
t(B, ..., B, A, B, ..., B) \subseteq B
\]

for each possible location of \(A\). We just say that \(B\) absorbs \(A\), written \(B \triangleleft A\), if there exists some idempotent term \(t\) such that \(B\) absorbs \(A\) with respect to \(t\).

More generally, we sometimes say that a set \(B\) absorbs a set \(A\) with respect to an idempotent term \(t\) if

\[
t(B, ..., B, A, B, ..., B) \subseteq B
\]

for each possible location of \(A\). Note that if \(B \subseteq A\), then \(B\) must be closed under \(t\).

The reason we avoid specifying the idempotent term \(t\) in the notation \(B \triangleleft A\) is that there exists a common term \(t\) which witnesses all absorption within any finite collection of pairs \(B_i \triangleleft A_i\).

**Proposition 22.2.** If \(B_1 \triangleleft A_1\) with respect to \(t_1\) and \(B_2 \triangleleft A_2\) with respect to \(t_2\), then each \(B_i \triangleleft A_i\) with respect to the star composition \(t_1 \ast t_2\) (see Definition 6.3). If \(A_1 = B_2\), then \(B_1 \triangleleft A_2\) with respect to \(t_1 \ast t_2\).

**Corollary 22.3.** A finite algebra \(A\) has a near-unanimity term iff for all \(a \in A\), the singleton \(\{a\}\) absorbs \(A\).

A common strategy in arguments involving absorbing operations \(t\) of high arity \(n\) is to consider expressions of the form

\[
t(x, ..., x, y, z, ..., z),
\]

where just a single \(y\) occurs, and iteratively march the location of the \(y\) one step to the left at a time. We can make such arguments more transparent by phrasing them in terms of the sequence of ternary terms

\[
d_i(x, y, z) := t(x, ..., x, y, z, ..., z)_{n-1 \rightarrow i-1}^{n-1 \rightarrow 0},
\]

with \(d_0(x, y, z) := x\) and \(d_{n+1}(x, y, z) := z\), so that the \(d_i\) satisfy the system of identities

\[
\begin{align*}
d_0(x, y, z) &\approx x, \\
d_i(x, y, z) &\approx d_{i+1}(x, x, y), \\
d_{n+1}(x, y, z) &\approx z.
\end{align*}
\]

154
If \( B \) absorbs \( A \) with respect to the term \( t \), then we will additionally have

\[
d_i(B, A, B) \subseteq B
\]

for all \( i \).

**Definition 22.4.** A Jónsson absorption chain is a sequence of ternary terms \( d_1, \ldots, d_n \) which satisfy the identities

\[
d_1(x, x, y) \approx x, \\
d_i(x, y, y) \approx d_{i+1}(x, x, y), \\
d_n(x, y, y) \approx y.
\]

We say that \( B \) Jónsson absorbs \( A \) with respect to the Jónsson chain \( d_1, \ldots, d_n \) if for each \( i \in [n] \) we have

\[
d_i(B, A, B) \subseteq B.
\]

If \( B \) Jónsson absorbs \( A \) with respect to some Jónsson chain, then we write \( B \triangleleft_J A \).

**Proposition 22.5.** If \( B \triangleleft A \), then \( B \triangleleft_J A \).

As with absorption, we can witness several instances of Jónsson absorption simultaneously with a single Jónsson absorption chain \( d_1, \ldots, d_n \).

**Proposition 22.6.** If \( B_1 \triangleleft_J A_1 \) with respect to \( d_1, \ldots, d_m \) and \( B_2 \triangleleft_J A_2 \) with respect to \( e_1, \ldots, e_n \), then the sequence of terms \( f_1, \ldots, f_{mn} \) defined by

\[
f_{ni+1+j} := d_i(x, e_j(x, y, z), z)
\]

is a Jónsson absorption chain which witnesses both \( B_1 \triangleleft_J A_1 \) and \( B_2 \triangleleft_J A_2 \). If \( A_1 = A_2 \), then \( B_1 \triangleleft_J A_2 \) with respect to \( f_1, \ldots, f_{mn} \).

**Corollary 22.7.** A finite algebra \( A \) generates a congruence distributive variety iff for all \( a \in A \), the singleton \( \{a\} \) Jónsson absorbs \( A \).

**Proof.** A Jónsson absorbing chain which witnesses \( \{a\} \triangleleft_J A \) for all \( a \in A \) is the same as a sequence of terms \( d_1, \ldots, d_m \) which satisfy the system of identities

\[
d_1(x, x, y) \approx x, \\
d_i(x, y, x) \approx x \text{ for all } i, \\
d_i(x, y, y) \approx d_{i+1}(x, x, y) \text{ for all } i, \\
d_m(x, y, y) \approx y,
\]

that is, \( d_1, \ldots, d_m \) are a sequence of directed Jónsson terms. By Theorem [A.50], a variety is congruence distributive iff it has directed Jónsson terms. \( \square \)

**Example 22.1.** Suppose \( B \) is an ideal of a majority algebra \( A = (A, m) \). Then \( B \triangleleft_J A \) with respect to the Jónsson absorption chain \( d_1(x, y, z) = m(x, y, z) \) (of length 1):

\[
m(x, x, y) \approx x, \\
m(x, y, y) \approx y, \\
m(B, A, B) \subseteq B.
\]

155
In fact, the converse holds: if $B \triangleleft_J A$, then there must be a majority term $m' \in \text{Clo}(m)$ such that $m'(B, A, B) \subseteq B$. This follows from the fact that every ternary term in a majority algebra is either a projection or another majority operation. If $A$ is locally finite, then by applying the construction of Proposition 22.6 iteratively to all the majority operations in $\text{Clo}(m)$, we can find a single majority term $\hat{m} \in \text{Clo}(m)$ such that for any $C \leq B \in HSP(A)$ we have

$$C \triangleleft_J B \iff \hat{m}(C, B, C) \subseteq C.$$

As we will see later, $B \triangleleft_J A$ together with the fact that finite majority algebras are finitely related implies that $B \triangleleft A$ - possibly with respect to a term of very high arity (for instance, in the case where $A$ is the dual discriminator algebra from Example 7.5 and $|B| = |A| - 1$, the minimal arity of a term $t$ which witnesses $B \triangleleft A$ is $|A| + 1$). So ideals of finite majority algebras are actually the same thing as absorbing subalgebras!

Like ideals of majority algebras, absorbing subalgebras play nice with primitive positive formulas.

**Proposition 22.8.** Suppose that a relation $R$ is defined by a primitive positive formula $\Phi$ involving the relations $R_1, \ldots, R_k$. If we replace each $R_i$ with an absorbing subalgebra $R'_i \triangleleft R_i$ to make a primitive positive formula $\Phi'$, then the relation $R'$ which is defined by $\Phi'$ is an absorbing subalgebra of $R$. The same is true with “absorbing” replaced by “Jónsson absorbing”.

**Proof.** Since only finitely many relations $R_i$ show up in $\Phi$, we can find a single absorbing term (or Jónsson chain) which witnesses all absorptions $R'_i \triangleleft R_i$ (or $R'_i \triangleleft_J R_i$) simultaneously. From here the proof is similar to the proof in the case of ideals of majority algebras.

Now we will illustrate how Jónsson absorption is used, by proving a few connectivity results. Recall that every binary relation $R \leq A \times A$ can be visualized as a graph in two different ways: we can either think of $R$ as a bipartite graph on the disjoint union $A \sqcup A$, or we can think of $R$ as a directed graph on $A$. The next result is perhaps the most crucial.

**Theorem 22.9 (Absorbing directed paths [13]).** If $S, R \leq A \times A$ are binary relations with $S \triangleleft_J R$, and $a, b \in A$ satisfy

- $(a, a), (b, b) \in S$, and
- $(a, b) \in R$,

then if we think of $S$ as a directed graph on $A$, there is a directed path from $a$ to $b$ in $S$, that is, $(a, b) \in S^\circ n$ for some $n$.

**Proof.** Suppose $S \triangleleft_J R$ with respect to the Jónsson chain $d_1, \ldots, d_n$. Then for each $i$ we have

$$\left[\begin{array}{c}
  d_i(a, a, b) \\
  d_{i+1}(a, a, b)
\end{array}\right] = \left[\begin{array}{c}
  d_i(a, a, b) \\
  d_i(a, b, b)
\end{array}\right] = d_i\left(\left[\begin{array}{c}
  a \\
  a
\end{array}\right], \left[\begin{array}{c}
  a \\
  b
\end{array}\right], \left[\begin{array}{c}
  b \\
  b
\end{array}\right]\right) \in d_i(S, R, S) \subseteq S.$$

Stringing these together, we get a directed path from $d_1(a, a, b) = a$ to $d_n(a, b, b) = b$ of length $n$, so in fact $(a, b) \in S^\circ n$.

Applying the above to $S^\circ m \triangleleft_J R^\circ m$ for a sufficiently large $m$, we get the following stronger-looking corollary.
Corollary 22.10. If $\mathcal{S}, \mathcal{R} \leq \mathbb{A} \times \mathbb{A}$ have $\mathcal{S} \triangleleft J \mathcal{R}$, and $a, b \in \mathbb{A}$ satisfy
\begin{itemize}
  \item each of $a, b$ is contained in a directed cycle of the digraph $\mathcal{S}$, and
  \item there is a directed path from $a$ to $b$ in the digraph $\mathcal{R}$,
\end{itemize}
then there is a directed path from $a$ to $b$ in the digraph $\mathcal{S}$.

For the sake of applying the previous result, it is useful to keep in mind the following basic fact about finite directed graphs.

Proposition 22.11. If $(A, R)$ is a finite directed graph such that each vertex of $A$ has in-degree at least 1 (in other words, such that $\pi_2(R) = A$), then for every vertex $a \in A$ there is some $a' \in A$ and some $n$ such that $a'$ is contained in a directed cycle of length $n$ and such that there is a directed path from $a'$ to $a$ of length $n$ (that is, $(a', a'), (a', a) \in R^n$).

Proof. Define a function $\varphi : A \to A$ such that for each $a \in A$ we have $(\varphi(a), a) \in R$. Then there is some $n$ such that $\varphi^{2n} = \varphi^n$ by the finiteness of $A$: in fact, we may take $n = \text{lcm}\{1, \ldots, |A|\}$.

In the next result, we think of binary relations as bipartite graphs. Recall that the linked components of a binary relation $R \leq \mathbb{A} \times \mathbb{B}$ are the connected components of $R$ considered as a bipartite graph on $\mathbb{A} \sqcup \mathbb{B}$, and that the linked components of size greater than 1 are the same as the congruence classes of the linking congruence $\ker \pi_1 \vee \ker \pi_2$ on $R$.

Theorem 22.12 (Absorbing linked components [13]). If $\mathcal{S}, \mathcal{R} \leq \mathbb{A} \times \mathbb{B}$ are binary relations with $\mathcal{S} \triangleleft J \mathcal{R}$, and $a, b \in \pi_1(\mathcal{S})$ are in the same linked component of $\mathcal{R}$, then $a, b$ are in the same linked component of $\mathcal{S}$.

Proof. If $a, b$ are linked in $\mathcal{R}$, then there is some $m$ such that $(a, b) \in (R \circ R^-)^m$. Since $(S \circ S^-)^{om} \triangleleft J (R \circ R^-)^{om}$ and $(a, a), (b, b) \in S \circ S^-$ by $a, b \in \pi_1(S)$, we can apply Theorem 22.9 to see that there is some $n$ such that $(a, b) \in (S \circ S^-)^{omn}$. Thus $a, b$ are in the same linked component of $\mathcal{S}$.

The next result is an analogue of Theorem [18.9] and Theorem 21.10 for Jónsson absorption.

Theorem 22.13 (Loop Lemma, finite absorbing case [11]). If $\mathcal{R} \leq_{sd} \mathbb{A} \times \mathbb{A}$ is subdirect, $\mathbb{A}$ is finite, and $\mathcal{R}$ Jónsson absorbs the diagonal $\Delta_{\mathbb{A}}$, then $\mathcal{R} \cap \Delta_{\mathbb{A}} \neq \emptyset$.

Proof. We may assume without loss of generality that $\mathbb{A}$ is idempotent. As long as $|\mathbb{A}| > 1$, we will try to find a proper subalgebra $\mathbb{B} \leq \mathbb{A}$ with $\mathcal{R} \cap (\mathbb{B} \times \mathbb{B})$ subdirect. Then $\mathcal{R} \cap (\mathbb{B} \times \mathbb{B})$ will Jónsson absorb $\Delta_{\mathbb{B}}$, and we can show by induction that $\mathcal{R} \cap \Delta_{\mathbb{B}} \neq \emptyset$.

Let $b$ be any element of $\mathbb{A}$, and define a sequence of subalgebras $\mathbb{B}_i$ by $\mathbb{B}_0 = \{b\}, \mathbb{B}_{i+1} = \mathbb{B}_i + \mathcal{R}$, i.e. $\mathbb{B}_{i+1} = \pi_2(\mathcal{R} \cap (\mathbb{B}_i \times \mathbb{A}))$. If there is any $i$ such that $\mathbb{B}_i \neq \mathbb{A}$ but $\mathbb{B}_{i+1} = \mathbb{A}$, then for every $C \leq \mathbb{A}$ we have $(C + R^-) \cap \mathbb{B}_i \neq \emptyset$, so by the finiteness of $\mathbb{B}$ we may take
\[ \mathbb{B} = \bigcap_{k \geq 0} \mathbb{B}_i + (R^-)^{\circ k} = \{a \mid \exists a_0, a_1, \ldots \in \mathbb{B}_i \text{ s.t. } a_0 = a \text{ and } \forall j (a_j, a_{j+1}) \in \mathcal{R} \}. \]

Otherwise, each $\mathbb{B}_i \neq \mathbb{A}$, and by the finiteness of $\mathbb{A}$ there must be some $m, n$ such that $\mathbb{B}_{m+i} = \mathbb{B}_m$ for each $i$, so we may take $\mathbb{B} = \mathbb{B}_m$.

157
Consider any directed cycle \(a_0, ..., a_{kn} = a_0\) of \(\mathbb{R}\) (considered as a digraph) with \(a_0 \in \mathbb{B}_m\). We will show that each \(a_i \in \mathbb{B}_m\). Note that \((a_0, a_0), (a_i, a_i) \in \mathbb{R}^{\text{kn}}\), that \(\mathbb{R}^{\text{kn}}\) Jónsson absorbs \((\mathbb{R} \cup \Delta_A)^{\text{kn}}\), and that \((a_0, a_i) \in \mathbb{R}^{\text{oi}} \subseteq (\mathbb{R} \cup \Delta_A)^{\text{kn}}\). Thus by Theorem 22.9 there is some \(l\) such that \((a_0, a_i) \in \mathbb{R}^{\text{ln}}\), and since \(\mathbb{B}_m + \mathbb{R}^{\text{ln}} = \mathbb{B}_{m+l} = \mathbb{B}_m\), we see that \(a_i \in \mathbb{B}_m\).

Since \(\mathbb{B}_{m+i} + \mathbb{R}^{\text{on}} = \mathbb{B}_{m+i}\) and \(\mathbb{B}_{m+i}\) is finite, for each element \(a\) of \(\mathbb{B}_{m+i}\) there is an \(a_i\) contained in a directed cycle of \(\mathbb{R}^{\text{on}}\) and a directed path of \(\mathbb{R}^{\text{on}}\) from \(a_i\) to \(a\), so in fact we have \(a \in \mathbb{B}_m\) as well, and we see that \(\mathbb{B}_{m+i} \subseteq \mathbb{B}_m\). Similarly we have \(\mathbb{B}_m \subseteq \mathbb{B}_{m+i}\), so \(\mathbb{B}_m = \mathbb{B}_{m+i}\).

**Corollary 22.14.** If \(\mathbb{R} \leq_{\text{sd}} A \times A\) is subdirect, \(A\) is finite and has no proper absorbing subalgebra, and \(\mathbb{R}\) absorbs the diagonal \(\Delta_A\), then \(\Delta_A \subseteq \mathbb{R}\).

**Proof.** Since \(\mathbb{R} \cap \Delta_A \neq \emptyset\) is an absorbing subalgebra of \(\Delta_A\) and \(\Delta_A \cong A\) has no proper absorbing subalgebra, we must have \(\mathbb{R} \cap \Delta_A = \Delta_A\).

**Definition 22.15.** We say that \(B\) is a minimal absorbing subalgebra of \(A\), written \(B \ll A\), if \(B \vartriangleleft A\) and \(B\) has no proper absorbing subalgebra.

**Proposition 22.16.** Every finite absorbing subalgebra of \(A\) contains a minimal absorbing subalgebra of \(A\), and any pair of distinct minimal absorbing subalgebras of \(A\) are disjoint.

**Proof.** This follows from the fact that \(C \vartriangleleft B \vartriangleleft A\) implies \(C \vartriangleleft A\), and the fact that the intersection of any pair of absorbing subalgebras is an absorbing subalgebra.

**Theorem 22.17.** Suppose that \(X\) is an arc-consistent instance of a CSP, and suppose that for each variable domain \(A_v\) there is a minimal absorbing subalgebra \(A'\vartriangleleft A_v\) such that the reduced instance \(X'\) with variable domains replaced by \(A'\) and relations \(R \leq_{\text{sd}} A_{v_1} \times \cdots \times A_{v_n}\) replaced by \(R' = R \cap (A'_{v_1} \times \cdots \times A'_{v_n})\) is arc-consistent.

Then for any path \(p\) in \(X\) from a variable \(v\) to itself such that \(P_p \supseteq \Delta_{A_v}\), the corresponding path \(p'\) of \(X'\) has \(P_{p'} \supseteq \Delta_{A'_v}\). In particular, if \(X\) is cycle-consistent then so is \(X'\).

**Proof.** Note that \(P_{p'}\) is an absorbing subalgebra of \(P_p\), so \(P_{p'}\) absorbs \(\Delta_{A'_v}\). Since \(P_{p'}\) is subdirect in \(A'_{v_1} \times A'_{v_2}\) by the arc-consistency of \(X'\) and \(A'_{v_2}\) has no proper absorbing subalgebra, we may apply Corollary 22.14 to see that \(P_{p'} \supseteq \Delta_{A'_v}\).

Later we will show that any cycle-consistent instance \(X\) has an arc-consistent reduction \(X'\) where all variable domains are replaced by minimal absorbing subalgebras, which will set us up to apply Theorem 22.17. The argument strategy will be fairly generic, not using any specific properties of absorbing subalgebras other than Theorem 22.9 and the fact that absorption is compatible with primitive positive formulas. Additionally, we will be able to weaken cycle-consistency to a property known as \(pq\)-consistency, which says that for any pair of paths \(p, q\) from a variable \(v\) to itself, there is some \(j \geq 0\) such that \(P_{p+q+j} \supseteq \Delta_{A_v}\).

**References**


[85] Paolo Lipparini. Difference terms and commutators.


163


A  Commutator theory in congruence modular varieties

Before diving into commutator theory, we’ll review of some of the theory of modular lattices. The theory really begins with the observation that in any module, the lattice of submodules is always ranked (so long as there are no infinite chains of submodules). In fact, not only is this lattice ranked, but also every (finite) sublattice of the lattice of submodules is ranked as well. So it is natural to study lattices which have this property.

Definition A.1. The length of a finite chain is the number of elements in the chain minus 1. The length of a poset is the supremum of the lengths of all of its chains.

Definition A.2. A poset satisfies the Jordan-Dedekind chain condition if for any \( a \leq b \), any two maximal chains from \( a \) to \( b \) have equal length.


The simplest situation to consider is the situation where some element \(a\) has two distinct covers \(b, c\). Then \(a = b \land c\), and we may start by considering sublattices of the interval \([a, b \lor c]\). The claim is that in this scenario, we need \(b \lor c\) to cover both \(b\) and \(c\), i.e. the sublattice generated by \(b\) and \(c\) must have length two. If \(b \lor c\) does not cover \(c\), say \(c < d < b \lor c\) for some \(d\), then we have a problem: the sublattice generated by \(b, c, d\) is a copy of the pentagon lattice \(N_5\), which is not ranked. The only hard part of verifying this is checking that \(b \land d = a\), but this follows from \(a \leq b \land d \leq b\) and \(b \not\leq d\).

**Definition A.3.** A poset is called *upper semimodular* if whenever an element \(a\) has two distinct covers \(b, c\), there is some element \(d\) which covers both \(b\) and \(c\).

Surprisingly, it turns out that any upper semimodular poset which has no infinite chains satisfies the Jordan-Dedekind chain condition. Note that every chain is contained in a maximal chain (by Zorn’s Lemma).

**Proposition A.4.** If \(a\) is any element of an upper semimodular poset which has no infinite chains, then any two maximal chains starting at \(a\) (going upwards) have the same length.

**Proof.** Let \(a < a_1 < \cdots\) and \(a < a'_1 < \cdots\) be two maximal chains starting from \(a\) of lengths \(m, n\), and induct on \(\min(m, n)\). We may assume without loss of generality that \(m \leq n\). By upper semimodularity, there is some element \(a''_2\) which covers both \(a_1\) and \(a'_1\). Pick some maximal chain \(a''_2 < a''_3 < \cdots\) starting from \(a''_2\). Then the maximal chains \(a_1 < a_2 < \cdots\) and \(a_1 < a''_2 < \cdots\) must both have length \(m - 1\) by the induction hypothesis. Since the maximal chain \(a'_1 < a''_2 < \cdots\) then also has length \(m - 1\), we can apply the induction hypothesis to see that the maximal chain \(a'_1 < a''_2 < \cdots\) has length \(m - 1\) as well, so \(m = n\). \(\square\)

**Corollary A.5 (Birkhoff [22]).** An upper semimodular poset which has no infinite chains satisfies the Jordan-Dedekind chain condition.

**Proof.** If \(a \leq b\), then we can pick some fixed maximal chain \(b < b_1 < \cdots\) starting from \(b\). By appending it to any two maximal chains from \(a\) to \(b\) of different lengths, we obtain two maximal chains starting from \(a\) which have different lengths, contradicting the previous proposition. \(\square\)

On any poset of finite length which satisfies the Jordan-Dedekind chain condition and has upper or lower bounds, we can define a *height function* \(h\) such that whenever \(a\) is covered by \(b\), we have \(h(b) = h(a) + 1\).

**Proposition A.6 (Birkhoff [22]).** A ranked lattice of finite length is upper semimodular if and only if its height function satisfies the inequality

\[
h(x) + h(y) \geq h(x \lor y) + h(x \land y).
\]

**Proof.** The inequality clearly implies upper semimodularity. Now suppose our lattice is upper semimodular, and pick maximal chains

\[
x \land y = x_0 < x_1 < \cdots < x_m = x,
\]
\[
x \land y = y_0 < y_1 < \cdots < y_n = y.
\]
We claim that for each \( i, j \), \( x_i \lor y_j \) is either covered by or equal to \( x_{i+1} \lor y_j \) and \( x_i \lor y_{j+1} \). We can prove this by induction on \( i, j \): if it’s true for \( i, j \), then by upper semimodularity \( x_{i+1} \lor y_{j+1} \) will either cover or be equal to both of \( x_{i+1} \lor y_j \) and \( x_i \lor y_{j+1} \).

Thus, the sequence

\[
x = x \lor y_0 \leq x \lor y_1 \leq \cdots \leq x \lor y_n = x \lor y
\]

has every adjacent pair either equal or a cover, so

\[
h(x \lor y) - h(x) \leq h(y) - h(x \land y).
\]

There is also a corresponding notion of lower semimodularity, and a dual version of the above result. Putting them together, we get the following.

**Theorem A.7** (Birkhoff [22]). A lattice of finite length is modular iff it satisfies the Jordan-Dedekind chain condition and its height function satisfies

\[
h(x) + h(y) = h(x \lor y) + h(x \land y).
\]

**Proof.** Since modular implies both upper and lower semimodular, it implies the chain condition and the condition on the height function. For the other direction, suppose that we have a ranked lattice whose height function satisfies the given condition.

Suppose for contradiction that there is a sublattice isomorphic to the pentagon \( N_5 \) (recall from the discussion around Definition 8.7 that a lattice is modular iff it doesn’t have \( N_5 \) as a sublattice). Suppose this sublattice is generated by \( a, b, c \), with \( b < c \) and \( a \land b = a \land c \), \( a \lor b = a \lor c \). Then we have

\[
h(a) + h(b) = h(a \lor b) + h(a \land b) = h(a \lor c) + h(a \land c) = h(a) + h(c),
\]

so \( h(b) = h(c) \), contradicting \( b < c \).

The next result can be viewed as a strengthening of the fact that a modular lattice is both upper and lower semimodular.

**Theorem A.8** (Diamond Isomorphism Theorem). If \( a, b \) are elements of a modular lattice, then the maps \( \phi : [a, a \lor b] \rightarrow [a \land b, b] \) and \( \varphi : [a \land b, b] \rightarrow [a, a \lor b] \) given by

\[
\phi : x \mapsto x \land b \quad \text{and} \quad \varphi : y \mapsto y \lor a
\]

are lattice isomorphisms.

**Proof.** First we check that \( \phi, \varphi \) are inverse to each other. By the modular law, for \( x \in [a, a \lor b] \) we have

\[
\varphi(\phi(x)) = (x \land b) \lor a = x \land (b \lor a) = x,
\]

and for \( y \in [a \land b, b] \) we have

\[
\phi(\varphi(y)) = (y \lor a) \land b = y \lor (a \land b) = y.
\]

It is clear that \( \phi \) respects meets and that \( \varphi \) respects joins, so from the fact that they are inverse to each other we see that they are both lattice isomorphisms. □
Definition A.9. If \(a, b\) are elements of a lattice, then we say that the intervals \([a, a \lor b]\) and \([a \land b, b]\) are perspective to each other, and we abbreviate this with either the notation

\[
[a, a \lor b] \succ [a \land b, b]
\]
or the notation

\[
[a \land b, b] \succ [a, a \lor b].
\]

If two intervals in a lattice can be connected by a chain of perspectivities, then we say that they are projective to each other.

The fact that all maximal chains in a finite length semimodular lattice have the same length can be strengthened to a lattice version of the Jordan-Hölder Theorem.

Theorem A.10 (Jordan-Hölder for semimodular lattices [57]). Suppose we have two maximal chains

\[
0 = a_0 < a_1 < \cdots < a_n = 1,
\]

\[
0 = b_0 < b_1 < \cdots < b_n = 1
\]
in an upper semimodular lattice of finite length. Then there is a permutation \(\sigma \in S_n\) such that each \([a_{i-1}, a_i]\) is projective in two steps (going \(\succ\), \(\prec\)) to \([b_{\sigma(i)-1}, b_{\sigma(i)}]\).

Proof. We induct on the length \(n\). If \(a_1 = b_1\) then we can apply the inductive hypothesis. Otherwise, for each \(i\), let \(c_i = a_1 \lor b_i\). If \(k\) is maximal such that \(a_1 \not\leq b_k\), then

\[
a_1 = c_0 < c_1 < \cdots < c_k = c_{k+1} < \cdots < c_n = 1
\]

where the strict inequalities up to \(c_k\) follow from upper semimodularity, and in the portion after \(c_{k+1}\) we have \(c_j = b_j\).

Applying the induction hypothesis, we get a bijection \(\sigma' : [n] \setminus \{1\} \to [n] \setminus \{k+1\}\) such that each \([a_{i-1}, a_i]\) is projective going \(\succ\), \(\prec\) to \([c_{\sigma'(i)-1}, c_{\sigma'(i)}]\). Since \([c_{\sigma'(i)-1}, c_{\sigma'(i)}]\) \(\succ\) \([b_{\sigma'(i)-1}, b_{\sigma'(i)}]\), and since \([a_0, a_1] \succ [b_k, b_{k+1}]\), we can take \(\sigma\) to be the extension of \(\sigma'\) given by setting \(\sigma(1) = k+1\).

To relate this to the usual Jordan-Hölder Theorem, we have to consider the lattice of subnormal subgroups of a group. A subgroup \(M \leq G\) is called subnormal if there is a finite chain of subgroups connecting it to \(G\), such that each is a normal subgroup of the next.

Proposition A.11. A subgroup \(M \leq G\) is subnormal iff the sequence of groups \(G = G_0 \triangleright G_1 \triangleright \cdots\) defined by taking \(G_{i+1}\) to be the normal closure of \(M\) inside \(G_i\) eventually reaches \(M\). As a consequence, the intersection of two subnormal subgroups is also subnormal.

Proposition A.12. If \(G\) is a group of finite composition length, then the collection of subnormal subgroups of \(G\) forms a lower semimodular lattice. If \([N_1, M_1]\), \([N_2, M_2]\) are \(\prec\), \(\succ\) projective covers in this lattice, then \(M_1/N_1 \cong M_2/N_2\).

Note that the modular law is equivalent to the following identity, which recovers the usual modular law in the case \(a \leq b\) by replacing \(a \land b\) with \(a\):

\[
(a \land b) \lor (c \land b) \approx ((a \land b) \lor c) \land b.
\]

Thus modular lattices form a variety of lattices. We finish our review of modular lattices by mentioning a famous result of Dedekind.
Proposition A.13 (Dedekind [47]). The free modular lattice on 3 generators is finite, with exactly 28 elements and length 8. It is isomorphic to a subdirect product of six copies of the two-element lattice and a single copy of the diamond lattice $\mathcal{M}_3$.

In particular, one can test whether a given 3-variable lattice identity is a consequence of modularity in finite time, by testing whether it holds on $\mathcal{M}_3$.

The corresponding result for 4 generators is false: the free modular lattice on 4 generators is infinite. To see this, note that if you start with four generic points on the projective plane and repeatedly generate new points and lines, the resulting set of points and lines you obtain is infinite. Determining whether a 4-variable lattice identity follows from the modular law is undecidable in general [63].

A.1 The Shifting Lemma and the Day terms

We will follow Freese and McKenzie [54], with some arguments taken from Gumm [59] and some from [94]. The starting point for proving things in congruence modular varieties is the Shifting Lemma (this is the main place in the theory where the modular law is actually used).

Lemma A.14 (Shifting Lemma). If $\mathbb{A}$ is congruence modular, $x, y, z, w \in \mathbb{A}$ and $\alpha, \beta, \gamma \in \text{Con}(\mathbb{A})$ with $\alpha \land \beta \leq \gamma$ and $x \equiv_\alpha y, z \equiv_\alpha w, x \equiv_\beta z, y \equiv_\beta w$, then $z \equiv_\gamma w \implies x \equiv_\gamma y$.

Proof. We have $(x, y) \in \alpha \land (\beta \circ (\alpha \land \gamma) \circ \beta) \subseteq \alpha \land (\beta \lor (\alpha \land \gamma))$. Since $\alpha \land \gamma \leq \alpha$, we can apply the modular law to get $\alpha \land (\beta \lor (\alpha \land \gamma)) = (\alpha \land \beta) \lor (\alpha \land \gamma)$, and this is contained in $\gamma$ by the assumption $\alpha \land \beta \leq \gamma$, so $(x, y) \in \gamma$. \hfill $\Box$

Corollary A.15 (Day terms). In any congruence modular variety $\mathcal{V}$, if $\mathcal{F}_V(x, y, z, w)$ is the free algebra on four generators, and if we let $\theta_{a,b}$ be the congruence generated by identifying $a, b$, then there are quaternary terms $m_0, \ldots, m_n \in \mathcal{F}_V(x, y, z, w)$ such that

\[
\begin{align*}
m_0 &= x, \\
m_i \ (\theta_{x,y} \lor \theta_{z,w}) \land (\theta_{x,z} \lor \theta_{y,w}) & m_{i+1} \text{ for } i \text{ even}, \\
m_i \theta_{z,w} m_{i+1} & m_{i+1} \text{ for } i \text{ odd}, \\
m_n &= y.
\end{align*}
\]

In other words, the $m_i$ satisfy the following system of identities:

\[
\begin{align*}
m_0(x, y, z, w) & \approx x, \\
m_i(x, x, z, z) & \approx x \text{ for all } i, \\
m_i(x, y, x, y) & \approx m_{i+1}(x, y, x, y) \text{ for } i \text{ even}, \\
m_i(x, y, z, z) & \approx m_{i+1}(x, y, z, z) \text{ for } i \text{ odd}, \\
m_n(x, y, z, w) & \approx y.
\end{align*}
\]
Proof. Apply the Shifting Lemma with $\alpha = \theta_{x,y} \lor \theta_{z,w}$, $\beta = \theta_{x,z} \lor \theta_{y,w}$, and $\gamma = (\alpha \land \beta) \lor \theta_{z,w}$ to see that $(x, y) \in (\alpha \land \beta) \lor \theta_{z,w} = \bigcup_n ((\alpha \land \beta) \lor \theta_{z,w})^n$. \hfill $\square$

Lemma A.16. Let $\mathbb{A}$ be an algebra with Day terms $m_0, ..., m_n$, $\theta \in \text{Con}(\mathbb{A})$, and $a, b, c, d \in \mathbb{A}$ with $(c, d) \in \theta$. Then $(a, b) \in \theta$ iff for all $i \leq n$ we have $m_i(a, b, a, b) \equiv_\theta m_i(a, a, b, b)$.

Proof. If $(a, b) \in \theta$, then for each $i$ we have $m_i(a, b, a, b) \equiv_\theta m_i(a, a, a, a) = a$ and $m_i(a, b, c, d) \equiv_\theta m_i(a, a, c, d) = a$. For the converse direction, we will show that if $c \equiv_\theta d$ and $m_i(a, b, a, b) \equiv_\theta m_i(a, b, c, d)$ for all $i$, then $m_i(a, b, c, d) \equiv_\theta m_{i+1}(a, b, c, d)$ for all $i$, and then we can conclude $a = m_0(a, b, c, d) \equiv_\theta m_n(a, b, c, d) = b$.

For $i$ even, we use $m_i(a, b, a, b) = m_{i+1}(a, b, a, b)$ together with the assumed congruences relating $m_i(a, b, a, b)$ to $m_i(a, b, c, d)$, while for $i$ odd we use $m_i(a, b, c, c) = m_{i+1}(a, b, c, c)$ together with $c \equiv_\theta d$. \hfill $\square$

The existence of Day terms implies a result slightly stronger than the Shifting Lemma, called the Shifting Principle.

Lemma A.17 (The Shifting Principle). If $\mathbb{A}$ has Day terms $m_0, ..., m_n$, then $\mathbb{A}$ satisfies the Shifting Principle: if $x, y, z, w \in \mathbb{A}$ and $\alpha, \gamma \in \text{Con}(\mathbb{A})$ and $\Lambda \leq \mathbb{A}^2$ is a reflexive relation preserved by the $m_i$ with $\alpha \cap \Lambda \subseteq \gamma$ and $x \equiv_\alpha y, z \equiv_\alpha w, (x, z) \in \Lambda, (y, w) \in \Lambda$, then $z \equiv_\gamma w \implies x \equiv_\gamma y$.

![Shifting Principle Diagram](image)

Proof. By Lemma A.16 it’s enough to show that $m_i(x, y, x, y) \equiv_\gamma m_i(x, y, z, w)$ for each $i$. Since $\Lambda$ is preserved by the $m_i$ and is reflexive, we have

$$
\begin{align*}
\begin{bmatrix}
m_i(x, y, x, y) \\
m_i(x, y, z, w)
\end{bmatrix} &= m_i\left(\begin{bmatrix}x \\ x\end{bmatrix}, \begin{bmatrix}y \\ y\end{bmatrix}, \begin{bmatrix}x \\ z\end{bmatrix}, \begin{bmatrix}y \\ w\end{bmatrix}\right) \in \Lambda, \\
\end{align*}
$$

while $m_i(x, y, x, y) \equiv_\alpha m_i(x, y, z, w)$ by Lemma A.16, so $(m_i(x, y, x, y), m_i(x, y, z, w)) \in \alpha \cap \Lambda \subseteq \gamma$. \hfill $\square$

Lemma A.18. If the Shifting Principle holds for an algebra $\mathbb{A}$ in the special case where $\alpha \geq \gamma$, then $\mathbb{A}$ is congruence modular.

Proof. Suppose that $\alpha, \beta, \gamma \in \text{Con}(\mathbb{A})$ with $\alpha \geq \gamma \geq \alpha \land \beta$, then to verify congruence modularity we just need to check that $\alpha \land (\beta \lor \gamma) \leq \gamma$, as this rules out the existence of a sublattice of $\text{Con}(\mathbb{A})$ isomorphic to the pentagon $\Lambda_5$. Defining reflexive, symmetric relations $\Lambda_i$ by $\Lambda_i = \beta \circ (\gamma \circ \beta)^{\circ 2}$, we see that we just need to prove that $\alpha \cap \Lambda_i \subseteq \gamma$ for each $i$.

We will prove this by induction on $i$: note that the base case $i = 0$ is trivial, since $\Lambda_0 = \beta$. For the inductive step, we apply the Shifting Principle to $\alpha, \Lambda_i$, and $\gamma$ see that if $\alpha \cap \Lambda_i \subseteq \gamma$, then

$$
\alpha \cap \Lambda_{2i+1} = \alpha \cap (\Lambda_i \circ \gamma \circ \Lambda_i) = \alpha \cap (\Lambda_i \circ (\alpha \land \gamma) \circ \Lambda_i) \subseteq \gamma.
$$

Corollary A.19. A variety is congruence modular iff it has Day terms.
Example A.1. If $p(x, y, z)$ is a Mal’cev term, then we can take
\[
\begin{align*}
m_0(x, y, z, w) &= x, \\
m_1(x, y, z, w) &= p(z, w, y), \\
m_2(x, y, z, w) &= y
\end{align*}
\]
as a sequence of Day terms. Rather than laboriously checking the Day identities, it is easier to verify that this sequence of terms can be used in the Shifting Lemma setup to show that $x \equiv \gamma y$.

We have $x (\alpha \land \beta) p(z, w, y) \gamma y$, so from $\alpha \land \beta \leq \gamma$ we get $x \gamma y$.

Example A.2. If $g(x, y, z)$ is a majority term, then we can take
\[
\begin{align*}
m_0(x, y, z, w) &= x, \\
m_1(x, y, z, w) &= g(x, y, z), \\
m_2(x, y, z, w) &= g(x, y, w), \\
m_3(x, y, z, w) &= y
\end{align*}
\]
as a sequence of Day terms. Again, in the Shifting Lemma setup, we have $x (\alpha \land \beta) g(x, y, z) \gamma g(x, y, w) (\alpha \land \beta) y$, so $x \gamma y$.

The next corollary gives us a large class of examples of congruence modular varieties, generalizing groups and rings.

Definition A.20. An algebra is congruence regular if every congruence on $\mathbb{A}$ is uniquely determined by any of its congruence classes.

Corollary A.21 (Gumm [59]). If every subalgebra of $\mathbb{A}^2$ is congruence regular, then $\mathbb{A}$ is congruence modular.

Proof. We just need to verify the Shifting Principle for $\mathbb{A}$. Let $\Lambda \leq \mathbb{A}^2$ be reflexive, let $\alpha \geq \gamma$ be congruences on $\mathbb{A}$ with $\alpha \cap \Lambda \subseteq \gamma$, and consider the congruences $\alpha \times \gamma$ and $\gamma \times \gamma$ restricted to $\Lambda$. We will show that for any $a \in \mathbb{A}$, the congruence classes containing $(a, a)$ in these restrictions are equal, so congruence regularity will imply that $\alpha \times \gamma|\Lambda = \gamma \times \gamma|\Lambda$, which is the Shifting Principle for $\alpha, \Lambda, \gamma$.

So suppose that $(a, a) \equiv_{\alpha \times \gamma} (b, c) \in \Lambda$. Then $(b, c) \in \alpha \circ \gamma = \alpha$, so $(b, c) \in \alpha \cap \Lambda \subseteq \gamma$, and this implies that $(a, b) \in \gamma \circ \gamma = \gamma$. Thus $(a, a) \equiv_{\gamma \times \gamma} (b, c)$ as well, and we are done.

To finish this subsection, we will prove one of Gumm’s “geometric” results on congruence modular varieties, which generalizes the result used to prove associativity of the loop operation in the case of abelian Mal’cev algebras.

Lemma A.22 (The Cube Lemma [59]). Suppose every subalgebra of $\mathbb{A}^2$ satisfies the Shifting Lemma. If $\alpha, \beta, \gamma \in \text{Con}(\mathbb{A})$ with $\gamma \geq \alpha \land \beta$, and if $a, b, c, d, a', b', c', d' \in \mathbb{A}$ with $(a, b), (c, d), (a', b'), (c', d') \in \alpha, (a, d), (b, c), (a', d'), (b', c') \in \beta$, and $(a, a'), (b, b'), (c, c') \in \gamma$, then $(d, d') \in \gamma$. 

171
Proof. We apply the Shifting Lemma to the algebra $\beta \leq A^2$, and the congruences $\gamma \times 1_A|\beta, \alpha \times \alpha|\beta$, and $\gamma \times \gamma|\beta$. By the Shifting Lemma applied to $\alpha, \beta, \gamma$, we have $(\alpha \times \alpha|\beta) \land (\gamma \times 1_A|\beta) \leq \gamma \times \gamma|\beta$, so the Shifting Lemma applies to $\gamma \times 1_A|\beta, \alpha \times \alpha|\beta, \gamma \times \gamma|\beta$.

Thus, from $((b, c), (b', c')) \in \gamma \times \gamma|\beta$, $((a, d), (a', d')) \in \gamma \times 1_A|\beta$, and $((b, c), (a, d)), ((b', c'), (a', d')) \in \alpha \times \alpha|\beta$, the Shifting Lemma allows us to conclude that $((a, d), (a', d')) \in \gamma \times \gamma|\beta$, so $(d, d') \in \gamma$.

A.2 The modular commutator

First we go over a slick proof of the main properties of the commutator, using the Day terms to construct an explicit set of generators $X(\alpha, \beta)$ for the congruence $[\alpha, \beta]$ - however, as the approach feels somewhat ad-hoc, we will also prove these properties via a different approach based on the Shifting Lemma applied to congruences (as in the proof of the Cube Lemma). The definition of $X(\alpha, \beta)$ is based on the algebra of matrices $M(\alpha, \beta)$ used to visualize the term condition (Definition 10.28) and Lemma A.16.

Definition A.23. Suppose $A$ has Day terms $m_0, \ldots, m_n$. For $\alpha, \beta \in \text{Con}(A)$, we define $X(\alpha, \beta)$ to be the set of pairs $(m_i(a, b, a, b), m_i(a, b, c, d))$ for $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in M(\alpha, \beta)$ and $i \leq n$.

Example A.3. If we have a Mal’cev term $p(x, y, z)$ and take $m_0 = x, m_1 = p(z, w, y), m_2 = y$ as our sequence of Day terms, then $X(\alpha, \beta)$ is the set of pairs $(a, p(c, d, b))$ for $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in M(\alpha, \beta)$.

Example A.4. If we have a majority term $g(x, y, z)$ and take $m_0 = x, m_1 = g(x, y, z), m_2 = g(x, y, w), m_3 = y$ as our sequence of Day terms, then $X(\alpha, \beta)$ is the set of pairs $(a, g(a, b, c))$ for $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in M(\alpha, \beta)$. For $(a, b) \in \alpha \land \beta$, we have

$$g\left(\begin{bmatrix} a & a \\ b & b \end{bmatrix}, \begin{bmatrix} a & b \\ b & b \end{bmatrix}\right) = \begin{bmatrix} a & b \\ b & b \end{bmatrix} \in M(\alpha, \beta),$$

so $(a, g(a, b, b)) = (a, b) \in X(\alpha, \beta)$. Thus $X(\alpha, \beta) = \alpha \land \beta$ for majority algebras.

Theorem A.24 (Commutator via Day terms). If $A$ has Day terms $m_0, \ldots, m_n$ and $\alpha, \beta, \delta \in \text{Con}(A)$, then the following are equivalent.

(i) $X(\alpha, \beta) \subseteq \delta$,

(ii) $X(\beta, \alpha) \subseteq \delta$,

(iii) $C(\alpha, \beta; \delta)$ holds,

(iv) $C(\beta, \alpha; \delta)$ holds,

(v) $[\alpha, \beta] \leq \delta$.

Proof. It’s enough to show (iii) $\implies$ (i) $\implies$ (iv). For (iii) $\implies$ (i), suppose that $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in M(\alpha, \beta)$, then

$$m_i\left(\begin{bmatrix} a & a \\ a & a \end{bmatrix}, \begin{bmatrix} a & a \\ b & b \end{bmatrix}, \begin{bmatrix} a & c \\ b & d \end{bmatrix}\right) = \begin{bmatrix} a & a \\ m_i(a, b, a, b) & m_i(a, b, c, d) \end{bmatrix} \in M(\alpha, \beta),$$

172
so \( C(\alpha, \beta; \delta) \) implies that we have \((m_i(a, b, a, b), m_i(a, b, c, d)) \in \delta. \)

For (i) \( \implies \) (iv), we apply Lemma [A.16] to see that if \( \begin{bmatrix} a & c \\ b & d \end{bmatrix} \in M(\alpha, \beta), (c, d) \in \delta, \) and \( X(\alpha, \beta) \subseteq \delta, \) then we must have \((a, b) \in \delta \) as well, so \( C(\beta, \alpha; \delta) \) holds.

Now we can finally prove some useful properties of commutators.

**Proposition A.25.** If \( \mathbb{A} \) is contained in a congruence modular variety, then for congruences on \( \mathbb{A} \) we have

(a) \( \left[ \alpha, \beta \right] = \left[ \beta, \alpha \right], \)

(b) \( \left[ \alpha \land \gamma, \beta \right] \leq \left[ \alpha, \beta \right] \land \gamma, \)

(c) \( \bigvee_i \alpha_i, \beta = \bigvee_i \left[ \alpha_i, \beta \right], \)

(d) if \( f : \mathbb{A} \to \mathbb{B} \) is surjective, then \( f(\left[ \alpha, \beta \right] \lor f) = \left[ f(\alpha \lor \ker f), f(\beta \lor \ker f) \right], \)

(e) if \( \mathbb{B} \leq \mathbb{A}, \) then \( \left[ \alpha | \mathbb{B}, \beta | \mathbb{B} \right] \leq \left[ \alpha, \beta \right] | \mathbb{B}, \)

(f) if \( \mathbb{A} = \prod_i \mathbb{A}_i, \) then \( \bigoplus_i \alpha_i, \bigoplus_i \beta_i = \bigoplus_i \left[ \alpha_i, \beta_i \right], \) where \( \bigoplus_i \alpha_i \) is the set of pairs \((a, b)\) in \( \prod_i \mathbb{A}_i \) such that for all but finitely many \( i \) we have \( a_i = b_i, \)

(g) if \( \mathbb{A} = \prod_i \mathbb{A}_i, \) then \( \prod_i \alpha_i, \prod_i \beta_i \leq \prod_i \left[ \alpha_i, \beta_i \right]. \)

**Proof.** Part (a) follows from Theorem [A.24], part (b) follows from Proposition [10.30] (d), and part (e) is Proposition [10.30] (g). For part (c), Theorem [A.24] shows that \( C(\alpha_j, \beta; \bigvee_i \left[ \alpha_i, \beta \right]) \) holds for each \( j, \) so we can use Proposition [10.30] (e) to see that \( \bigvee_i \alpha_i, \beta \leq \bigvee_i \left[ \alpha_i, \beta \right], \) while the other inequality follows from monotonicity of the commutator.

For part (d), note that part (c) implies \( \left[ \alpha, \beta \right] \lor \ker f = \left[ \alpha \lor \ker f, \beta \lor \ker f \right] \lor \ker f, \) so we may assume that \( \alpha, \beta \geq \ker f \) without loss of generality. By Theorem [A.24] \( \left[ \alpha, \beta \right] \lor \ker f \) is the congruence generated by \( X(\alpha, \beta) \lor \ker f, \) and \( \left[ f(\alpha), f(\beta) \right] \) is the congruence generated by \( X(f(\alpha), f(\beta)) = f(X(\alpha, \beta)), \) so \( f(\left[ \alpha, \beta \right] \lor \ker f) = \left[ f(\alpha), f(\beta) \right]. \)

Parts (f) and (g) follow directly from Theorem [A.24] but they can also be proved using only parts (a) - (d) (left as an exercise to the reader).

Proposition [A.25] (d) tells us that we can compute commutators on quotients of \( \mathbb{A} \) directly in \( \mathbb{A}. \)

Since \( \text{Con}(\mathbb{A}/\pi) \) is naturally isomorphic to the interval \([\pi, 1_{\mathbb{A}}]\) in \( \text{Con}(\mathbb{A}), \) computing commutators on \( \mathbb{A}/\pi \) is equivalent to computing relative commutators on \( \mathbb{A}. \) Recall that if \( \alpha, \beta \geq \pi, \) then their relative commutator \( \left[ \alpha, \beta \right]_\pi \) is defined to be the least \( \delta \geq \pi \) which satisfies the term condition \( C(\alpha, \beta; \delta). \)

**Corollary A.26.** If \( \alpha, \beta \geq \pi \) are congruences in a congruence modular variety, then their relative commutator is given by the formula \( \left[ \alpha, \beta \right]_\pi = \left[ \alpha, \beta \right] \lor \pi. \)

**Theorem A.27** (Diamond Isomorphism Theorem for relative commutators). If \( \mathbb{A} \) is in a congruence modular variety and \( \alpha, \beta \in \text{Con}(\mathbb{A}), \) then the maps \( \phi : [\alpha, \alpha \lor \beta] \to [\alpha \land \beta, \beta] \) and \( \varphi : \left[ \alpha \land \beta, \beta \right] \to \left[ \alpha, \alpha \lor \beta \right] \) given by

\[
\phi : x \mapsto x \land \beta \quad \text{and} \quad \varphi : y \mapsto y \lor \alpha
\]

are lattice isomorphisms which respect the relative commutators \( [\cdot, \cdot]_\alpha, [\cdot, \cdot]_{\alpha \land \beta}. \)

Furthermore, in this case we have the equality of relative centralizers \( (\alpha : \alpha \lor \beta) = (\alpha \land \beta : \beta). \)
Proof. By Theorem A.8, \( \phi \) and \( \varphi \) are lattice isomorphisms. If \( \gamma, \delta \geq \alpha \land \beta \), then from \( [\gamma \lor \alpha, \delta \lor \alpha] \leq [\gamma, \delta] \lor \alpha \) we have
\[
\varphi([\gamma, \delta]_{\alpha \land \beta}) = [\gamma, \delta]_{\alpha \land \beta} \lor \alpha = [\gamma \lor \alpha, \delta \lor \alpha] \lor \alpha = [\varphi(\gamma), \varphi(\delta)]_{\alpha}.
\]
If \( \gamma, \delta \in [\alpha, \alpha \lor \beta] \), then from \( \gamma = \phi(\gamma) \lor \alpha, \delta = \phi(\delta) \lor \alpha \) and \( [\phi(\gamma), \phi(\delta)] \leq \beta \) we have
\[
\phi([\gamma, \delta]_{\alpha}) = [\gamma, \delta]_{\alpha} \lor \beta = ([\phi(\gamma), \phi(\delta)] \lor \alpha) \land \beta = [\phi(\gamma), \phi(\delta)]_{\alpha \land \beta}.
\]

For the last statement, note that
\[
[\delta, \alpha \lor \beta] \leq \alpha \iff [\delta, \alpha] \lor [\delta, \beta] \leq \alpha \iff [\delta, \beta] \leq \alpha \land \beta,
\]
so \( \delta \leq (\alpha : \alpha \lor \beta) \iff \delta \leq (\alpha \land \beta : \beta) \).

For the second approach to the commutator, we will follow Gumm [59] and take the transitive closure of \( M(\alpha, \beta) \) to produce a congruence on \( \alpha \), considered as a subalgebra of \( A^2 \).

**Definition A.28.** For \( \alpha, \beta \in \text{Con}(A) \), if we consider \( \alpha \leq A^2 \) as an algebra of column vectors then we can treat \( M(\alpha, \beta) \) (from Definition 10.28) as a binary relation on \( \alpha \), so \( \left( \begin{array}{cc} a & c \\ b & d \end{array} \right) \in M(\alpha, \beta) \)
means that \( \left[ \begin{array}{cc} a & c \\ b & d \end{array} \right] \in M(\alpha, \beta) \). We define \( \Delta^\beta_\alpha \) to be the transitive closure of this binary relation on \( \alpha \).

Note that \( \Delta^\beta_\alpha \) is the least congruence on \( \alpha \) which contains the binary relation \( \beta \times \beta \vert_{\Delta_A} \). When \( A \) has a Mal’cev term, \( \Delta^\beta_\alpha \) simplifies to \( M(\alpha, \beta) \).

**Proposition A.29.** If \( A \) has a Mal’cev polynomial \( p \) and \( \alpha, \beta \in \text{Con}(A) \), then \( \Delta^\beta_\alpha = M(\alpha, \beta) \).

**Proof.** We just need to check that \( M(\alpha, \beta) \) is transitively closed, so supposed that \( \left[ \begin{array}{cc} a & c \\ b & d \end{array} \right], \left[ \begin{array}{cc} c & e \\ d & f \end{array} \right] \in M(\alpha, \beta) \). Then we have
\[
\left[ \begin{array}{cc} a & e \\ b & f \end{array} \right] = p \left( \left[ \begin{array}{cc} a & c \\ b & d \end{array} \right], \left[ \begin{array}{cc} c & c \\ d & d \end{array} \right], \left[ \begin{array}{cc} c & e \\ d & f \end{array} \right] \right) \in M(\alpha, \beta). \]

**Theorem A.30.** Suppose that the Shifting Lemma holds for every subalgebra of \( A^2 \). Then for \( x, y \in A \) and \( \alpha, \beta \in \text{Con}(A) \), the following are equivalent:

(a) \((x, y) \in [\beta, \alpha]\),

(b) \([x \ y \ y y] \in \Delta^\beta_\alpha\),

(c) there exists \( a \in A \) such that \([x \ y \ a a] \in \Delta^\beta_\alpha\),

(d) there exists \( b \in A \) such that \([x \ b \ b y] \in \Delta^\beta_\alpha\).

174
Proof. That (b) implies (c), (d) are clear, and (c) implies (a) directly from the term condition $C(\beta, \alpha; [\beta, \alpha])$ and the definition of $\Delta_\alpha^\beta$. That (c) implies (b) follows from the fact that $(a, y) \in \beta \implies \begin{bmatrix} a & y \\ a & y \end{bmatrix} \in \mathcal{M}(\alpha, \beta)$, and since $\Delta_\alpha^\beta$ is the transitive closure of $\mathcal{M}(\alpha, \beta)$ we have $\begin{bmatrix} x & y \\ y & y \end{bmatrix} \in \Delta_\alpha^\beta \circ \mathcal{M}(\alpha, \beta) = \Delta_\alpha^\beta$.

For $(d) \implies (b)$ we apply the Shifting Lemma to the algebra $\alpha \leq_{sd} A \times A$, the congruences $\ker \pi_1, \ker \pi_2, \Delta_\alpha^\beta \in \text{Con}(\alpha)$, and the elements $\begin{bmatrix} x \\ b \end{bmatrix}, \begin{bmatrix} y \\ b \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} y \\ y \end{bmatrix} \in \alpha$ (that $x \equiv_\alpha y$ follows from $x \equiv_\alpha b \equiv_\alpha y$).

For $(a) \implies (b)$, we will show that the relation $\Theta$ defined by $(x, y) \in \Theta \iff \begin{bmatrix} x & y \\ y & y \end{bmatrix} \in \Delta_\alpha^\beta$ is a congruence which satisfies $C(\beta, \alpha; \Theta)$, which will show that $[\beta, \alpha] \leq \Theta$. That $\Theta$ is reflexive is obvious, that it is symmetric follows from the equivalence of (b) with (c) or (d). If $(x, y), (y, z) \in \Theta$, then from $\begin{bmatrix} x & y \\ y & y \end{bmatrix}, \begin{bmatrix} y & z \\ y & y \end{bmatrix} \in \Delta_\alpha^\beta$ and the fact that $\Delta_\alpha^\beta$ is transitively closed, we get $\begin{bmatrix} x & z \\ y & y \end{bmatrix} \in \Delta_\alpha^\beta$, so $(x, z) \in \Theta$ by the equivalence of (b) and (d).

To finish, we just need to show that $\Theta$ satisfies $C(\beta, \alpha; \Theta)$, that is, if $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in \mathcal{M}(\alpha, \beta)$ with $(c, d) \in \Theta$, then $(a, b) \in \Theta$. But if $(c, d) \in \Theta$, then $\begin{bmatrix} c & d \\ d & d \end{bmatrix} \in \Delta_\alpha^\beta$, so since $\Delta_\alpha^\beta$ is the transitive closure of $\mathcal{M}(\alpha, \beta)$, we see that $\begin{bmatrix} a & d \\ b & d \end{bmatrix} \in \Delta_\alpha^\beta$, so by the equivalence of (b) with (c) we have $(a, b) \in \Theta$. \qed

Corollary A.31. If every subalgebra of $A^2$ satisfies the Shifting Lemma, then for $\alpha, \beta_i \in \text{Con}(A)$ we have $[\bigvee_i \beta_i, \alpha] = \bigvee_i [\beta_i, \alpha]$.

Proof. We have $\Delta_{\bigvee_i \beta_i}^{\alpha} = \bigvee_i \Delta_{\alpha}^{\beta_i}$, so $(x, y) \in [\bigvee_i \beta_i, \alpha]$ iff $\begin{bmatrix} x & z \\ y & z \end{bmatrix} \in \bigvee_i \Delta_{\alpha}^{\beta_i}$ for some $z$. So there must be a sequence $(x_i, y_i) \in \alpha, j_i$, with $\begin{bmatrix} x_i & x_{i+1} \\ y_i & y_{i+1} \end{bmatrix} \in \Delta_{\alpha}^{\beta_{j_i}}$ and $(x, y) = (x_n, y_n), x_0 = y_0$.

We show by induction on $i$ that $\begin{bmatrix} x_i & y_i \\ y_i & y_i \end{bmatrix} \in \bigvee_j (\Delta_{\alpha}^{\beta_j} \land \ker \pi_2)$, this will show that $(x_i, y_i) \in \bigvee_j [\beta_j, \alpha]$. For the inductive step, we apply the Shifting Lemma to $\alpha \leq A^2$ with the congruences $\ker \pi_2, \Delta_{\alpha}^{\beta_{j_i}}, \bigvee_j (\Delta_{\alpha}^{\beta_{j}} \land \ker \pi_2)$.
Corollary A.32. If every subalgebra of $\mathbb{A}^2$ satisfies the Shifting Lemma, then for $f : \mathbb{A} \to \mathbb{B}$ surjective and $\alpha, \beta \geq \ker f$, we have $f([\beta, \alpha] \vee \ker f) = [f(\beta), f(\alpha)]$.

Proof. The hard direction is to check that if $(f(x), f(y)) \in [f(\beta), f(\alpha)]$, then $(x, y) \in [\beta, \alpha] \vee \ker f$. In this case we have $\begin{bmatrix} x & y \\ y & y \end{bmatrix} \in \Delta^\beta_\alpha \vee (\ker f \times \ker f |_\alpha)$. Using a similar argument to the previous corollary, we can show that this implies $\begin{bmatrix} x & y \\ y & y \end{bmatrix} \in (\Delta^\beta_\alpha \wedge \ker \pi_2) \vee (\ker f \times \ker f |_\alpha)$ by repeatedly applying the Shifting Lemma on $\alpha$. Thus we have $(x, y) \in [\beta, \alpha] \vee \ker f$ by Theorem A.30. \qed

To prove the symmetry of the commutator, we will actually prove a stronger statement: $\Delta^\beta_\alpha$ is in fact the transpose of $\Delta^\alpha_\beta$. In particular, if we view $\Delta^\beta_\alpha$ as a binary relation on row vectors in $\beta$, then $\Delta^\beta_\alpha$ will be transitively closed (which is far from obvious from the definition!).

Theorem A.33. Suppose that the Shifting Lemma holds for every subalgebra of $\mathbb{A}^4$. If $\overline{\Delta}_\alpha^\beta$ denotes the set of transposes of matrices from $\Delta^\beta_\alpha$, then $\overline{\Delta}_\alpha^\beta$ is transitively closed as a binary relation on $\beta$ and we have $\overline{\Delta}_\alpha^\beta = \Delta^\beta_\alpha$. In particular, we have $[\alpha, \beta] = [\beta, \alpha]$.

Proof. It’s enough to prove that $\overline{\Delta}_\alpha^\beta$ is transitively closed as a binary relation on $\beta$, as we will then have $\Delta^\beta_\alpha = \bigcup_n M(\beta, \alpha)^n \subseteq \overline{\Delta}_\alpha^\beta$, and a symmetric argument with $\alpha, \beta$ swapped will show that $\Delta^\beta_\alpha \subseteq \overline{\Delta}_\alpha^\beta$, so $\Delta^\beta_\alpha \subseteq \overline{\Delta}_\alpha^\beta \subseteq \Delta^\beta_\alpha$.

Suppose that $\begin{bmatrix} a & b \\ c & d \end{bmatrix}, \begin{bmatrix} e & f \\ c & d \end{bmatrix} \in \Delta^\beta_\alpha$. To finish, we just need to show that $\begin{bmatrix} a & b \\ e & f \end{bmatrix} \in \Delta^\beta_\alpha$. This follows from the following application of the Cube Lemma (Lemma A.22) applied to the congruences $\ker \pi_1, \ker \pi_2, \Delta^\beta_\alpha$ on $\alpha$. 

176
Theorem A.34. In a congruence modular variety, any alternative commutator $[\cdot,\cdot]'$ which satisfies $[\alpha,\beta]' \leq \alpha \wedge \beta$ and $f([\alpha,\beta]' \lor \ker f) = [f(\alpha),f(\beta)]'$ for $f$ surjective and $\alpha,\beta \geq \ker f$ has $[\alpha,\beta]' \leq [\alpha,\beta]$ for all $\alpha,\beta$.

Proof. Consider congruences on $\alpha \leq \ker \pi_2 \times \ker \pi_1$. We have $[\Delta_\alpha^\beta \land \ker \pi_2] \leq \Delta_\alpha^\beta \land \ker \pi_2 \leq \pi_1^{-1}[\beta,\alpha]$ by Theorem A.30. Also, $\alpha = \pi_1(\ker \pi_2 \lor \ker \pi_1)$, so

$$[\beta,\alpha]' = [\pi_1(\Delta_\alpha^\beta \lor \ker \pi_1),\pi_1(\ker \pi_2 \lor \ker \pi_1)]' = \pi_1([\Delta_\alpha^\beta \lor \ker \pi_2]' \lor \ker \pi_1) \leq [\beta,\alpha].$$

A.3 The Gumm difference term

In this subsection we prove that congruence modular varieties have a ternary term $p$, called a Gumm difference term, which acts like a Mal’cev operation on all abelian algebras. This will imply that abelian algebras in congruence modular varieties are affine.

Theorem A.35 (Gumm difference term). For any variety with Day terms $m_0,\ldots,m_n$, there is a ternary term $p$ satisfying the following two properties:

(i) $p$ satisfies the identity $p(y,y,x) \approx x$, and

(ii) for any $(x,y) \in \Theta$, $\Theta$ any congruence, we have $p(x,y,y) [\Theta,\Theta] x$.

Furthermore, in a congruence modular variety, a ternary term $p$ satisfies (i) and (ii) iff it satisfies the following property:

(iii) for any congruences $\alpha,\beta,\gamma$ with $\alpha \land \beta \leq \gamma$, the implication in the following picture holds.
Finally, if a variety has a term $p$ which satisfies (iii), then it is congruence modular.

Proof. Recall the identities satisfied by Day terms:

$$m_0(x, y, z, w) \approx x,$$
$$m_i(x, x, z, z) \approx x \text{ for all } i,$$
$$m_i(x, y, x, y) \approx m_{i+1}(x, y, y) \text{ for } i \text{ even},$$
$$m_i(x, y, z, z) \approx m_{i+1}(x, y, z, z) \text{ for } i \text{ odd},$$
$$m_n(x, y, z, w) \approx y.$$

We inductively define a sequence of ternary terms $q_i(x, y, z)$ by

$$q_{i+1}(x, y, z) = \begin{cases} m_{i+1}(q_i(x, y, z), q_i(x, y, z), x) & \text{if } i \text{ odd}, \\ m_{i+1}(q_i(x, y, z), q_i(x, y, z), x, y) & \text{if } i \text{ even}, \end{cases}$$

and we set $p(x, y, z) = q_n(x, y, z)$.

To see that (i) holds, we just check inductively that $q_i(y, y, x) \approx x$:

$$q_{i+1}(y, y, x) = m_{i+1}(q_i(y, y, x), q_i(y, y, x), y, y) \approx m_{i+1}(x, y, y) \approx x.$$ 

For (ii), we will inductively check that

$$q_i(y, x, x) [\theta, \theta] m_{i+1}(m_i(y, x, x), m_i(y, x, y), y, x).$$

Using the term condition $C(\theta, \theta; [\theta, \theta])$, from

$$m_{i+1}(m_i(x, y, x, y), m_i(x, y, y), y, \overline{y}) = m_i(x, y, x, y) = m_{i+1}(x, y, y) = m_{i+1}(m_i(x, x, x), m_i(y, y, y), x, \overline{y}),$$

we conclude

$$m_{i+1}(m_i(x, y, y), m_i(x, y, y), y, \overline{x}) [\theta, \theta] m_{i+1}(m_i(x, x, x), m_i(y, y, y), x, \overline{x}) = m_{i+1}(x, y, x),$$

so $q_{i+1}(y, x, x) [\theta, \theta] m_{i+1}(x, y, x)$.

When $i$ is odd, the proof is very similar. Inductively, we have

$$q_{i+1}(y, x, x) = m_{i+1}(q_i(y, x, x), q_i(y, x, x), y, y) [\theta, \theta] m_{i+1}(m_i(x, y, x), m_i(x, y, x), x, y).$$
Using the term condition $C(\theta, \theta; [\theta, \theta])$, from

$$m_{i+1}(m_i(x, y, x, x), m_i(x, y, x, x), x, [x]) = m_i(x, y, x, x) = m_{i+1}(x, y, x, x)$$

$$= m_{i+1}(m_i(x, x, x, x), m_i(x, y, y, y), x, [x]),$$

we conclude

$$m_{i+1}(m_i(x, y, x, x), m_i(x, y, x, x), x, [y]) = m_{i+1}(x, y, y, y),$$

so $q_{i+1}(x, y, x) = m_{i+1}(x, y, y)$. This conclude the proof of (ii).

Now we show that (i) and (ii) imply (iii). Suppose we have the configuration

```
\begin{array}{|c|c|c|}
\hline
\gamma & \alpha & \beta \\
\hline
x & y & z' \\
\hline
\end{array}
```

with $\gamma \geq \alpha \wedge \beta$. From $x \equiv_\beta y \equiv_\beta z$, we have $p(x, y, z) \equiv_\beta z$. Additionally, we have $p(x, y, z) \equiv_\gamma p(y', y, z)$, so we just need to prove that $p(y', y, z) \equiv_\gamma z'$ to finish.

We have $p(y', y, z) \equiv_\alpha p(y', y', z') = z'$, and $p(y', y, z) \equiv_\beta p(z', z, z)$. From $(z, z') \in \alpha \wedge (\beta \lor \gamma)$, we have

$$p(z', z, z) [\alpha \wedge (\beta \lor \gamma), \alpha \wedge (\beta \lor \gamma)] z'.$$

The commutator above is bounded by

$$[\alpha \wedge (\beta \lor \gamma), \alpha \wedge (\beta \lor \gamma)] \leq [\alpha, \beta \lor \gamma] = [\alpha, \beta] \lor [\alpha, \gamma] \leq (\alpha \wedge \beta) \lor (\alpha \wedge \gamma) = \alpha \wedge \gamma.$$

Thus, we have $(p(y', y, z), z') \in \alpha \wedge (\beta \lor (\alpha \wedge \gamma))$, and by the modular law this is $(\alpha \wedge \beta) \lor (\alpha \wedge \gamma) = \alpha \wedge \gamma \leq \gamma$.

Finally, assume that $d$ is a term which satisfies (iii). Taking $x = y = y'$, $z = z'$, $\alpha = \gamma = 0_A$, $\beta = 1_A$, we get $p(y, y, z) = z$, which is (i). Taking $x = y$ and using $p(y, y, z) = z$, we see that (iii) implies the Shifting Lemma in every algebra, so our variety is congruence modular.

To prove that (iii) implies (ii), suppose $(x, y) \in \theta$, and consider the congruences $\ker \pi_1, \ker \pi_2, \Delta_\theta^\theta$ on $\theta$. Applying (iii) in the picture

```
\begin{array}{|c|c|c|}
\hline
[y] & y \\
\hline
[x] & y \\
\hline
\end{array}
```

and

```
\begin{array}{|c|c|c|}
\hline
y & y \\
\hline
x & x \\
\hline
\end{array}
```

we see that $\begin{bmatrix} x & y \\ p(x, y, y) & y \end{bmatrix} \in \Delta_\theta^\theta$, so by Theorem A.30 we have $p(x, y, y) [\theta, \theta] x$. 

□
Corollary A.36 (Factor Permutability). If $\mathbb{A} = \mathbb{A}_1 \times \mathbb{A}_2$ is contained in a congruence modular variety, then the factor congruences $\ker \pi_1, \ker \pi_2$ permute with every congruence $\gamma \in \text{Con}(\mathbb{A})$.

Proof. A pair of congruences $\alpha, \beta \in \text{Con}(\mathbb{A})$ correspond to a pair of factor congruences iff they satisfy $\alpha \land \beta = 0_\mathbb{A}$ and $\alpha \circ \beta = 1_\mathbb{A}$. Thus, if $x \gamma y' \alpha z'$, then by $\alpha \circ \beta = 1_\mathbb{A}$ we can find $y, z \equiv_\alpha x$ with $(y, y'), (z, z') \in \beta$. Then from $\gamma \geq 0_\mathbb{A} = \alpha \land \beta$ we can use property (iii) of a difference term to see that $x \alpha d(x, y', z') \gamma z'$, so $(x, z') \in \gamma \circ \alpha \implies (x', z') \in \alpha \circ \gamma$. \hfill \Box

Corollary A.37. Any abelian algebra which is contained in a congruence modular variety is affine.

Corollary A.38. A nontrivial algebra $\mathbb{A}$ in a congruence modular variety is abelian iff there is some $\mathbb{B} \leq_{sd} \mathbb{A} \times \mathbb{A}$ such that $\mathcal{M}_3$ is a $0,1$-sublattice of $\text{Con}(\mathbb{B})$.

Proof. If $\mathbb{A}$ is abelian, then it is affine and we can take $\mathbb{B} = \mathbb{A} \times \mathbb{A}$. For the other direction, it suffices to prove that $\mathbb{B}$ is abelian if $\mathcal{M}_3$ is a $0,1$-sublattice of $\text{Con}(\mathbb{B})$, since then $\mathbb{B}$ is affine and $\mathbb{A}$ is a quotient of $\mathbb{B}$, so $\mathbb{A}$ is also affine.

Let $\alpha, \beta, \gamma \in \text{Con}(\mathbb{B})$ generate a copy of $\mathcal{M}_3$ which is a $0,1$-sublattice. Then

$$[1, 1] = \langle \alpha \lor \beta, \alpha \lor \gamma \rangle = \langle \alpha, \alpha \rangle \lor \langle \alpha, \gamma \rangle \lor \langle \beta, \alpha \rangle \lor \langle \beta, \gamma \rangle \leq \alpha \lor (\beta \land \gamma) = \alpha.$$ 

Similarly we have $[1, 1] \leq \beta$, so $[1, 1] \leq \alpha \land \beta = 0$. \hfill \Box

By plugging a difference term into itself, we can strengthen property (ii) of a Gumm difference term, to get terms which act as Mal’cev operations on solvable algebras.

Definition A.39. For any congruence $\alpha$, define $[\alpha]^n$ inductively by $[\alpha]^0 = \alpha, [\alpha]^{n+1} = [[\alpha]^n, [\alpha]^n]$.

Proposition A.40. If $p$ is a Gumm difference term, and if we define terms $p_n$ inductively by $p_0 = p$ and

$$p_{n+1}(x, y, z) = p_n(x, p_n(x, y, y), p_n(x, y, y)), \quad \text{then each } p_n \text{ is also a Gumm difference term, and for any } (x, y) \in \theta \text{ we have } p_n(x, y, y) \theta^{2^n} x.$$ 

Proof. Inductively, we have

$$p_{n+1}(y, y, x) = p_n(y, p_n(y, y, y), p_n(y, y, y)) = p_n(y, y, x) = x,$$

and from $(x, p_n(x, y, y)) \in \theta^{2^n}$, we have

$$p_{n+1}(x, y, y) = p_n(x, p_n(x, y, y), p_n(x, y, y)) \theta^{2^n} x.$$ \hfill \Box

The last result of this subsection is useful for understanding the center of an algebra in terms of the difference term.

Theorem A.41. Suppose $p$ is a Gumm difference term for a congruence modular variety and $\alpha \geq \beta$. Then $\Delta^\alpha_\beta$ is given by

$$\begin{bmatrix} x & w \\ y & z \end{bmatrix} \in \Delta^\alpha_\beta \iff (p(x, y, z) \theta \beta w \land (x \beta y \alpha z)).$$
Proof. If \( \begin{bmatrix} x & w \\ y & z \end{bmatrix} \in \Delta_\alpha^\beta \) then clearly \( (x, y, z) \), and from
\[
p \left( \begin{bmatrix} x & x \\ x & y \\ x & y \\ x & w \\ y & z \end{bmatrix} \right) = \begin{bmatrix} x \ p(x, y, y) & w \\ p(x, y, y) \ p(x, y, z) \end{bmatrix} \in \Delta_\beta^\alpha,
\]
we see that from \( p(x, y, y) [\beta, \beta] x, [\beta, \beta] \leq [\alpha, \beta] \), and the term condition for \( [\alpha, \beta] \) we have \( w [\alpha, \beta] p(x, y, z) \).

For the other direction, if \( x \beta y \alpha z \) then from \( \alpha \geq \beta \) we have \( (x, y), (x, z) \in \alpha \), so we can apply the defining property (iii) of the difference term to congruences on \( \alpha \) to see the implication in the following picture.

Taking transposes, we have \( \begin{bmatrix} x & p(x, y, z) \\ y & z \end{bmatrix} \in \Delta_\alpha^\beta \) by Theorem A.33. By Theorems A.30 and A.33, if \( p(x, y, z) [\alpha, \beta] w \) then \( \begin{bmatrix} p(x, y, z) & w \\ z & z \end{bmatrix} \in \Delta_\alpha^\beta \), so \( \begin{bmatrix} x & w \\ y & z \end{bmatrix} \in \Delta_\alpha^\beta \) by the fact that \( \Delta_\beta^\alpha \) is transitively closed.

\( \square \)

Corollary A.42. If \( \alpha \geq \beta \) and \( [\alpha, \beta] = 0 \), then the restriction of the graph of \( p(x, y, z) \) to triples with \( x \beta y \alpha z \) is preserved by all polynomial operations of \( A \).

Using this, it’s possible to show that if \( A \) has center \( \zeta \), then we can write \( A \) as an extension of the quotient \( A/\zeta \) by the abelian algebra \( \zeta/\Delta_1^\zeta \) after making a choice of a section \( s : A/\zeta \to A \), with each \( n \)-ary basic operation \( f \) inducing a map \( t : (A/\zeta)^n \to \zeta/\Delta_1^\zeta \) so that the action of \( f \) on \( A \) can be decomposed as \( (x, y) \mapsto (f^\zeta/\Delta_1^\zeta(x) + t(y), f^\zeta/\Delta_1^\zeta(y)) \). If \( A \) is idempotent, then we can simplify this description slightly by noting that in this case, \( \zeta/\Delta_1^\zeta \) is isomorphic to any congruence class of \( \zeta \).

As a consequence of the decomposition of an algebra via its center, nilpotent algebras in congruence modular varieties turn out to be very well-behaved (e.g. they are always Mal’tsev and they have regular congruences), and after selecting an element to serve as the identity, one can define an associated nilpotent loop. See Chapter 7 of Freese and McKenzie [54] for details.

A.4 (Directed) Jónsson and Gumm terms

First we give Jónsson’s [69] characterization of congruence distributive varieties.

Definition A.43. A variety \( V \) is congruence distributive if for every \( A \in V \), \( \text{Con}(A) \) is a distributive lattice, that is, if the inequality
\[
\alpha \land (\beta \lor \gamma) \leq (\alpha \land \beta) \lor (\alpha \land \gamma)
\]
holds for all \( \alpha, \beta, \gamma \in \text{Con}(A) \).
The prototypical modular lattice which is not distributive is the lattice $\mathcal{M}_3$, as the next proposition shows.

**Proposition A.44** (Birkhoff [22]). In any modular lattice, if $a, b, c$ do not satisfy the distributive law, and if we define elements $d, e, f$ by

$$d = (b \land c) \lor (a \land (b \lor c)) = ((b \land c) \lor a) \land (b \lor c),$$

with $e, f$ defined by cyclic permutations of the variables $a, b, c$ in the above formula, then $d, e, f$ generate a sublattice isomorphic to the diamond lattice $\mathcal{M}_3$.

**Proof.** Using the modular law, we can check the formulas

$$d \land e = e \land f = f \land d = (a \land b) \lor (b \land c) \lor (c \land a)$$

and

$$d \lor e = e \lor f = f \lor d = (a \lor b) \land (b \lor c) \land (c \lor a).$$

If any two of $d, e, f$ are equal, then so are the two displayed expressions, and if we take the wedge of both with $a$ we get

$$a \land ((a \lor b) \land (b \lor c) \land (c \lor a)) = a \lor (b \lor c)$$

and (using the modular law again)

$$a \lor ((a \land b) \lor (b \land c) \lor (c \land a)) = (a \land b) \lor (a \land c).$$

\[\square\]

**Proposition A.45.** A variety is congruence distributive iff it is congruence modular and none of its algebras has a nontrivial abelian congruence. In particular, the commutator is given by $[\alpha, \beta] = \alpha \land \beta$ and $p(x, y, z) = z$ is a Gumm difference term in any congruence distributive variety.

**Proof.** If $\alpha$ is an abelian congruence, then $\ker \pi_1, \ker \pi_2, \Delta^0_\alpha \in \text{Con}(\alpha)$ generate a sublattice isomorphic to $\mathcal{M}_3$, with top element $\alpha \times \alpha |_\alpha$. The other direction follows from Proposition [10.32] since $\mathcal{M}_3$ does not satisfy the meet-semidistributive law $\text{SD}(\lor)$.

\[\square\]

**Example A.5.** The variety of unital rings is not congruence distributive, even though it is congruence modular (in fact, congruence permutable, since it has a Mal’cev term $x - y + z$) and contains no nontrivial abelian algebras (any such algebra would have $x \cdot y = 0$ for all $x, y$, and plugging in $y = 1$ would give $x = 0$ for all $x$). The reason for this is that the congruence on the ring $\mathbb{Z}/p^2$ corresponding to the ideal $(p)$ is abelian, but no congruence class of this ideal forms a unital subring of $\mathbb{Z}/p^2$.

**Theorem A.46** (Jónsson terms). A variety is congruence distributive iff it has ternary terms $q_0, ..., q_n$ satisfying the system of identities

$$q_0(x, y, z) \approx x,$$

$$q_i(x, y, x) \approx x \text{ for all } i,$$

$$q_i(x, y, y) \approx q_{i+1}(x, y, y) \text{ for } i \text{ odd},$$

$$q_i(x, x, y) \approx q_{i+1}(x, x, y) \text{ for } i \text{ even},$$

$$q_n(x, y, z) \approx z.$$
Proof. Consider the congruences $\theta_{x,y}, \theta_{y,z}, \theta_{x,z}$ corresponding to identifying pairs of variables on the free algebra $F(x, y, z)$ in a congruence distributive variety. From $(x, z) \in \theta_{x,z} \land (\theta_{x,y} \lor \theta_{y,z})$ and distributivity, we have

$$x (\theta_{x,z} \land \theta_{x,y}) \lor (\theta_{x,z} \land \theta_{y,z}) z.$$ 

Thus there exist $q_0, ..., q_n \in F(x, y, z)$ with $q_0 = x, q_i (\theta_{x,z} \land \theta_{x,y}) q_{i+1}$ for $i$ even, $q_i (\theta_{x,z} \land \theta_{y,z}) q_{i+1}$ for $i$ odd, and $q_n(x, y, z) = z$. In particular, we have $q_i \theta_{x,z} x$ for all $i$ by induction on $i$. Thus $q_0, ..., q_n$ satisfy the desired system of identities.

For the converse, suppose that $\alpha, \beta, \gamma$ are congruences on any algebra and that $(a, c) \in \alpha \land (\beta \lor \gamma)$. We need to show that $(a, c) \in (\alpha \land \beta) \lor (\alpha \land \gamma)$.

From $(a, c) \in \beta \lor \gamma$, there is a sequence $b_0, ..., b_m$ with $a = b_0, b_j \beta \lor \gamma b_{j+1}$ for all $j$, and $b_m = c$. Since $q_i(a, b_j, c) \alpha q_i(a, b_j, a) = a$ for all $i, j$, we then have

$$q_i(a, b_j, c) (\alpha \land \beta) \lor (\alpha \land \gamma) q_i(a, b_{j+1}, c)$$

for each $i, j$, so $q_i(a, a, c) (\alpha \land \beta) \lor (\alpha \land \gamma) q_i(a, c, c)$ for all $i$. Stringing these together with the identities relating $q_i$ to $q_{i+1}$, we see that $a = q_0(a, c, c) (\alpha \land \beta) \lor (\alpha \land \gamma) q_n(a, a, c) = c$. 

Example A.6. The variety of lattices is congruence distributive. For the Jónsson terms, we may take $n = 2$ and $q_1(x, y, z)$ to be the majority term $(x \land y) \lor (y \land z) \lor (z \land x)$. More generally, any variety with a near-unanimity term is congruence distributive.

We now prove a permutability result which is directly related to the fact that every congruence modular variety has a sequence of ternary terms known as Gumm terms, which look like Jónsson terms “glued to” a Mal’cev term.

Theorem A.47. If $\alpha, \beta$ are any two congruences in a congruence modular variety, then

$$\alpha \circ \beta \subseteq \left[ a, a \right] \circ \beta \circ \alpha.$$ 

Proof. If $(a, c) \in \alpha \circ \beta$, then there is some $b$ with $a \alpha b \beta c$. Applying the Gumm difference term $p$, we have

$$a \left[ a, a \right] p(a, b, b) \beta p(a, b, c) \alpha p(b, b, c) = c.$$ 

Corollary A.48. If $\alpha, \beta, \gamma$ are congruences in a congruence modular variety, then

$$(\alpha \circ \beta) \land \gamma \subseteq ((\alpha \land \beta) \lor (\alpha \land \gamma)) \circ \beta \circ \alpha.$$ 

Proof. We have $(\alpha \circ \beta) \land \gamma = ((\alpha \land (\beta \lor \gamma)) \circ \beta) \land \gamma$, and

$$[\alpha \land (\beta \lor \gamma), \alpha \land (\beta \lor \gamma)] \leq [\alpha, \beta \lor \gamma] = [\alpha, \beta] \lor [\alpha, \gamma] \leq (\alpha \land \beta) \lor (\alpha \land \gamma).$$

Thus, by the previous theorem we have

$$(\alpha \circ \beta) \land \gamma \subseteq (\alpha \land (\beta \lor \gamma)) \circ \beta \subseteq ((\alpha \land \beta) \lor (\alpha \land \gamma)) \circ \beta \circ \alpha.$$ 

A very similar argument shows that

$$(\alpha \circ \beta) \land \gamma \subseteq ((\alpha \land \gamma) \lor (\beta \land \gamma)) \circ \beta \circ \alpha,$$

which we will use to prove the following result (the corollary above could also be used to prove it, but there is an extra step of reordering the variables if we do it that way). Note that this containment can be viewed as a combination of a distributivity result with a permutability result.
\textbf{Theorem A.49} (Gumm terms). A variety is congruence modular iff it has ternary terms \(q_0, \ldots, q_n, p\) satisfying the system of identities
\begin{align*}
q_0(x, y, z) &\approx x, \\
q_i(x, y, x) &\approx x \text{ for all } i, \\
q_i(x, y, y) &\approx q_{i+1}(x, y, y) \text{ for } i \text{ odd}, \\
q_i(x, x, y) &\approx q_{i+1}(x, x, y) \text{ for } i \text{ even}, \\
q_n(x, y, y) &\approx p(x, y, y), \\
p(x, x, y) &\approx y.
\end{align*}

Furthermore, a ternary term \(p\) is a Gumm difference term iff there exist terms \(q_0, \ldots, q_n\) satisfying the above system of identities.

\textbf{Proof.} Consider the congruences \(\theta_{x,y}, \theta_{y,z}, \theta_{x,z}\) corresponding to identifying pairs of variables on the free algebra \(F(x, y, z)\) in a congruence distributive variety. From \((x, z) \in \theta_{x,z} \land (\theta_{x,y} \lor \theta_{y,z})\) and
\[\theta_{x,z} \land (\theta_{x,y} \lor \theta_{y,z}) \leq (\theta_{x,z} \land \theta_{x,y}) \lor (\theta_{x,z} \land \theta_{y,z}),\]
which is proved as in the previous corollary, we see that for any Gumm difference term \(p\) we have
\[x (\theta_{x,z} \land \theta_{x,y}) \lor (\theta_{x,z} \land \theta_{y,z}) \approx p(x, z, z).\]

Thus there exist \(q_0, \ldots, q_n \in F(x, y, z)\) with \(q_0 = x, q_i (\theta_{x,z} \land \theta_{x,y}) \approx q_{i+1}\) for \(i\) even, \(q_i (\theta_{x,z} \land \theta_{y,z}) \approx q_{i+1}\) for \(i\) odd, and \(q_n(x, y, z) = p(x, z, z)\). Therefore \(q_0, \ldots, q_n, p\) satisfy the desired system of identities.

To see that Gumm terms imply congruence modularity, we just need to show that they imply the existence of Day terms. If we assume without loss of generality that \(n\) is odd and take
\begin{align*}
m_0(x, y, z, w) &= x, \\
m_{2i-1}(x, y, z, w) &= q_i(x, w, y) \text{ for } i \text{ even}, \\
m_{2i}(x, y, z, w) &= q_i(x, z, y) \text{ for } i \text{ even}, \\
m_{2i-1}(x, y, z, w) &= q_i(x, z, y) \text{ for } i \text{ odd}, \\
m_{2i}(x, y, z, w) &= q_i(x, w, y) \text{ for } i \text{ odd}, \\
m_{2n+1}(x, y, z, w) &= p(z, w, y), \\
m_{2n+2}(x, y, z, w) &= y,
\end{align*}
then we have \(m_i(x, x, z, z) \approx x\) for all \(i\), \(m_i(x, y, x, y) \approx m_{i+1}(x, y, x, y)\) for \(i\) even, and \(m_i(x, y, z, z) \approx m_{i+1}(x, y, z, z)\) for \(i\) odd, so \(m_0, \ldots, m_{2n+2}\) are Day terms.

To show that any such \(p\) is a Gumm difference term, we just need to show that if \((x, y) \in \theta\), then \(p(x, y, y) \approx [\theta, \theta] x\). We will show by induction that \(q_i(x, y, y) \approx [\theta, \theta] q_i(x, x, y)\) for all \(i\). For the inductive step, we just need to show that for all \(i\), we have \(q_i(x, y, y) \approx [\theta, \theta] q_i(x, x, y)\). This follows from the term condition for \([\theta, \theta]\):
\[q_i(x, y, y) = q_i(x, x, y) \implies q_i(x, y, y) \approx [\theta, \theta] q_i(x, x, y).\]

The need to constantly divide into cases for even vs. odd \(i\) can be eliminated by the main result of [70], which establishes the existence of directed Jónsson and Gumm terms. The idea behind the directed variants is that if we have idempotent ternary terms \(f, g\) which satisfy
\[f(x, y, y) \approx g(x, x, y),\]
184
then they can also be indirectly connected by a ternary term \( h \) which satisfies \( h(x, y, x) \approx x \) and joins \( f, g \) by \( f \theta_{y,z} \ h \theta_{x,y} g \), that is,

\[
\begin{align*}
  f(x, y, y) & \approx h(x, y, y), \\
  h(x, x, y) & \approx g(x, x, y).
\end{align*}
\]

In fact, we can just take \( h(x, y, z) = f(x, z, z) \): then we will have \( h(x, y, y) = h(x, x, y) = f(x, y, y) = g(x, x, y) \), and \( h(x, y, x) = f(x, x, x) = x \). The goal of the directed Jónsson and Gumm terms is to cut out the middleman \( h \), to obtain a substantially stronger system of identities.

Another reason to prefer the directed equations \( f_i(x, y, y) \approx f_{i+1}(x, x, y) \) is that they have a clearer connection to higher arity terms, especially near-unanimity terms. Suppose that \( \phi \) is an \( n \)-ary operation, and define terms \( f_i \) by

\[
  f_i(x, y, z) = \phi(x, ..., x, y, z, ..., z),
\]

where the lone \( y \) occurs in the \( i \)-th position from the right (so there are \( i - 1 \) zs). Then the \( f_i \) will automatically satisfy

\[
  f_i(x, y, y) = \phi(x, ..., x, y, y, ..., y) = f_{i+1}(x, x, y),
\]

and if \( \phi \) is idempotent they will satisfy \( f_1(x, x, y) \approx x \) and \( f_n(x, y, y) \approx y \). Finally, \( \phi \) will be a near-unanimity term iff each \( f_i \) satisfies \( f_i(x, y, x) \approx x \).

**Theorem A.50 (Directed Gumm terms [70]).** A variety is congruence modular iff it has ternary terms \( f_1, ..., f_m, p \) with

\[
\begin{align*}
  f_1(x, x, y) & \approx x, \\
  f_i(x, y, x) & \approx x \text{ for all } i, \\
  f_i(x, y, y) & \approx f_{i+1}(x, x, y) \text{ for all } i, \\
  f_m(x, y, y) & \approx p(x, y, y), \\
  p(x, x, y) & \approx y,
\end{align*}
\]

and if the variety is congruence distributive then we can take \( f_m(x, y, y) \approx y \) (directed Jónsson terms).

**Proof.** Assume without loss of generality that our variety is idempotent. Suppose that there are Gumm terms \( q_1, ..., q_{2k+1}, p_1 \) with

\[
\begin{align*}
  q_1(x, x, y) & \approx x, \\
  q_i(x, y, x) & \approx x \text{ for all } i, \\
  q_{2i-1}(x, y, y) & \approx q_{2i}(x, y, y) \text{ for all } i, \\
  q_{2i}(x, x, y) & \approx q_{2i+1}(x, x, y) \text{ for all } i, \\
  q_{2k+1}(x, y, y) & \approx p_1(x, y, y), \\
  p_1(x, x, y) & \approx y.
\end{align*}
\]

Let \( \mathcal{F} \) be the free algebra on \( x, y \). Let \( \rightsquigarrow \) be the transitive closure of the binary relation on \( \mathcal{F} \) generated by \( x \rightsquigarrow x, x \rightsquigarrow y, y \rightsquigarrow y \), so binary terms \( a(x, y), b(x, y) \) have \( a \rightsquigarrow b \) iff there is a sequence of ternary terms \( t_i \) with \( t_1(x, x, y) = a(x, y), t_i(x, y, y) = t_{i+1}(x, x, y), \) and \( t_n(x, y, y) = b(x, y) \).

185
Additionally, let $\to$ be the relation on $F$ with $a \to b$ if there is a sequence of ternary terms $t_i$ with $t_1(x, x, y) = a(x, y)$, $t_i(x, y, y) = t_{i+1}(x, x, y)$, $t_n(x, y, y) = b(x, y)$, and additionally $t_i(x, y, x) = x$ for all $i$. Then for any ternary term $q$ satisfying $q(x, y, x) = x$, we have

$$q(\rightarrow, \leadsto, \rightarrow) \subseteq \to .$$

For any binary term $a(x, y)$, we define $a^n(x, y)$ recursively by $a^0(x, y) = y, a^1(x, y) = a(x, y)$, and

$$a^{n+1}(x, y) = a(x, a^n(x, y))$$

for each $n$.

Setting $b_k(x, y) = q_{2k+1}(x, y, y) = p_1(x, y, y)$, our goal will be to prove that

$$\exists b \in F \ x \to b_k^{2k}(b(x, y), b_k^{2k-1}(x, y)).$$

It will then be easy to construct a ternary term $p$ with $p(x, y, y) = b_k^p(b, b_k^{2k-1})$ and $p(x, y, y) = y$, by recursively plugging $p_1$ into itself in a similar way to the way we constructed Mal’cev terms on solvable algebras.

**Claim 1:** If $a \leadsto b$ and $c(x, y) \to d(x, y)$, then $c(a, b) \to d(a, b)$.

**Proof of Claim 1:** We just have to check this in the case where $c \to d$ in one step. So suppose that $t(x, y) = c(x, y), t(x, y, y) = d(x, y), t(x, y, x) = x$. Then

$$[c(a, b), d(a, b)] = t ([a], [a], [b], [b]) \in t(\to, \leadsto, \to) \subseteq \to .$$

**Claim 1.5:** If $a \leadsto b$ and $c(x, y) \leftarrow d(x, y)$, then $c(b, a) \to d(b, a)$.

**Proof of Claim 1.5:** This follows from Claim 1 and the fact that $c(x, y) \leftarrow d(x, y) \iff c(y, x) \to d(y, x)$.

**Claim 2:** If $a \to b$, then $a^n \to b^n$ for every $n$.

**Proof of Claim 2:** Induct on $n$. For the inductive step, we have

$$a^{n+1}(x, y) = a(x, a^n(x, y)) \to b(x, a^n(x, y)) \to b(x, b^n(x, y)) = b^{n+1}(x, y),$$

where the first $\to$ follows from $x = a^n(x, x) \leadsto a^n(x, y)$ and Claim 1, while the second $\to$ follows from the fact that $\to$ is preserved by $b$ and the inductive hypothesis.

The sequence of Gumm terms $q_1, ..., q_{2k+1}$ gives us a $k$-fence:

$$x = a_0 \to b_0 \leftarrow a_1 \to b_1 \leftarrow a_2 \to \cdots \leftarrow a_k \to b_k,$$

where $a_i(x, y) = q_{2i+1}(x, x, y) = q_{2i}(x, y), b_1(x, y) = q_{2i+1}(x, y, y) = q_{2i+2}(x, y, y)$. Our strategy will be to use Claims 1 and 1.5 to iteratively reduce the length of the fence.

**Claim 3:** If $x \to b \leftarrow a \to c$ is a 1-fence, then $x \to b_k(b, c(b, c))$.

**Proof of Claim 3:** We define a sequence of terms $d_i$ by $d_0 = x$ and

$$d_{i+1} = b(d_i, a),$$

and define terms $e_i$ by

$$e_i = a(d_i, a).$$

We claim that for each $i$ we have

186
• $d_i \rightsquigarrow a$, $d_i \rightsquigarrow d_{i+1}$, $d_i \rightsquigarrow e_i$, $d_i \rightarrow b$,
• $e_i \rightarrow d_{i+1}$, $e_i \rightarrow e_{i+1}$, $e_i \rightarrow c(b,c)$.

$$x = d_0 \rightarrow d_1 \rightarrow d_2 \rightarrow \cdots \rightarrow e_0 \rightarrow e_1 \rightarrow e_2 \rightarrow c(b,c)$$

To see this, note first that $d_0 = x \rightsquigarrow a$, so by induction on $i$ we have $d_{i+1} = b(d_i, a) \rightsquigarrow b(a, a) = a$ for each $i$. So from $x \rightarrow b \leftarrow a$ we get $d_i \rightarrow b(d_i, a) \leftarrow a(d_i, a)$ by Claim 1, that is, $d_i \rightarrow d_{i+1} \leftarrow e_i$ for each $i$.

Then we have $e_i = a(d_i, a) \rightarrow a(d_{i+1}, a) = e_{i+1}$ for each $i$, and $d_i = a(d_i, d_i) \rightsquigarrow a(d_i, a) = e_i$ for each $i$. This finishes up all of the arrows other than the rightmost two in the picture.

For $d_i \rightarrow b$, note that $d_0 = x \rightarrow b$ by assumption, and $d_{i+1} = b(d_i, a) \rightarrow b(b, b) = b$ inductively. Finally, for each $i$ we have

$$e_i = a(d_i, a) \rightarrow c(d_i, a) \rightarrow c(b, c),$$

where the first arrow follows from Claim 1.

Now we can use all these arrows to see that

$$x = d_0 = a_0(d_0, e_0) \rightarrow b_0(d_0, e_0) \rightarrow b_0(d_1, e_0) \rightarrow a_1(d_1, e_0) \rightarrow a_1(d_1, e_1) \rightarrow b_1(d_1, e_1) \rightarrow \cdots,$$

where we have used Claim 1 and Claim 1.5 several times. Chaining these together, we get

$$x \rightarrow b_k(d_k, e_k) \rightarrow b_k(b(c(b,c))).$$

This completes the proof of Claim 3.

**Claim 4**: For each $i < k$, there is a $k-i$-fence

$$x \rightarrow b_{0,i} \leftarrow a_{1,i} \rightarrow b_{1,i} \leftarrow a_{2,i} \rightarrow \cdots \leftarrow a_{k-i,i} \rightarrow b_{k-i,i} = b_k^{i+1-1}.$$

**Proof of Claim 4**: We prove this by induction on $i$. The base case $i = 0$ comes from the Gumm terms. Suppose it is known for $i$, then by Claim 3 we have

$$x \rightarrow b_k(b_{0,i}, b_{1,i}(b_{0,i}, b_{1,i})),

and from $b_{0,i} \leftarrow x$ we have

$$b_k(b_{0,i}, b_{1,i}(b_{0,i}, b_{1,i})) \leftarrow b_k(x, b_{1,i}(x, b_{1,i})) = b_k(x, b_{1,i}^2).$$

By Claim 2, we have $b_{1,i}^2 \leftarrow a_{2,i}^2 \rightarrow b_{2,i}^2 \leftarrow \cdots$, so if we take

$$b_{0,i+1} = b_k(b_{0,i}, b_{1,i}(b_{0,i}, b_{1,i}))$$

and

$$a_{j,i+1} = b_k(x, a_{j+1,i}^2), \quad b_{j,i+1} = b_k(x, b_{j+1,i}^2),$$

we get

$$x \rightarrow b_{0,i+1} \leftarrow a_{1,i+1} \rightarrow b_{1,i+1} \leftarrow a_{2,i+1} \rightarrow \cdots \leftarrow a_{k-i-1,i+1} \rightarrow b_{k-i-1,i+1},$$

187
and
\[ b_{k-1,i+1} = b_k(x, b_{k-1,i}^2) = b_k(x, (b_{k-1,i}^{2i+1-1})^2) = b_k^{2i+2-1}. \]
This completes the proof of Claim 4.

By Claim 4 applied with \( i = k - 1 \), we get a 1-fence
\[ x \to b_{0,k-1} \leftrightarrow a_{1,k-1} \to b_{1,k-1} = b_k^{2k-1}. \]
Applying Claim 3, we get
\[ x \to b_k(b_{0,k-1}, b_k^{2k-1}(b_{0,k-1}, b_k^{2k-1})) = b_k^{2k}(b_{0,k-1}, b_k^{2k-1}). \]
Letting \( b = b_{0,k-1} \), we see that we have succeeded in showing that \( x \to b_k^{2k}(b, b_k^{2k-1}) \). Thus there exist ternary terms \( f_i \) with
\[
\begin{align*}
  f_1(x, x, y) &\approx x, \\
  f_i(x, y, x) &\approx x \text{ for all } i, \\
  f_i(x, y, y) &\approx f_{i+1}(x, x, y) \text{ for all } i, \\
  f_m(x, y, y) &\approx b_k^{2k}(b(x, y), b_k^{2k-1}(x, y)).
\end{align*}
\]
Note that if \( b_k(x, y) = y \), then we also have \( b_k^{2k}(b(x, y), b_k^{2k-1}(x, y)) = y \), so the above becomes a sequence of directed Jónsson terms.

To finish, we just need to construct \( p \) with \( p(x, y, y) = b_k^{2k}(b(x, y), b_k^{2k-1}(x, y)) \) and \( p(x, y, y) = y \). Recall that \( p_1 \) satisfied \( p_1(x, y, y) = b_k(x, y) \) and \( p_1(x, y, y) = y \). We construct terms \( p_i \) inductively. For \( 2 \leq i + 1 < 2^k \), we set
\[ p_{i+1}(x, y, z) = p_1(x, p_i(x, y, y), p_i(x, y, z)), \]
and for \( 2^k \leq i + 1 \), we set
\[ p_{i+1}(x, y, z) = p_1(b(x, y), p_i(x, y, y), p_i(x, y, z)), \]
and finally we set \( p(x, y, z) = p_{2k+1-1}(x, y, z) \).

A.5 Subdirectly irreducible algebras, ultraproducts, and residually small varieties

In this subsection, we go over the proof of an extension of Jónsson’s Lemma [69], which shows that subdirectly irreducible algebras in a finitely generated congruence distributive variety have bounded size, to the congruence modular case. The key technical tool is the concept of an ultraproduct, and the fact that any ultrapower of a finite algebra \( A \) is isomorphic to \( A \).

Before we discuss ultraproducts, we first review some basic results about subdirect representations of algebras due to Birkhoff [23]. The following result is elementary.

**Proposition A.51.** If \( A \leq_{sd} \prod_{i \in I} A_i \) is a subdirect product, then \( \bigwedge_{i \in I} \ker \pi_i = 0_A \). In particular, if no \( \pi_i \) is an isomorphism then the congruence \( 0_A \) can be written as a meet of some family of nontrivial congruences.

Conversely, if \( 0_A \) can be written as a meet of congruences \( \alpha_i \in \text{Con}(A) \) for \( i \in I \), then \( A \leq_{sd} \prod_{i \in I} A_i \).
Definition A.52. An algebraic structure $\mathbb{A}$ is **subdirectly irreducible** if every way of writing $\mathbb{A}$ as a subdirect product $\mathbb{A} \leq_{sd} \prod_{i \in I} \mathbb{A}_i$ has at least one coordinate $i$ such that the projection map $\pi_i : \mathbb{A} \to \mathbb{A}_i$ is an isomorphism. The least nontrivial congruence on a subdirectly irreducible algebra is called its **monolith**.

The preceding proposition can now be rephrased as saying that $\mathbb{A}$ is subdirectly irreducible iff $0_\mathbb{A}$ is **meet-irreducible**.

Definition A.53. An element $\alpha$ of a complete lattice $\mathcal{L}$ is **meet-irreducible** if for any set of elements $\alpha_i \in \mathcal{L}$ with $\bigwedge_{i \in I} \alpha_i = \alpha$, some $\alpha_i$ is equal to $\alpha$. In this case, we define the **cover** of $\alpha$, written $\alpha^*$, to be the least element of $\mathcal{L}$ with $\alpha < \alpha^*$.

In particular, the monolith of a subdirectly irreducible algebra is the cover $0^*_\mathbb{A}$ of $0_\mathbb{A}$.

Theorem A.54 (Birkhoff’s Subdirect Representation Theorem). Any algebraic structure $\mathbb{A}$ can be represented as a subdirect product of subdirectly irreducible algebras.

Proof. For any $a \neq b \in \mathbb{A}$, Zorn’s Lemma implies that there is a maximal congruence $\theta'_{a,b}$ such that $(a,b) \notin \theta'_{a,b}$. Any such $\theta'_{a,b}$ is necessarily meet-irreducible, since any congruence which properly contains $\theta'_{a,b}$ necessarily contains $(a,b)$, and therefore contains the congruence generated by $\theta'_{a,b}$ and the pair $(a,b)$.

Since we clearly have $0_\mathbb{A} = \bigwedge_{a \neq b} \theta'_{a,b}$, we have the subdirect representation $\mathbb{A} \leq_{sd} \prod_{a \neq b} \mathbb{A}/\theta'_{a,b}$. □

Birkhoff’s subdirect representation theorem has a purely lattice-theoretic generalization to algebraic lattices.

Definition A.55. An element $\alpha$ of a complete lattice is called **compact** if for any family $\alpha_i$ such that $\alpha \leq \bigvee_{i \in I} \alpha_i$, there is some finite subset $\{i_1, \ldots, i_k\} \subseteq I$ such that $\alpha \leq \alpha_{i_1} \vee \cdots \vee \alpha_{i_k}$. A complete lattice is called **algebraic** if every element can be written as a join of compact elements.

Every congruence lattice $\text{Con}(\mathbb{A})$ is an algebraic lattice, since for any $a,b \in \mathbb{A}$ the congruence $\theta_{a,b}$ generated by $(a,b)$ is compact, and every congruence is a join of such congruences.

Proposition A.56. Let $\mathcal{L}$ be an algebraic lattice. Then every element $\alpha$ of $\mathcal{L}$ can be written as a meet of some family of meet-irreducible elements of $\mathcal{L}$.

Proof. Let $\theta$ be any compact element of $\mathcal{L}$ with $\alpha \not\geq \theta$. By Zorn’s Lemma and the compactness of $\theta$, there is some $\theta' \geq \alpha$ which is maximal such that $\theta' \not\geq \theta$, and this $\theta'$ is necessarily meet-irreducible with cover $\theta' \vee \theta$. Then $\bigwedge_{\theta \leq \alpha} \theta'$ is $\geq \alpha$, and is not $\geq$ any compact element $\theta$ with $\alpha \not\geq \theta$, so it must be equal to $\alpha$. □

Corollary A.57. If $\alpha < \beta$ in an algebraic lattice, then there is a meet-irreducible $\gamma$ such that $\gamma \geq \alpha$ but $\gamma \not\geq \beta$.

Now we can briefly discuss ultrafilters and ultraproducts before moving on to the main result of this subsection.

Definition A.58. If $I$ is a set, then a collection of subsets $\mathcal{U} \subseteq \mathcal{P}(I)$ is a **filter** if $\mathcal{U}$ does not contain $\emptyset$, $U,V \in \mathcal{U} \implies U \cap V \in \mathcal{U}$, and $U \subseteq V, U \in \mathcal{U} \implies V \in \mathcal{U}$. We say that $\mathcal{U}$ is an **ultrafilter** if additionally for every $U \subseteq I$, one of $U, I \setminus U$ is in $\mathcal{U}$. 189
Proposition A.59. Any filter is contained in an ultrafilter.

Proof. We apply Zorn’s Lemma to see that any filter is contained in a maximal filter. To finish, we just need to show that any maximal filter is an ultrafilter. Suppose that \( U, I \setminus U \not\in \mathcal{U} \), and let \( \mathcal{U}' \) be the collection of \( V \subseteq I \) such that \( V \cup U \in \mathcal{U} \). Then \( \mathcal{U}' \) is a filter which strictly contains \( \mathcal{U} \). 

Definition A.60. If \( A_i \) is a collection of structures which share a common signature \( \sigma \) and are indexed by \( i \in I \), and if \( \mathcal{U} \) is an ultrafilter on \( I \), then we define the ultraproduct \( \prod_i A_i / \mathcal{U} \) to be the quotient of \( \prod_i A_i \) by the congruence defined by

\[
a \equiv_U b \iff \{i \mid a_i = b_i\} \in \mathcal{U}.
\]

That \( \equiv_U \) is compatible with functions \( f \in \sigma \) follows from the fact that \( \mathcal{U} \) is a filter. If \( R \in \sigma \) is an \( m \)-ary relation, then \( R \) is interpreted on \( \prod_i A_i / \mathcal{U} \) by

\[
R(a^1 / \mathcal{U}, ..., a^m / \mathcal{U}) \iff \{i \mid R(a^1_i, ..., a^m_i)\} \in \mathcal{U}.
\]

If all the \( A_i \) are isomorphic to \( A \), then we call \( A_i / \mathcal{U} \) an ultrapower of \( A \).

Note that in terms of the congruence lattice \( \text{Con}(\prod_i A_i) \), the congruence \( \equiv_U \) is equal to the join

\[
\bigvee_{U \in \mathcal{U}} \ker \pi_U,
\]

where \( \pi_U : \prod_{i \in I} A_i \to \prod_{i \in U} A_i \) is projection onto the coordinates in \( U \). That this join is equal to the union \( \bigcup_{U \in \mathcal{U}} \ker \pi_U \) follows from the fact that \( \mathcal{U} \) is a filter.

Proposition A.61. If \( \mathcal{U} \) is an ultrafilter on \( I \) and \( U_1, ..., U_k \) partition \( I \) into \( k \) disjoint parts, then exactly one of \( U_1, ..., U_k \) is in \( \mathcal{U} \).

Corollary A.62. If \( |A_i| \leq n \) for all \( i \in I \), then \( |\prod_i A_i / \mathcal{U}| \leq n \) as well. If each \( A_i \) is finite and only finitely many isomorphism classes occur among the \( A_i \), then \( \prod_i A_i / \mathcal{U} \) is isomorphic to some \( A_i \).

In fact, much more is true about ultraproducts, and the corollary above also follows from the following result from model theory.

Theorem A.63 (Loś’s Theorem). Let \( \varphi(x_1, ..., x_n) \) be any first order formula in the signature \( \sigma \) with parameters \( x_1, ..., x_n \), then for any \( a^1, ..., a^n \in \prod_{i \in I} A_i \) and any ultrafilter \( \mathcal{U} \) on \( I \), we have

\[
\prod_i A_i / \mathcal{U} \models \varphi(a^1 / \mathcal{U}, ..., a^n / \mathcal{U}) \iff \{i \mid A_i \models \varphi(a^1_i, ..., a^n_i)\} \in \mathcal{U}.
\]

Proof. If \( \varphi \) is atomic, then this follows directly from the definitions. Otherwise, \( \varphi \) can be built up from atomic formulas via \( \lnot, \land, \exists \), and we can induct on the structure of \( \varphi \): for \( \lnot \), we use the ultrafilter property that exactly one of \( U, I \setminus U \) is in \( \mathcal{U} \) for each \( U \), for \( \land \) we use the filter property that intersections of sets in \( \mathcal{U} \) are in \( \mathcal{U} \), and for \( \exists \) we just need the fact that supersets of sets in \( \mathcal{U} \) are in \( \mathcal{U} \). 

190
Now for the main result. We extend Birkhoff’s $H, S, P$ notation by the operation $P_\alpha$, where $P_\alpha(\{A_i\})$ is the collection of ultraproducts of the $A_i$s. Recall that for $\beta$ a congruence, the centralizer $(0 : \beta)$ of $\beta$ is defined as the largest $\alpha$ such that $[\alpha, \beta] = 0$, and more generally $(\delta : \beta)$ is defined as the largest $\alpha$ such that $[\alpha, \beta] \leq \delta$.

**Theorem A.64.** Let $\{A_i\}$ be a family of algebras, and let $V = V(\{A_i\})$ be the variety they generate. If $V$ is congruence modular, $B \in V$ is subdirectly irreducible, and $\alpha = (0_B : 0_B)$ is the centralizer of the monolith $0_B^*$ of $B$, then $B/\alpha$ is a homomorphic image of a subalgebra of an ultraproduct of the $A_i$s, that is, $B/\alpha \in HSP_u(\{A_i\})$.

**Proof.** (From [54], where a stronger statement is proved.) By Birkhoff’s HSP Theorem, we can write $B = C/\theta$ for $C \leq \prod_i A_i$. Then $B$ will be subdirectly irreducible iff $\theta$ is meet-irreducible in $\text{Con}(C)$, so $\theta$ will have a cover $\theta^*$. The preimage $\varphi$ of $\alpha$ under $C \to B$ is the largest congruence on $C$ such that $[\varphi, \theta^*] \leq \theta$ (i.e. $\varphi = (\theta : \theta^*)$), and we have $B/\alpha = C/\varphi$.

The main step of the proof is the following claim: if $\beta \wedge \gamma \leq \theta$ but $\gamma \not\leq \theta$, then $\beta \leq \varphi$.

**Proof of claim:** We have

$$[\beta, \theta^*] \leq [\beta, \gamma \vee \theta] = [\beta, \gamma] \vee [\beta, \theta] \leq (\beta \wedge \gamma) \vee \theta = \theta,$$

so $\beta \leq \varphi$ by $\varphi = (\theta : \theta^*)$.

Using the claim, we can now argue as follows: let $F$ be a maximal filter such that $U \in F$ implies $\ker \pi_U \leq \theta$, and let $U$ be any ultrafilter which extends $F$. Then for any $U \in U$, we were unable to adjoin its complement to $F$, so there is some $V \in F$ such that $\ker \pi_V \setminus U \leq \theta$. Then

$$\ker \pi_U \wedge \ker \pi_{V \setminus U} = \ker \pi_U \cup V \leq \ker \pi_V \leq \theta,$$

so by the claim we have $\ker \pi_U \leq \varphi$. Thus the congruence $\bigvee_{U \in U} \ker \pi_U$ corresponding to $U$ is also $\leq \varphi$, and we see that $B/\alpha = C/\varphi$ is a quotient of $C/U \leq \prod_i A_i/U$. □

**Corollary A.65** (Jónsson’s Lemma [69]). Let $\{A_i\}$ be a family of algebras, and let $V = V(\{A_i\})$ be the variety they generate. If $V$ is congruence distributive and $B \in V$ is subdirectly irreducible, then $B \in HSP_u(\{A_i\})$. In particular, if $\{A_i\}$ is a finite set of finite algebras, then $B \in HS(\{A_i\})$.

**Corollary A.66.** For any two finite subdirectly irreducible algebras $A, B$ with the same signature which generate congruence distributive varieties, we have $A \cong B$ iff the set of identities that hold in $A$ is the same as the set of identities that hold in $B$.

**Example A.7.** Consider the variety of distributive lattices, and the two-element lattice $\{0, 1\}$, max, min). It is easy to see that every identity that holds in the two-element lattice is implied by the lattice axioms together with distributivity (since these allow us to put every term into conjunctive normal form), so the variety of distributive lattices is generated by the two-element lattice.

By Jónsson’s Lemma, the only subdirectly irreducible distributive lattice is the two-element lattice itself, so we see that in fact every distributive lattice is a sublattice of $\{0, 1\}^I$ for some index set $I$, that is, every distributive lattice is a sublattice of the lattice of subsets of some set $I$.

In order to understand subdirectly irreducible algebras in congruence modular varieties, we need to combine the above results with an understanding of subdirectly irreducible modules over rings.

191
Proposition A.67. Let $G, M$ be abelian groups and let $R$ be a finite subgroup of $\text{Hom}(G, M)$, such that there is a nonzero element $a \in M$ so that for all $x \in G$ there is an $r \in R$ with $rx = a$. Then $|G|$ is a prime power dividing $|R|$.

Proof. First we show that $G$ is finite, following [54]. Let $r_1, ..., r_k$ be the nontrivial elements of $R$.

We will show by induction on $k$ that $|G| \leq (k + 1)!$. For the base case, if $k = 0$ then $G$ can have no nonzero elements, so $|G| = 1 = (k + 1)!$. For the inductive step, note that by the pigeonhole principle there is some $r_i$ such that at least $\frac{|G| - 1}{k}$ elements are mapped to $a$ by $r_i$, so $|\ker r_i| \geq \frac{|G| - 1}{k}$ (this is the ordinary group theoretic kernel), and every nonzero element of $\ker r_i$ can be mapped to $a$ by some $r_j$ with $j \neq i$, so $|\ker r_i| \leq k!$ by the induction hypothesis. Thus $|G| \leq 1 + k \cdot k! \leq (k + 1)!$.

Now that we know that $G$ is finite, we know that every element of $G$ has finite order, so some element $x$ has order $p$ for some prime $p$. Then there is some $r \in R$ with $rx = a$, so $a$ must also have order $p$. From this argument, we see that every element of $G$ must have order a power of $p$, so $|G|$ is also a power of $p$.

We may assume without loss of generality that $M$ is generated by the image of $G$ under all elements of $R$, so in particular that $M$ is finite. Then there exists an element $\pi \in \hat{M} = \text{Hom}(M, Q/Z)$ such that $\pi(a) \neq 0$.

Define a linear map $\phi : R \to \hat{G} = \text{Hom}(G, Q/Z)$ by $\phi : r \mapsto \phi_r$, where $\phi_r$ is the linear map $\phi_r : x \mapsto \pi(rx)$. Then $\phi$ must be surjective, or else the image will be a proper subgroup of $\hat{G}$ and so there will be some nonzero $x \in G$ with $\phi_r(x) = 0$ for all $r \in R$, which implies $rx \neq a$ for all $r$. Thus $|G| = |\hat{G}|$ divides $|R|$.

Corollary A.68. Let $R$ be a finite ring, and let $M$ be a subdirectly irreducible module over $R$. Then $|M|$ is a prime power dividing $|R|$.

Proof. If $M$ is subdirectly irreducible, then it has a least nontrivial submodule $N$, which is generated by some nonzero element $a \in N$. Then for each nonzero $x \in M$ we have $N \leq Rx$, so there is some $r \in R$ with $rx = a$. Thus we can apply the previous proposition with $G = M$.

Now we can use this result to bound the sizes of subdirectly irreducible algebras in congruence modular varieties in the special case where the centralizer of the monolith is abelian.

Theorem A.69. Suppose that $B \in V$ is subdirectly irreducible, and $V$ is locally finite and congruence modular. If $\alpha \in \text{Con}(B)$ is abelian and $|B/\alpha| = k$, then every congruence class of $\alpha$ has size a prime power bounded by $|\text{F}_V(k + 1)|$.

Proof. (Adapted from [90].) Assume $\alpha$ is nontrivial, so $0^*_\alpha \leq \alpha$. Let $p$ be a Gumm difference term. By Corollary A.42 the restriction of the graph of $p$ to the blocks of $\alpha$ is preserved by every operation of $B$. Choose elements $0 \neq a$ with $(0, a) \in 0^*_\alpha$, and note that $0, a$ are in the same congruence block of $\alpha$.

Pick constants $c_0, ..., c_{k-1}$ with $c_0 = 0$ such that each congruence class of $\alpha$ contains some $c_i$. We will treat each congruence class $c_i/\alpha$ of $\alpha$ as an abelian group with zero element $c_i$, addition given by $x + y = p(x, c_i, y)$, and subtraction given by $x - y = p(x, y, c_i)$.

Suppose that $x \neq y$ with $(x, y) \in \alpha$. Then since $0^*_\alpha$ is the least nontrivial congruence, the pair $(0, a)$ must be in the congruence generated by $(x, y)$, so there must be a chain of unary polynomials $f_i$ such that $0 = f_0(x), f_i(y) = f_{i+1}(x)$, and $f_m(y) = a$. Note that this implies that $f_i(x), f_i(y)$ are all in the congruence class $0/\alpha$. Thus, it makes sense to define a unary polynomial $f$ such that $f(z) = f_0(z) +_0 f_1(z) -_0 f_1(x) +_0 \cdots +_0 f_m(z) -_0 f_m(x)$
for $z$ in the congruence class $x/\alpha$. One explicit way to construct such an $f$ is given by

$$f(z) = p(p(\cdots(p(f_0(z), f_1(x), f_1(z)), f_2(z)), \cdots), f_m(x), f_m(z)).$$

It’s easy to check that we have $f(x) = 0$ and $f(y) = a$. Since $f$ preserves the graph of $p$ restricted to congruence classes of $\alpha$, if $x, y \in c_i/\alpha$ then we have $f(x - y) - f(c_i) = a$, and the unary polynomial $z \mapsto f(z) - f(c_i)$ defines a linear map in $\text{Hom}(c_i/\alpha, c_0/\alpha)$.

To finish, we just need to bound the size of the subgroup $R_{i,0}$ of linear maps in $\text{Hom}(c_i/\alpha, c_0/\alpha)$ which can be defined by unary polynomials $f$. Suppose that $f(z) = t(z, b_1, \ldots, b_m)$ for some $t$ and constants $b_1, \ldots, b_m \in B$, such that $f(c_i) = c_0$. For each $b_i$, we choose $j_i$ such that $b_i \in c_{j_i}/\alpha$. Define a unary polynomial $f'$ by

$$f'(z) = t(z, c_{j_1}, \ldots, c_{j_m}) - 0(t(c_{i_1}, c_{j_1}, \ldots, c_{j_m}).$$

Then for $z \in c_i/\alpha$, we have $f'(z) \in c_0/\alpha$, and since $t$ preserves the graph of $p$ restricted to congruence classes of $\alpha$, we have $f'(z) = f(z)$ for $z \in c_i/\alpha$ (alternatively, we could prove this by the term condition for $[\alpha, \alpha] = 0_B$). Thus every element of $\text{Hom}(c_i/\alpha, c_0/\alpha)$ which can be defined by a unary polynomial can also be defined by a polynomial $f'$ which has the form $f'(z) = t'(z, c_0, \ldots, c_{k-1})$ for some $k + 1$-ary term $t'$, so

$$|R_{i,0}| \leq |F_V(k + 1)|.$$ 

Applying the previous proposition, we see that $|c_i/\alpha|$ is a prime power dividing $|R_{i,0}|$. \qed

**Corollary A.70.** If $|A| = m$ is finite and $V(A)$ is congruence modular, and if $B \in V(A)$ is subdirectly irreducible with $(0_B : 0_B^*)$ abelian, then $|B| \leq m \cdot m^{m+1}$.

**Definition A.71.** A variety $V$ is called residually small if there is a cardinal $\kappa$ such that every subdirectly irreducible algebra $B \in V$ has $|B| < \kappa$, and residually finite if every subdirectly irreducible algebra in $V$ is finite. An algebra $A$ is called residually small if the variety $V(A)$ generated by $A$ is residually small.

First we show that if a locally finite variety contains an infinite subdirectly irreducible algebra, then it contains infinitely many distinct finite subdirectly irreducible algebras.

**Theorem A.72.** If $B$ is subdirectly irreducible, then $B$ is a subalgebra of an ultraproduct of a family of finitely generated subdirectly irreducible algebras in $HS(B)$.

**Proof.** (From [54].) Let the monolith $0_B^*$ of $B$ be generated (as a congruence) by the pair $(a, b)$. Let $I$ be the family of finitely generated subalgebras $S \leq B$ with $a, b \in S$, and for each $S \in I$, pick a congruence $\alpha_S$ on $S$ which is maximal among all congruences which do not contain $(a, b)$. Then each $S/\alpha_S$ is subdirectly irreducible, since every congruence which properly contains $\alpha_S$ contains $(a, b)$.

Let $U$ be an ultrafilter on $I$ such that the set $U_S = \{S \mid S \subseteq S\}$ is in $U$ for every finite $S \subseteq B$. Such an ultrafilter exists since for any $S_1, S_2$ we have $U_{S_1 \cap S_2} = U_{S_1 \cup S_2}$, and for $S$ finite $U_S$ is nonempty since it contains $S_{(S \cup \{a, b\})} = S_{(S \cup \{a, b\})} = S_{(S \cup \{a, b\})}$. Define a map $\varphi : B \rightarrow (\prod_{S \in I} S/\alpha_S)/U$ as the ultraproduct of the family of maps $\varphi_S$ given by $\varphi_S(x) = x/\alpha_S$ for $x \in S$ and $\varphi_S(x) = a/\alpha_S$ for $x \notin S$. Then for $x_1, \ldots, x_k \in B$ and $t$ a $k$-ary term of $B$, we have

$$\{S \mid \varphi_S(t(x_1, \ldots, x_k)) = t(\varphi_S(x_1), \ldots, \varphi_S(x_k))\} \in U,$$

since it contains $U_{\{x_1, \ldots, x_k\}}$. Thus $\varphi$ is a homomorphism. To see that it is injective, just note that $\varphi_S(a) \neq \varphi_S(b)$ for all $S \in I$. \qed

193
Corollary A.73. If a locally finite variety contains an infinite subdirectly irreducible algebra, then it contains arbitrarily large finite subdirectly irreducible algebras.

It turns out that finite residually small algebras can be understood in terms of a commutator condition. We say that an algebra $A$ satisfies a commutator identity hereditarily if every congruence lattice of every subalgebra of $A$ satisfies the identity.

Proposition A.74. The commutator identity $[\alpha \land \beta, \beta] = \alpha \land [\beta, \beta]$ is equivalent to the implication $\alpha \leq [\beta, \beta] \implies [\alpha, \beta] = \alpha$.

Proof. The implication clearly follows from the identity. For the other direction, we apply the implication to $\alpha \land [\beta, \beta] \leq [\beta, \beta]$ to see that

$$[\alpha, \beta] = [\alpha \land [\beta, \beta], \beta] \leq [\alpha \land \beta, \beta] \leq \alpha \land [\beta, \beta].$$

Proposition A.75. If $\mathbb{A}$ is in a congruence modular variety and satisfies the commutator identity $[\alpha \land \beta, \beta] = \alpha \land [\beta, \beta]$ hereditarily, then so does every quotient $\mathbb{B}$ of $\mathbb{A}$.

Proof. Suppose $\mathbb{B} = \mathbb{A}/\gamma$ and $\alpha, \beta \in \text{Con}(\mathbb{A})$ with $\alpha, \beta \geq \gamma$. We need to check that if $\alpha \leq [\beta, \beta]\gamma$, then $\alpha = [\alpha, \beta]\gamma$. By the modular law, if $\alpha \leq [\beta, \beta] \lor \gamma$ then

$$\alpha = \alpha \land ([\beta, \beta] \lor \gamma) = (\alpha \land [\beta, \beta]) \lor \gamma = [\alpha, \beta] \lor \gamma = [\alpha, \beta]_{\gamma}. \quad \square$$

Proposition A.76. If $\mathbb{A}_1, \mathbb{A}_2$ are in a congruence modular variety and satisfy the commutator identity $[\alpha \land \beta, \beta] = \alpha \land [\beta, \beta]$ hereditarily, then so does their product $\mathbb{A}_1 \times \mathbb{A}_2$.

Proof. Let $\mathbb{B} \leq \mathbb{A}_1 \times \mathbb{A}_2$, we can assume without loss of generality that this inclusion is subdirect by replacing the $A_i$ with $\pi_i(\mathbb{B})$. Suppose $\alpha, \beta \in \text{Con}(\mathbb{B})$ with $\alpha \leq [\beta, \beta]$, we will show that $[\alpha, \beta] = \alpha$. We have

$$\alpha \lor \ker \pi_1 \leq [\beta \lor \ker \pi_1, \beta \lor \ker \pi_1]_{\ker \pi_1},$$

so from the assumption on $\mathbb{A}_1$ we get

$$\alpha \lor \ker \pi_1 = [\alpha \lor \ker \pi_1, \beta \lor \ker \pi_1]_{\ker \pi_1} = [\alpha, \beta] \lor \ker \pi_1.$$

Thus by the modular law and $[\alpha, \beta] \leq \alpha$, we have

$$\alpha = \alpha \land (\ker \pi_1 \lor [\alpha, \beta]) = (\alpha \land \ker \pi_1) \lor [\alpha, \beta].$$

Similarly, we have $\alpha = (\alpha \land \ker \pi_2) \lor [\alpha, \beta]$. Since $\alpha \land \ker \pi_2 \leq \alpha \leq [\beta, \beta]$, we may apply the same reasoning to $\alpha \land \ker \pi_2$ to see that

$$\alpha \land \ker \pi_2 = (\alpha \land \ker \pi_2 \land \ker \pi_1) \lor [\alpha \land \ker \pi_2, \beta],$$

so $\alpha \land \ker \pi_2 \leq [\alpha, \beta]$, so

$$\alpha = (\alpha \land \ker \pi_2) \lor [\alpha, \beta] = [\alpha, \beta]. \quad \square$$

Theorem A.77. If $|\mathbb{A}| = m$ is finite and $\mathcal{V}(\mathbb{A})$ is congruence modular, and if $\mathbb{A}$ satisfies the commutator identity $[\alpha \land \beta, \beta] = \alpha \land [\beta, \beta]$ hereditarily, then every subdirectly irreducible algebra $\mathbb{B} \in \mathcal{V}(\mathbb{A})$ has $|\mathbb{B}| \leq m \cdot m^{m+1}$. 

194
Proof. By Corollary A.73 we just need to check the bound in the case where \( B \) is finite. In this case, we have \( B \in \text{HSP}_{\text{fin}}(\mathcal{A}) \), so \( B \) satisfies the commutator identity \([\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]\) by the previous propositions. Let \( 0_B^* \) be the monolith of \( B \), and let \( \alpha = (0_B : 0_B^*) \) be its centralizer.

We claim that \( \alpha \) is abelian. To see this, note that from \([\alpha, 0_B^*] = 0_B\) we have
\[
0_B = [0_B^* \wedge \alpha, \alpha] = 0_B^* \wedge [\alpha, \alpha],
\]
so \([\alpha, \alpha] = 0_B\). Now we can apply Corollary A.70 to see that \(|B| \leq m \cdot m^{m^{m+1}}\). \( \square \)

Example A.8. The symmetric group \( S_3 \) on three letters is residually small, since it satisfies the commutator identity \([\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]\) hereditarily: the only interesting case to check is that \([A_3, S_3] = A_3\), where \( A_3 \) is the alternating group on three letters. We have \( HS(S_3) = \{1, \mathbb{Z}/2, \mathbb{Z}/3, S_3\} \), and all three nontrivial elements are subdirectly irreducible.

The general theory shows that every subdirectly irreducible \( G \in \mathcal{V}(S_3) \) has an abelian normal subgroup \( N \) with \( G/N \in HS(S_3) \), with \(|N| \) a prime power bounded by \(|F_{\mathcal{V}(S_3)}([G/N] + 1)| \leq 6^7\). Since \( N \in \mathcal{V}(S_3) \) and every element of \( S_3 \) has order dividing 6, \( N \) has exponent 2 or 3. From here it is not too hard to check that the nontrivial subdirectly irreducible algebras in \( \mathcal{V}(S_3) \) are \( \mathbb{Z}/2, \mathbb{Z}/3, S_3 \), and all three of these are subgroups of \( S_3 \). Thus every group in \( \mathcal{V}(S_3) \) is a subgroup of a power of \( S_3 \).

Proposition A.78. If \( \mathcal{A} \) is contained in a congruence modular variety but does not satisfy the commutator identity \([\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]\) hereditarily, then there is some subdirectly irreducible \( B \in HS(\mathcal{A}) \) such that the centralizer of the monolith of \( B \) is not abelian.

Proof. Suppose that \( \mathcal{A} \) fails to satisfy the commutator identity. In this case there must be \( \alpha, \beta \in \text{Con}(\mathcal{A}) \) with \( \alpha \leq [\beta, \beta] \) and \([\alpha, \beta] < \alpha\). Let \( \theta \) be a meet-irreducible congruence such that \( \theta \geq [\alpha, \beta] \) but \( \theta \not\supseteq \alpha \), and let \( \theta^* \) be its cover. Then
\[
\theta^* \leq \alpha \lor \theta \leq [\beta, \beta] \lor \theta \leq [\beta \lor \theta, \beta \lor \theta]_\theta
\]
and
\[
[\theta^*, \beta \lor \theta]_\theta \leq [\alpha \lor \theta, \beta \lor \theta]_\theta = [\alpha, \beta] \lor \theta = \theta,
\]
so if we take \( B \) to be \( \mathcal{A}/\theta \), then the monolith of \( B \) is \( \theta^*/\theta \), and \( \beta \lor \theta/\theta \) is contained in the centralizer of the monolith of \( B \) but is not abelian. \( \square \)

Theorem A.79. If \( \mathcal{A} \) is contained in a congruence modular variety but does not satisfy the commutator identity \([\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]\), then \( \mathcal{V}(\mathcal{A}) \) is not residually small. In fact, for every cardinal \( \kappa \), \( \mathcal{V}(\mathcal{A}) \) contains a subdirectly irreducible algebra whose congruence lattice has size at least \( \kappa \).

Proof. (From [54].) By the proposition, we can reduce to the case where \( \mathcal{A} \) is subdirectly irreducible and the centralizer \( \beta \) of the monolith \( 0^* \) is not abelian.

Consider \( \beta \) as a subalgebra of \( \mathcal{A}^2 \), and \( \Delta_\beta^0 \) as a congruence on \( \beta \). From \([\beta, 0^*_\mathcal{A}] = 0_\mathcal{A}\) we have \( \Delta_\beta^0 \wedge \ker \pi_1 = \Delta_\beta^0 \wedge \ker \pi_2 = 0_\beta \), and from the definition of \( \Delta_\beta^0 \) we have \( \Delta_\beta^0 \lor \ker \pi_i = \pi_i^{-1}(0^*) \). Set \( 0_i^* = \pi_i^{-1}(0^*) \) and \( \theta = 0_i^* \wedge 0_2^* \lor \theta_i = 0_i^* \lor \ker \pi_{(1,2)} \backslash \{i\} \), then (after several applications of the modular law - don’t worry about the details just yet) we have the following sublattice in \( \text{Con}(\beta) \).

\[ \]
In the picture, we see that $\Delta^\beta_0$ appears to be meet-irreducible in $\text{Con}(\beta)$, and the interval $[\Delta^\beta_0, \beta]$ contains the incomparable elements $0^*_1, 0^*_2$. If $\Delta^\beta_0$ isn’t meet-irreducible, we can still try to find a meet-irreducible congruence $\lambda$ on $\beta$ which is above $\Delta^\beta_0$ but not above $\theta$, and then $\beta/\lambda$ should give us a subdirectly irreducible algebra whose congruence lattice contains two distinct elements coming from $\ker \pi_1 \vee \lambda$ and $\ker \pi_2 \vee \lambda$ (that neither of these is equal to $\beta\beta$ will come from the assumption that $\beta$ is not abelian). This is the basic idea behind the general construction, but we will need to scale up by considering higher dimensional analogues of $\beta \leq \mathbb{A}^2$.

Let $\kappa$ be any cardinal, considered as the set of all ordinals below $\kappa$. Define $\mathbb{B} \leq \mathbb{A}^\kappa$ by

$$\mathbb{B} = \{ a \in \mathbb{A}^\kappa \mid a_i \equiv a_j \forall i, j \in \kappa \}.$$ 

Then $\mathbb{B}$ has a natural map to $\mathbb{A}/\beta$, and we call the kernel of this map $\beta_\mathbb{B}$. Inside $\text{Con}(\mathbb{B})$, we have $\ker \pi_i \vee \ker \pi_j = \beta_\mathbb{B}$ for all $i \neq j \in \kappa$. The strategy is to construct a congruence $\lambda$ on $\mathbb{B}$ such that $\mathbb{B}/\lambda$ is subdirectly irreducible and $\ker \pi_i \vee \lambda \not\geq \beta_\mathbb{B}$ for all $i$, which will guarantee that the congruences $\ker \pi_i \vee \lambda/\lambda \in \text{Con}(\mathbb{B}/\lambda)$ are pairwise distinct. The congruence $\lambda$ will be constructed by first constructing congruences $\Delta, \theta$ with $\Delta < \theta$ and $\theta \vee \ker \pi_i \not\geq \beta_\mathbb{B}$.

We need a congruence on $\mathbb{B}$ generalizing $\Delta^\beta_0$ on $\beta$. We define $\Delta_i$ by

$$ (a, b) \in \Delta_i \iff \begin{bmatrix} a_0 & b_0 \\ a_i & b_i \end{bmatrix} \in \Delta^\beta_0 \land (a_j = b_j \forall j \neq 0, i), $$

and define $\Delta$ by

$$ \Delta = \bigvee_{0 < i < \kappa} \Delta_i.$$ 

We also define congruences $\theta_i$ by

$$ (a, b) \in \theta_i \iff \begin{bmatrix} a_i & b_i \end{bmatrix} \in 0^*_\kappa \land (a_j = b_j \forall j \neq i), $$

and define $\theta$ by

$$ \theta = \bigvee_{i < \kappa} \theta_i.$$ 

We need to check some basic properties of these congruences, to see that they behave as in the picture of $\text{Con}(\beta)$. First, we check that $\theta_0 \leq \theta_i \vee \Delta_i$ for all $i$. Letting $\pi_{i'}$ be the projection onto all coordinates other than $i$, then it’s easy to check that $\theta_0 \leq \ker \pi_{i'} \vee \Delta$ by reasoning about just the two coordinates $0, i$ and keeping all other coordinates fixed:

$$ \begin{bmatrix} a_0 \\ a_i \end{bmatrix} \ker \pi_{i'} \begin{bmatrix} a_0 \\ a_0 \end{bmatrix} \Delta_i \begin{bmatrix} b_0 \\ b_0 \end{bmatrix} \ker \pi_{i'} \begin{bmatrix} b_0 \\ b_i \end{bmatrix}. $$

196
Then by the modular law, if we let $0_i^* = \pi_i^{-1}(0\_\lambda)$ and note that $\Delta_i \leq 0_i^*$, we get

$$\theta_0 = \theta_0 \wedge 0_i^* \leq (\Delta_i \lor \ker \pi_i') \wedge 0_i^* = \Delta_i \lor (\ker \pi_i \wedge 0_i^*) = \Delta_i \lor \theta_i.$$  

Similarly, we get $\theta_i \leq \theta_0 \lor \Delta_i$ for all $i$.

Next, for each $i$ we have $\Delta \lor \theta_i = \theta$: for each $j \in \kappa$, we have

$$\Delta \lor \theta_i = \Delta \lor \Delta_i \lor \theta_i \geq \Delta \lor \theta_0 \geq \Delta_i \lor \theta_0 \geq \theta_j,$$

so $\Delta \lor \theta_i \geq \bigvee_{j \in \kappa} \theta_j = \theta$, while the other containment follows from $\Delta_i \leq \theta_0 \lor \theta_i$ for all $i$.

We now check that $\Delta \neq \theta$. It's enough to check that $\theta_0 \not\leq \Delta$, since $\Delta \lor \theta_0 = \theta$. Note first that $\theta_0$ is compact, since $0\_\lambda$ is compact. Thus we just need to check that $\theta_0 \not\leq \bigvee_{j \leq n} \Delta_i$ for all $i_1, \ldots, i_n$.

In fact, we can assume that $i_1, \ldots, i_n$ are 1, ..., $n$ by a symmetry argument.

We will show by induction on $n$ that $\theta_0 \wedge (\Delta_1 \lor \cdots \lor \Delta_n) = 0\_\beta$ for all $n$. The base case follows from the fact that $[\beta, 0\_\lambda] = 0\_\beta \implies \ker \pi_2 \lor \Delta_0^* = 0\_\beta$ in $\con{\beta}$, which in turn implies $\theta_0 \lor \Delta_1 = 0\_\beta$. For the inductive step, we argue as follows:

$$\theta_0 \wedge (\Delta_1 \lor \cdots \lor \Delta_n) = \theta_0 \wedge ((\theta_0 \lor \theta_n) \wedge (\Delta_1 \lor \cdots \lor \Delta_n))$$

$$= \theta_0 \wedge ((\theta_0 \lor \theta_n) \lor (\Delta_1 \lor \cdots \lor \Delta_{n-1}) \lor \Delta_n)$$

$$= \theta_0 \wedge ((\theta_0 \lor \Delta_1 \lor \cdots \lor \Delta_{n-1}) \lor \Delta_n)$$

$$= \theta_0 \wedge \Delta_n = 0\_\beta,$$

where the second equality used the modular law and the fact that $\Delta_n \leq \theta_0 \lor \theta_n$, the third equality used the fact that $\theta_n$ is independent of everything that happens on the coordinates 0, ..., $n-1$, and the last two equalities used the inductive hypothesis.

We have shown that $\Delta \lt \theta$. We can now apply Corollary A.57 to see that there is some meet-irreducible congruence $\lambda$ with $\lambda \geq \Delta$ but $\lambda \not\geq \theta$. To finish, we just need to check that $\lambda \lor \ker \pi_i \not\geq \beta\_\beta$. To see this, note that $\lambda \not\geq \theta_i$, since otherwise we would have $\lambda \geq \Delta \lor \theta_i = \theta$, a contradiction. Since $\theta_i$ is the minimal nonzero element of the interval $[0\_\beta, \ker \pi_i']$, this means that $\lambda \lor \ker \pi_i' = 0\_\beta$. Thus if (for contradiction) $\lambda \lor \ker \pi_i \geq \beta\_\beta$, then we would have

$$[\beta\_\beta, \beta\_\beta] \leq \ker \pi_i' \lor \ker \pi_i \lor \ker \pi_i = (\lambda \lor \ker \pi_i') \lor \ker \pi_i = \ker \pi_i,$$

and applying $\pi_i$ we would get $[\beta, \beta] = 0\_\lambda$, a contradiction to the assumption that $\beta$ was not abelian.

Putting it all together, we have a meet-irreducible congruence $\lambda$ such that $\lambda \lor \ker \pi_i \not\geq \beta\_\beta$ for each $i$, but $\ker \pi_i \lor \ker \pi_j \geq \beta\_\beta$ for all $i \neq j$. Thus $\beta/\lambda$ is subdirectly irreducible, and the congruences $\ker \pi_i \lor \lambda/\lambda$ are mutually distinct elements of $\con{\beta/\lambda}$.

**Corollary A.80.** Let $\mathcal{V}$ be a finitely generated congruence modular variety. Then the following are equivalent:

- $\mathcal{V}$ is residually small,
- every algebra in $\mathcal{V}$ satisfies the commutator identity $[\alpha \land \beta, \beta] = \alpha \land [\beta, \beta]$,
- $\mathcal{V}$ is generated by a finite algebra $\mathcal{A}$ such that for every subdirectly irreducible $\beta \in HS(\mathcal{A})$, the centralizer of the monolith of $\beta$ is abelian.
• $\mathcal{V}$ is generated by a finite algebra which satisfies the commutator identity $[\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]$ hereditarily,

• $\mathcal{V}$ has a finite bound on the size of its subdirectly irreducible elements.

Corollary A.81. If $\mathbb{A}$ is in a congruence modular variety and has size $|\mathbb{A}| \leq 3$, then $\mathcal{V}(\mathbb{A})$ is residually small.

Proof. Suppose for contradiction that $\mathbb{A}$ is subdirectly irreducible with a monolith $0^*_{\mathbb{A}}$ whose centralizer $(0_{\mathbb{A}} : 0^*_{\mathbb{A}})$ is not abelian. Then since $\text{Con}(\mathbb{A})$ has height at most 2, we necessarily have

$$0_{\mathbb{A}} < 0^*_{\mathbb{A}} < (0_{\mathbb{A}} : 0^*_{\mathbb{A}}) = 1_{\mathbb{A}}.$$ 

Thus $|\mathbb{A}| = 3$, and we may name the elements of $\mathbb{A}$ as $a, b, c$, such that $0^*_{\mathbb{A}}$ corresponds to the partition $\{a, b\}, \{c\}$ of $\mathbb{A}$. Letting $p(x, y, z)$ be a Gumm difference term, we see from Theorem A.41 that

$$\begin{bmatrix} a & p(a, b, c) \\ b & c \end{bmatrix} \in \Delta^1_{0^*_{\mathbb{A}}},$$

Modulo $0^*_{\mathbb{A}}$, we have $p(a, b, c) \equiv_{0^*_{\mathbb{A}}} p(a, a, c) = c$, so we must have $p(a, b, c) = c$. Then by Theorem A.30 we have $(a, b) \in [1_{\mathbb{A}}, 0^*_{\mathbb{A}}]$, which contradicts $(0_{\mathbb{A}} : 0^*_{\mathbb{A}}) = 1_{\mathbb{A}}$. \hfill \Box

Proposition A.82. If $\mathbb{A}$ satisfies the commutator identity $[\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]$, then every nilpotent congruence on $\mathbb{A}$ is abelian.

Proof. The commutator identity implies that

$$[[\alpha, \alpha], \alpha] = [[\alpha, \alpha] \wedge \alpha, \alpha] = [\alpha, \alpha] \wedge [\alpha, \alpha] = [\alpha, \alpha].$$ \hfill \Box

Proposition A.83 (Ol’šanskii [93]). If all the Sylow subgroups of a finite group $\mathbb{G}$ are abelian, then the center $Z(\mathbb{G})$ and the commutator subgroup $[\mathbb{G}, \mathbb{G}]$ intersect trivially, that is, $Z(\mathbb{G}) \wedge [\mathbb{G}, \mathbb{G}] = 0_{\mathbb{G}}$.

Proof. Fix a Sylow subgroup $\mathbb{S}$ of $\mathbb{G}$, and consider the transfer map $\mathbb{G} \to \mathbb{S}/[\mathbb{S}, \mathbb{S}]$. Recall that the transfer homomorphism from a finite group to the abelianization of a subgroup is defined by making a choice of coset representatives $x_i$ with $\mathbb{G} = \bigcup_i x_i \mathbb{S}$, and sending $g \in \mathbb{G}$ to $\prod_i x_i g x_i^{-1}$, where for each $i$, $s_i \in \mathbb{S}$ is given by $g x_i = x_j s_i$ for some $j$. Since $\mathbb{S}$ is assumed to be abelian, this gives us a homomorphism from $\mathbb{G}$ to $\mathbb{S}$.

Now consider any $g \in Z(\mathbb{G}) \cap \mathbb{S}$. The transfer homomorphism sends $g$ to $\prod_i g = g^{[\mathbb{G}, \mathbb{S}]}$ since $g x_i = x_i g$ for each $i$, and if $g \neq 1$ then $g^{[\mathbb{G}, \mathbb{S}]} \neq 1$ as well since $[\mathbb{G} : \mathbb{S}]$ is relatively prime to the order of $g$. Thus there is a map from $\mathbb{G}$ to an abelian group such that $g$ is not in the kernel, so $g \not\in [\mathbb{G}, \mathbb{G}]$. Since every nontrivial element of $Z(\mathbb{G}) \wedge [\mathbb{G}, \mathbb{G}]$ has a power which has prime order and is therefore contained in a Sylow subgroup of $\mathbb{G}$, we must have $Z(\mathbb{G}) \wedge [\mathbb{G}, \mathbb{G}] = 0_{\mathbb{G}}$ to avoid a contradiction. \hfill \Box

Corollary A.84 (Ol’šanskii [93]). A finite group is residually small iff all of its Sylow subgroups are abelian.

Proof. By Proposition A.82 all nilpotent subgroups of a finite residually small group must be abelian, so in particular the Sylow subgroups must be abelian since all $p$-groups are nilpotent.

For the other direction, note that for any $\mathbb{B} \in HS(\mathbb{A})$, the Sylow subgroups of $\mathbb{B}$ are quotients of subgroups of the Sylow subgroups of $\mathbb{A}$ by the Sylow theorems. Thus we just have to check that if the Sylow subgroups of a subdirectly irreducible group are abelian, then the centralizer $\mathbb{C}$ of its monolith $0^*$ is abelian.

Note that if $\mathbb{C}$ centralizes $0^*$, then $0^* \leq Z(\mathbb{C})$. By Proposition A.83 we have $Z(\mathbb{C}) \wedge [\mathbb{C}, \mathbb{C}] = 0$, so $0^* \wedge [\mathbb{C}, \mathbb{C}] = 0$, which implies that $[\mathbb{C}, \mathbb{C}] = 0$. \hfill \Box
A.5.1 Similarity

Even if a finitely generated congruence modular variety is not residually small, we can still classify its subdirectly irreducible algebras by using the concept of similarity from Freese and McKenzie [54]. We will use a different definition of similarity than their definition, but which they prove to be equivalent.

**Definition A.85.** We say that subdirectly irreducible algebras $A, B$ in a congruence modular variety $V$ are similar if there exists an algebra $C \in V$ with congruences $\alpha, \beta, \gamma, \delta \in C$ such that $C/\alpha \cong A$, $C/\beta \cong B$, and

$$[[\alpha, \alpha^*] \setminus [\gamma, \delta]] \nearrow [\beta, \beta^*].$$

If furthermore $C \leq_{sd} A \times B$ and $\alpha, \beta$ are the kernels of the projections to $A, B$, then we say that $C$ is the graph of a similarity from $A$ to $B$.

**Proposition A.86.** If $A, B$ are similar, then there is a witnessing algebra $C \leq_{sd} A \times B$ which is the graph of a similarity from $A$ to $B$. If $\alpha, \beta$ are the kernels of the projections to $A, B$, then $(\alpha : \alpha^*) = (\beta : \beta^*)$ and $C/(\alpha : \alpha^*)$ is the graph of an isomorphism

$$A/(0_A : 0_A^*) \rightarrow B/(0_B : 0_B^*).$$

If $A, B$ are similar but not isomorphic, then they must both have abelian monoliths.

**Proof.** For the first statement, let $C \in V$ and $\alpha, \beta, \gamma, \delta \in \text{Con}(C)$ be as in the definition of similarity. It’s enough to show that we have

$$[[\alpha, \alpha^*] \setminus [\alpha \land \beta, (\alpha \land \beta) \lor \delta]] \nearrow [\beta, \beta^*],$$

since then we can replace $C$ by $C/(\alpha \land \beta)$, which is a subdirect product of $C/\alpha \cong A$ and $C/\beta \cong B$.

We have

$$\alpha \lor ((\alpha \land \beta) \lor \delta) = \alpha \lor \delta = \alpha^*, $$

and by the modular law and the fact that $\gamma \leq \alpha \land \beta$, we have

$$\alpha \land ((\alpha \land \beta) \lor \delta) = (\alpha \lor \delta) \land (\alpha \land \beta) = \gamma \lor (\alpha \land \beta) = (\alpha \land \beta),$$

so $[[\alpha, \alpha^*] \setminus [\alpha \land \beta, (\alpha \land \beta) \lor \delta]]$, and the other perspectivity follows by a symmetric argument.

The remaining statements follow from the Diamond Isomorphism Theorem A.27: if $[[\alpha, \alpha^*] \setminus [\gamma, \delta]] \nearrow [\beta, \beta^*]$, then $(\alpha : \alpha^*) = (\gamma : \delta) = (\beta : \beta^*)$, so

$$A/(0_A : 0_A^*) \cong C/(\alpha : \alpha^*) = C/(\beta : \beta^*) \cong B/(0_B : 0_B^*),$$

and

$$[\alpha^*, \alpha^*]_\alpha = \alpha \iff [\delta, \delta]_\gamma = \gamma \iff [\beta^*, \beta^*]_\beta = \beta,$$

so $0_A^*$ is abelian iff $0_B^*$ is abelian, and if neither is abelian then $\alpha = (\alpha : \alpha^*) = (\beta : \beta^*) = \beta$ and $A \cong C/\alpha = C/\beta \cong B$. \qed

**Proposition A.87.** If $A, B$ are similar such that $\sigma$ is the corresponding isomorphism

$$\sigma : A/(0_A : 0_A^*) \rightarrow B/(0_B : 0_B^*),$$

then they are similar via the algebra $R = \{(x, y) \in A \times B \mid \sigma((0_A : 0_A^*)) = y/(0_B : 0_B^*)\}$. 199
Proof. Suppose \( \mathcal{C} \subseteq \mathbb{R} \) is the graph of a similarity from \( \mathcal{A} \) to \( \mathcal{B} \), with

\[
[\ker \pi_1, (\ker \pi_1)^*] \setminus \left[ 0_{\mathcal{C}}, \delta \right] \cup \left[ \ker \pi_2, (\ker \pi_2)^* \right]
\]

in \( \text{Con}(\mathcal{C}) \). We may assume that \( \mathcal{A}, \mathcal{B} \) have abelian monoliths, so \( [\delta, \delta] = 0_{\mathcal{C}} \) by the Diamond Isomorphism Theorem [A.27]. Then by Theorem [A.47], \( \delta \) permutes with all congruences in \( \text{Con}(\mathcal{C}) \), so in particular \( (\ker \pi_1)^* = \delta \circ \ker \pi_1 \). In other words, for any \((a, b) \in \mathcal{C}\) and any \(a' \in a/0_A^*\), there exists a \(b'\) such that

\[
\begin{bmatrix} a' \\ b' \end{bmatrix} = \delta \begin{bmatrix} a \\ b \end{bmatrix}.
\]

In fact, this \(b'\) is uniquely determined by \(a, b, a'\), since \(\delta \cap \ker \pi_1 = 0_{\mathcal{C}}\). Additionally, we must have \(b' \in b/0_B^*\), since \(\delta \leq (\ker \pi_2)^*\).

Now we can extend \(\delta\) to a congruence \(\delta_{\mathcal{R}} \in \text{Con}(\mathbb{R})\) as follows. For \((a, b), (a', b') \in \mathbb{R}\) with \(a 0_A^* a'\) and \(b 0_B^* b'\), we pick any \((u, v) \in \mathbb{C}\) with \(u (0_A : 0_A^*) a\) and write

\[
\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta_{\mathcal{R}} \iff \begin{bmatrix} p(a', a') & u \\ p(b', b') & v \end{bmatrix} \in \delta,
\]

where \(p\) is a Gumm difference term. Note that by Corollary [A.42] this choice of \(\delta_{\mathcal{R}}\) is preserved by the operations of \(\mathcal{A}\) so long as it is well-defined. To check that this is in independent of the choice of \((u, v) \in \mathbb{C}\), suppose \((u', v') \in \mathbb{C}\) with \(u' (0_A : 0_A^*) a\), and apply Corollary [A.42] again to see that

\[
p \left( p \begin{bmatrix} p(a', a') & u \\ p(b', b') & v \end{bmatrix} \right) = \left( p \begin{bmatrix} p(a', a') & u' \\ p(b', b') & v' \end{bmatrix} \right).
\]

where we have used \(0_A^*, 0_B^*\) abelian to see that \(p(a', a, a) = a'\) and \(p(b', b, b) = b'\).

We need to check that \(\delta_{\mathcal{R}}\) is a congruence on \(\mathbb{R}\). It clearly contains the equality relation on \(\mathbb{R}\). For symmetry and transitivity, note that

\[
p \left( \begin{bmatrix} p(a', a') & u \\ p(b', b') & v \end{bmatrix} \right) = \left( \begin{bmatrix} p(a', a', u') & u' \\ p(b', b', v') \end{bmatrix} \right).
\]

Finally, we need to check that \(\delta_{\mathcal{R}} \cap \ker \pi_1 = 0_{\mathcal{R}}\) and \(\delta_{\mathcal{R}} \cap \ker \pi_1 = (\ker \pi_1)^*\). That \(\delta_{\mathcal{R}} \cap \ker \pi_1 = 0_{\mathcal{R}}\) follows from the fact that if we pick \(u\) such that \((u, b') \in \mathbb{C}\), then

\[
\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta_{\mathcal{R}} \iff \begin{bmatrix} p(a, a, u) & u \\ p(b, b', u') \end{bmatrix} \in \delta,
\]

and so this can only occur when \(b = b'\) since \(\delta \cap \ker \pi_1 = 0_{\mathcal{C}}\) (by assumption). That \(\delta_{\mathcal{R}} \cap \ker \pi_1 = (\ker \pi_1)^*\) follows from \(\delta \subseteq \delta_{\mathcal{R}} \subseteq (\ker \pi_1)^*\) and \(\delta \not\subseteq \ker \pi_1\).

\[\Box\]

Corollary A.88. A similarity from \(\mathcal{A}\) to \(\mathcal{B}\) can be described by the following data: an isomorphism

\[
\sigma : \mathcal{A}/(0_A : 0_A^*) \xrightarrow{\sim} \mathcal{B}/(0_B : 0_B^*)
\]

together with a congruence \(\delta \in \text{Con}(\mathbb{R})\), where \(\mathbb{R} = \{(x, y) \in \mathcal{A} \times \mathcal{B} \mid (x/0_A : 0_A^*) = y/0_B^* \}\), such that for every \((a, b) \in \mathbb{R}\) and every \(a' \in a/0_A^*\), there exists a unique \(b' \in b/0_B^*\) such that

\[
\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta.
\]

In particular, if \(\mathcal{A}, \mathcal{B}\) are idempotent, then for any \((a, b) \in \mathbb{R}\) the congruence classes \(a/0_A^*\) and \(b/0_B^*\) are isomorphic to each other.

200
Corollary A.89. Similarity is an equivalence relation on subdirectly irreducible algebras.

Proof. Suppose we have similarities from $A$ to $B$ and from $B$ to $C$, described by isomorphisms

$$A/(0_A : 0^*_A) \xrightarrow{\sim} B/(0_B : 0^*_B) \xrightarrow{\sim'} C/(0_C : 0^*_C)$$

and congruences $\delta, \delta'$. We define a congruence $\delta \circ \delta'$ by

$$\begin{bmatrix} a & a' \\ c & c' \end{bmatrix} \in \delta \circ \delta' \iff \exists (b, b') \in 0^*_B \left( \begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta \right) \land \left( \begin{bmatrix} b & b' \\ c & c' \end{bmatrix} \in \delta' \right).$$

We need to check that for each $a, c, a'$ there exists a unique $c'$ satisfying the above. Existence is easy: for each $b$, we can fill in a unique $b'$ to satisfy $\delta$, and then there is a unique $c'$ which satisfies $\delta'$. We just need to show that the choice of $b$ doesn’t affect the final $c'$ we get. Suppose that instead of $b$ we had picked $v$. Then the claim is that if we leave $a, a', c, c'$ unchanged and replace $b$ by $v$ and $b'$ by $p(b', b, v)$, we get another valid solution. For $\delta$, this follows from

$$p\left( \begin{bmatrix} a & a' \\ b & b' \end{bmatrix}, \begin{bmatrix} a & a \\ b & b \end{bmatrix}, \begin{bmatrix} a & a \\ v & v \end{bmatrix} \right) = \begin{bmatrix} a & a' \\ v & p(b', b, v) \end{bmatrix},$$

and it follows for $\delta'$ similarly.

We will show that every subdirectly irreducible algebra $A$ with abelian monolith is similar to a subdirectly irreducible algebra $D(A)$ such that the monolith of $D(A)$ is equal to its own centralizer. The size of the algebra $D(A)$ can then be bounded using Theorem A.64 and the following proposition.

Proposition A.90. If $B \in \mathcal{V}(A)$ is subdirectly irreducible, $A$ is finite, and $\mathcal{V}(A)$ is congruence modular, then every congruence class of $0^*_B$ has size at most $|A|$.

Proof. By Theorem A.72 and Corollary A.62, we may assume without loss of generality that $B$ is finite. By Theorem A.64, we may also assume that $0^*_B$ is abelian. Take $m$ minimal such that there exists $C \leq A^m$ and $\theta \in \text{Con}(C)$ with $B \cong C/\theta$, so $[\theta^*, \theta^*] \leq \theta$.

Let $\pi_{1'}$ be the projection onto all but the first coordinate, then by the minimality of $m$ we have $\ker \pi_{1'} \leq \theta$. Thus we have

$$[\theta, \theta^*] \setminus [\theta \land \ker \pi_{1'}, \theta^* \land \ker \pi_{1'}].$$

By Theorem A.47, the congruences $\theta$ and $\theta^* \land \ker \pi_{1'}$ permute. Thus for every congruence class $C^*$ of $\theta^*$ containing some $c \in C$, the size of $C^*/\theta$ is equal to the size of $C'/\theta$ where $C'$ is the congruence class of $\theta^* \land \ker \pi_{1'}$ containing $c$. But $|C'/\theta | \leq |C/\ker \pi_{1'}| = |A|$, so every congruence class of $0^*_B$ has size bounded by $|A|$.

Definition A.91. Suppose $A$ is a subdirectly irreducible algebra in a congruence modular variety. If $0^*_A$ is nonabelian, define $D(A)$ to be $A$. Otherwise, consider $0^*_A$ as a subalgebra of $A^2$ and $\Delta^{(0,0^*_A)}_{0^*_A}$ as a congruence on $0^*_A$, and define $D(A) = 0^*_A/\Delta^{(0,0^*_A)}_{0^*_A}$.

Recall that by Theorem A.41, if $0^*_A$ is abelian and $p$ is a Gumm difference term, then $(0_A : 0^*_A) \geq 0^*_A$ and $[(0_A : 0^*_A), 0^*_A] = 0_A$, so we have

$$\begin{bmatrix} x & w \\ y & z \end{bmatrix} \in \Delta^{(0,0^*_A)}_{0^*_A} \iff (p(x, y, z) = w) \land (x \equiv_{0^*_A} y \equiv_{(0,0^*_A)} z).$$
In this case, the subalgebra \( \{(x, x)/\Delta_{0_\Delta}^{(0,0^*)}\} \leq D(\mathbb{A}) \) meets every congruence class of \((0_\Delta : 0_\Delta^*)_{D(\mathbb{A})}\) (that is, the congruence \((0_\Delta : 0_\Delta^*)\) considered as a congruence on \(D(\mathbb{A})\)) exactly once, and is isomorphic to \(\mathbb{A}/(0_\Delta : 0_\Delta^*)\).

**Proposition A.92.** If \(\mathbb{A}\) is a subdirectly irreducible algebra in a congruence modular variety with an abelian monolith, then \(D(\mathbb{A})\) is subdirectly irreducible with monolith \((0_\Delta : 0_\Delta^*)_{D(\mathbb{A})}\), and \(\mathbb{A}, D(\mathbb{A})\) are similar via the algebra \(0_\Delta^*\) and the congruences \(\ker \pi_1, \Delta_{0_\Delta}^{(0,0^*)} \in \text{Con}(0_\Delta^*)\). Furthermore, the monolith \((0_\Delta : 0_\Delta^*)_{D(\mathbb{A})}\) of \(D(\mathbb{A})\) is its own centralizer.

**Proof.** Note that \(\ker \pi_1\) is covered by \(\ker \pi_1 \vee \ker \pi_2\), since \(\ker(\ker \pi_1 \vee \ker \pi_2) = 0_\Delta^*\). First we check that in \(\text{Con}(0_\Delta^*)\) we have the perspectivities

\[
[\ker \pi_1, \ker \pi_1 \vee \ker \pi_2] \sim [0_1^*, \ker \pi_2] \supset [\Delta_{0_\Delta}^{(0,0^*)}, (0_\Delta : 0_\Delta^*)_{0_\Delta^*}].
\]

The hardest step here is checking that \(\ker \pi_2 \vee \Delta_{0_\Delta}^{(0,0^*)} = (0_\Delta : 0_\Delta^*)_{0_\Delta^*}\) if \((x, y), (w, z) \in 0_\Delta^*\). First we check that \((0_\Delta : 0_\Delta^*)_{0_\Delta^*}\) is a subdirectly irreducible algebra in \(\text{Con}(0_\Delta^*)\).

We show that \(0_\Delta^*\) is the unique cover of \(\Delta_{0_\Delta}^{(0,0^*)}\) in \(\text{Con}(0_\Delta^*)\). Note first that \((0_\Delta : 0_\Delta^*_{0_\Delta^*})\) is a cover of \(\Delta_{0_\Delta}^{(0,0^*)}\), since the interval \([\Delta_{0_\Delta}^{(0,0^*)}, (0_\Delta : 0_\Delta^*)_{0_\Delta^*}]\) is isomorphic to \([\ker \pi_1, \ker \pi_1 \vee \ker \pi_2] \approx [0_\Delta, 0_\Delta^*]\) by the Diamond Isomorphism Theorem [A.27].

Suppose that \(\psi\) is any congruence in \(\text{Con}(0_\Delta^*)\) with \(\psi > \Delta_{0_\Delta}^{(0,0^*)}\). If \(\psi \geq \ker \pi_2\), then \(\psi \geq \Delta_{0_\Delta}^{(0,0^*)} \vee \ker \pi_2 = (0_\Delta : 0_\Delta^*)_{0_\Delta^*}\), and we are done. Otherwise, since \(\ker \pi_2\) is a cover of \(0_\Delta^*\), we must have \(\psi \wedge \ker \pi_2 = 0_\Delta^*\). Then we have

\[
[\psi \wedge \ker \pi_1, \ker \pi_2 \vee \ker \pi_1] \mid \ker \pi_1 \leq [\psi, \ker \pi_2] \vee \ker \pi_1 \leq (\psi \wedge \ker \pi_2) \vee \ker \pi_1 = \ker \pi_1.
\]

Applying \(\pi_1\) to both sides, we see that \(\pi_1(\psi \wedge \ker \pi_1) \leq (0_\Delta : 0_\Delta^*)\), so \(\psi \wedge \ker \pi_1 \leq (0_\Delta : 0_\Delta^*)_{0_\Delta^*}\). Thus \(\psi \in [\Delta_{0_\Delta}^{(0,0^*)}, (0_\Delta : 0_\Delta^*)_{0_\Delta^*}]\), so again we must have \(\psi = (0_\Delta : 0_\Delta^*)_{0_\Delta^*}\). We have finished showing that \(D(\mathbb{A})\) is subdirectly irreducible.

To see that the monolith \((0_\Delta : 0_\Delta^*)_{D(\mathbb{A})}\) of \(D(\mathbb{A})\) is its own centralizer, note that by the Diamond Isomorphism Theorem [A.27] we have

\[
(\Delta_{0_\Delta}^{(0,0^*)} : (0_\Delta : 0_\Delta^*)_{0_\Delta^*}) = (\ker \pi_1 : \ker \pi_1 \vee \ker \pi_2) = \pi_1^{-1}((0_\Delta : 0_\Delta^*)) = (0_\Delta : 0_\Delta^*)_{0_\Delta^*}.
\]

**Proposition A.93.** If \(\mathbb{A}, \mathbb{B}\) are subdirectly irreducible algebras in a congruence modular variety, then \(\mathbb{A}\) is similar to \(\mathbb{B}\) iff \(D(\mathbb{A}) \cong D(\mathbb{B})\).

**Proof.** Since similarity is an equivalence relation, we may as well replace \(\mathbb{A}, \mathbb{B}\) by \(D(\mathbb{A}), D(\mathbb{B})\). Thus we just need to prove that if \(\mathbb{A}, \mathbb{B}\) have monoliths equal to their own centralizers, and have
subalgebras \( X_A, X_B \) which intersect their monoliths transversely, then they are similar iff they are isomorphic.

Let \( \sigma : A/0_A^* \to B/0_B^* \) be the isomorphism and \( \delta \in \text{Con}(R) \), where \( R = \{(x, y) \in A \times B \mid \sigma(x/(0_A^* : 0_A^*)) = y/(0_B^* : 0_B^*)\} \), be the data describing a similarity from \( A \) to \( B \). Then \( \sigma \) induces an isomorphism \( \sigma_X : X_A \to X_B \), and the graph of \( \sigma_X \) is a subalgebra of \( R \). Let \( S \) be the subalgebra of \( (a, b) \in R \) such that \((a, b)\) is congruent to some element of \( \sigma_X \) modulo \( \delta \). Then \( S \) must be the graph of an isomorphism from \( A \) to \( B \).

\[ \text{Theorem A.94. If } B \in \mathcal{V}(A) \text{ is subdirectly irreducible, } A \text{ is finite, and } \mathcal{V}(A) \text{ is congruence modular, then } B \text{ is similar to a subdirectly irreducible algebra in } HS(A). \]

\[ \text{Proof. We may as well replace } B \text{ by } D(B), \text{ so assume without loss of generality that the monolith of } B \text{ is either nonabelian or equal to its own centralizer. If the monolith of } B \text{ is nonabelian, then } B \in HS(A) \text{ by Theorem A.64 so we just need to handle the case where } 0_B^* = (0_B : 0_B^*). \text{ In this case, Theorem A.64 implies that } B/0_B^* \in HS(A), \text{ so by Proposition A.90 we have } |B| \leq |A|^2 < \infty. \]

Since \( B \) is finite, we can write \( B = R/\theta \) for some \( R \leq A^n \) and \( \theta \in \text{Con}(R) \). Then we can write \( R \) as a subdirect product \( R \leq_{sd} A_1 \times \cdots \times A_m \) of finitely many subdirectly irreducible algebras \( A_i \in HS(A_i) \). We assume that the \( A_i \) are chosen such that none of them can be replaced by a subdirect product of some number of proper quotients of \( A_i \) while still keeping the isomorphism \( R/\theta \cong B \).

Then for any \( i \), we must have \( \theta \wedge \ker \pi_{[m]\{i\}} = 0_R; \) if not, we could replace \( A_i \) with a subdirect representation of \( R/\ker \pi_i \vee (\theta \wedge \ker \pi_{[m]\{i\}}) \), since by the modular law we have

\[ \ker \pi_{[m]\{i\}} \wedge (\ker \pi_i \vee (\theta \wedge \ker \pi_{[m]\{i\}})) = (\ker \pi_{[m]\{i\}} \wedge \ker \pi_i) \vee (\theta \wedge \ker \pi_{[m]\{i\}}). \]

Since \( \ker \pi_{[m]\{i\}} \neq 0_R \), we have \( \theta \vee \ker \pi_{[m]\{i\}} \geq \theta^* \), so \( \theta^* \wedge \ker \pi_{[m]\{i\}} \) is a cover of \( 0_R \), and we have

\[ \theta^* \wedge \ker \pi_{[m]\{i\}} \nsubseteq [0_R, \theta^* \wedge \ker \pi_{[m]\{i\}}]. \]

so \( B = R/\theta \) is similar to \( A_i = R/\ker \pi_i \).

\[ \text{Example A.9. Let’s work out what } D(G) \text{ is when } G \text{ is a subdirectly irreducible group. Let } M \lhd G \text{ be the normal subgroup corresponding to the monolith } 0_G^* \text{, and let } N = C_G(M) \lhd G \text{ be the normal subgroup corresponding to the centralizer } (0_G : 0_G^*). \text{ First off, what is the group structure on the congruence } 0_G^*? \]

By definition, we have

\[ 0_G^* = \{(x, y) \in G^2 \mid x^{-1}y \in M\}. \]

We have a natural exact sequence of groups

\[ 0 \to M \hookrightarrow 0_G^* \to G \to 0, \]

where the inclusion is the map \( m \mapsto (1, m) \) and the quotient map is the first projection \( \pi_1 \). The quotient \( 0_G^* \to G \) has a section \( \Delta : G \leftarrow 0_G^* \) given by \( g \mapsto (g, g) \). Thus we can write \( 0_G^* \) as a semidirect product

\[ 0_G^* \cong M \rtimes G, \]

where the action of \( G \) on \( M \) is the standard conjugation action.
How about the congruence $\Delta^{(0:0^r)}_{G'G} \in \Con(0^r_G)$? By Theorem A.41, we have

$$\begin{bmatrix} x & w \\ y & z \end{bmatrix} \in \Delta^{(0:0^r)}_{G'G} \iff (xy^{-1}z = w) \land (x \equiv_M y \equiv_N z).$$

Since this is a congruence on a group, we just need to understand the congruence class of the identity, so we plug in $x = y = 1$ and ask what values $(w, z)$ can take. We find that $\Delta^{(0:0^r)}_{G'G}$ corresponds to the normal subgroup $\{ (n, n) \mid n \in \mathbb{N} \}$, so under the isomorphism $0^r_G \cong M \times G$ it corresponds to $\mathbb{N}$, considered as a subgroup of $G$. Thus we have

$$D(G) = 0^r_G/\Delta^{(0:0^r)}_{G'G} \cong (M \rtimes G)/\mathbb{N} \cong M \rtimes (G/\mathbb{N}).$$

That any of this makes sense follows from $N = C_G(M)$. We see that $M$ is the normal subgroup corresponding to the monolith of $D(G)$, that $M$ is equal to its own centralizer in $D(G)$, and that the natural map $G/\mathbb{N} \hookrightarrow D(G)$ has image transverse to the monolith, and induces an isomorphism

$$\sigma : G/\mathbb{N} \xrightarrow{\sim} D(G)/M.$$ 

To complete the description of the similarity from $G$ to $D(G)$, we let $\mathbb{R}$ be the fiber product of $G$ and $D(G)$ over $G/\mathbb{N}$, and define the congruence $\delta \in \Con(\mathbb{R})$ as the 4-ary relation

$$\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta \iff \begin{bmatrix} a \\ b \end{bmatrix} \begin{bmatrix} a' \\ b' \end{bmatrix} \in \mathbb{R} \land a^{-1}a' = b^{-1}b' \in M.$$ 

That $\delta$ is closed under multiplication must be checked - it follows from the fact that $\mathbb{N}$ centralizes $M$, and the fact that for any $a, b, a', b'$ satisfying the above conditions all of $a, b, a', b'$ must necessarily map to the same element of $G/\mathbb{N}$.

What are the possible values for $D(G)$, assuming the monolith is abelian? Note that if we consider $M$ as a module via the $G/\mathbb{N}$ action, then it must be a simple module, since if it has any nontrivial submodule $M'$, then $M'$ will be a smaller normal subgroup of $G$. Thus the general situation is that $M$ is some simple module over the ring $\mathbb{Z}[G/\mathbb{N}]$ (where $G/\mathbb{N}$ acts faithfully on $M$), and $D(G) \cong M \rtimes (G/\mathbb{N})$.

**Example A.10.** If we take $G = S_3$ in the above, we find that $D(S_3) \cong \mathbb{Z}/3 \times \mathbb{Z}/2 \cong S_3$. The 4-ary relation $\delta \leq S_3^{2 \times 2}$ corresponding to the trivial similarity from $S_3$ to itself is given by

$$\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta \iff \sigma(a) = \sigma(b) = \sigma(a') = \sigma(b') \land a^{-1}a' = b^{-1}b',$$

where $\sigma : S_3 \to \{ \pm 1 \}$ is the sign homomorphism.

We can think of the relation $\delta$ as having two “strands” corresponding to the two possible signs of permutations, and if we restrict to either strand then $\delta$ becomes an affine relation over $\mathbb{Z}/3$. The fact that we can multiply elements of $\delta$ which come from different strands and still get an element of $\delta$ is worth thinking about.

Now suppose that $G$ is some other subdirectly irreducible group such that $D(G) \cong S_3$, with monolith corresponding to $M \triangleleft G$ and $N = C_G(M)$. Then since $G$ is similar to $S_3$, we must have $M \cong \mathbb{Z}/3$ and $G/\mathbb{N} \cong \mathbb{Z}/2$ by Corollary A.88 with $G/\mathbb{N}$ acting on $M$ by negation since

204
$D(G) \cong M \times (G/N) \cong S_3$. If the action of $G/N$ on $N$ is given by an involution $\tau$, then for any $n \in N \setminus \{1\}$ we must have $M$ contained in the normal subgroup of $N$ generated by $n, n^\tau$.

In particular, if $N$ is abelian then we see that $n + n^\tau, n - n^\tau \in M$ for all $n \in N$, and additionally in this case $N$ must have prime power order by Theorem A.69. Thus if $N$ is abelian then we must actually have $N = M$, and $G \cong S_3$.

**Example A.11.** If we take $G = Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$ the quaternion group with $i^2 = j^2 = k^2 = \pm 1, ijk = -1$, then the monolith is equal to the center, corresponding to the normal subgroup $\{\pm 1\}$, and the centralizer of the monolith is the full congruence $1_{Q_8}$. Thus

$$D(Q_8) \cong \{\pm 1\} \cong \mathbb{Z}/2.$$  

The relation $\delta \leq (Q_8 \times \mathbb{Z}/2)^2$ is then given by

$$\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta \iff a' = (\pm 1)^{b+b'} a.$$  

This relation closely resembles an affine relation over $\mathbb{Z}/2$. 

205